Localisation Focus

THE INTERNATIONAL JOURNAL OF LOCALISATION

ISSN 1649-2358

The peer-reviewed and indexed localisation journal

VOL. 7 Issue 1

EDITORIAL BOARD

PUBLISHER INFORMATION

Editor: Reinhard Schäler, *Director*, Localisation Research Centre, University of Limerick, Limerick, Ireland
Production Editor: Karl Kelly, *Manager* Localisation Research Centre, University of Limerick, Limerick, Ireland
Published by: Localisation Research Centre, CSIS Department, University of Limerick, Limerick, Ireland

AIMS AND SCOPE

Localisation Focus – The International Journal of Localisation provides a forum for localisation professionals and researchers to discuss and present their localisation-related work, covering all aspects of this multi-disciplinary field, including software engineering, tools and technology development, cultural aspects, translation studies, project management, workflow and process automation, education and training, and details of new developments in the localisation industry. Proposed contributions are peer-reviewed thereby ensuring a high standard of published material. Localisation Focus is distributed worldwide to libraries and localisation professionals, including engineers, managers, trainers, linguists, researchers and students. Indexed on a number of databases, this journal affords contributors increased recognition for their work. Localisation-related papers, articles, reviews, perspectives, insights and correspondence are all welcome.

To access previous issues online go to http://www.localisation.ie/resources/locfocus/pdf.htm and click on the issue you wish to download. Use the following logon details - username: locfocsub and password: V711208

Members of **The Institute of Localisation Professionals (TILP)** receive Localisation Focus – The International Journal of Localisation as part of their membership benefits. Membership applications can be filed electronically from ***www.tilponline.org*** Change of address details should be sent to LRC@ul.ie

Subscription: To subscribe to Localisation Focus - The International Journal of Localisation visit www.localisationshop.com (subscriptions tab). For more information visit www.localisation.ie/lf

Localisation Focus – The International Journal of Localisation (ISSN 1649-2358) is published and distributed annually and has been published since 1996 by the Localisation Research Centre, University of Limerick, Limerick, Ireland. Articles are peer reviewed and indexed by major scientific research services.

FROM THE EDITOR

Localisation research is now firmly established as a field of academic and industrial research. A quarter of a century after the first localisation service providers emerged, allowing multinational companies to translate and adapt their products to the requirements of foreign markets, and developing the techniques and technologies to manage the global information flow, academia and industry have recognised that under the cover of "applications" there are fundamental issues waiting to be resolved by scientific research efforts. Years of persistent work to encourage new and established researchers to explore the underlying, principle issues in localisation through the establishment of academic competitions (Best Thesis and Best Scholar Awards), collaborative projects, summer schools, conferences and this international journal have been rewarded. Apart from the significant body of scientific research that is now available, it is the establishment of the Centre for Next Generation Localisation (CNGL) that provides the conclusive evidence that localisation is firmly on the scientific research agenda: four Irish universities and nine Irish and international companies, supported by significant funding from the Irish Government's Science Foundation, have brought together more than 100 researchers investigating the blueprint for next generation localisation paradigms. Localisation Focus - The International Journal of Localisation will dedicate its next issue exclusively to report on the work of the CNGL.

Meanwhile, the current issue offers an insight into four extremely challenging aspects of localisation, presented by researchers working both in industry and academia.

Martin Ørsted of Microsoft Ireland reports on his company's strategy to systematically capture and fix language, layout and functional problems arising when a product is being localised into an ever increasing number of languages. He shows that patterns can be identified that enable the identification of inconsistencies and issues across multiple versions that, with a single language approach, would have been very costly and difficult to identify and correct. As a consequence, the linear dependency between the number of languages a product is being localised into and the total cost of defect correction can be broken.

Few topics are currently being discussed as much as the post-editing of machine translation. Both machine translated segments and matches from translation memories are now often included in pre-processed content presented to translators. It is widely assumed that this approach improves both productivity and the quality of the end product. Yet, few attempts have been made to prove this assumption. Ana Guerberof Arenas reports on the results of a study she conducted and presents some surprising results.

There is a clear requirement to automate the post-editing of machine translated text to remove this often tedious and repetitive task from increasingly frustrated translators, and to increase the efficiency of the post-editing process. Midori Tatsumi and Yanli Sun report on the results of their experiment that compared an automated statistical post-editing approach for English text that was machine-translated into Chinese and Japanese. In addition to efficiency, they also looked into issues closer to the heart of the eventual readers, namely fluency and adequacy.

Patrick Cadwell's contribution also looks at the requirements of the consumer of the localised material, examining the question whether controlled language can increase the readability of technical texts. Thus, he expands the coverage of controlled language in localisation beyond that of a useful pre-processing step for machine translation.

We would like to encourage you to submit contributions to Localisation Focus - The International Journal of Localisation and to encourage colleagues to do likewise (for details see the back pages of this edition). This would be an excellent way to show your support for our continued efforts to develop localisation as an exciting and interesting field of academic and industrial research.

Reinhard Schäler

Systematic validation of localisation across all languages

Martin Ørsted
Microsoft Ireland
Martin.Orsted@microsoft.com

Abstract

As software companies increase the number of markets and languages that they release their products in, it may become necessary to change the localisation process for these products. Quality assurance (QA) is often viewed as an area where processes could be streamlined through automation and one method for doing this would be through the design of a localisation verification system that can validate single resources across languages as well as check for generic issues across multiple resources and languages. This article outlines a graduated approach to systemically capture and fix issues when a product is being localised into an increasing number of languages. By examining multiple languages, patterns can be identified that enable the identification of inconsistencies and issues that, with a single language approach, would have been very costly and difficult to unearth.

Keywords: *Localisation, Resources, Verification, Systematic, Multiple languages, Controlled language*

Introduction

The best place to address localisation issues is upstream. Much can be done here; the use of newer programming languages with more built-in error checking, the use of pseudo localisation[1] upstream, educating developers, the use of controlled English and source reuse systems can all help. There are however many reasons why the above options will never be implemented perfectly; deadlines, tradeoffs, the inadequacy of the development languages used, and so on. For these reasons systematic validation can improve localisation and noticeably drive down costs for multi-language releases so that the more languages produced the better the return.

In most traditional localisation efforts languages are treated rather independently, with little or no ability to leverage the testing performed for one language on another. The most common forms of leveraging are highly manual or risk based[2] or a combination of both. One way of leveraging is, for example, to not run low priority test cases on certain languages; another is manual regression of bugs found in one language against others. The use of pseudo localisation is also gaining broad acceptance in the industry and serves many needs. Using pseudo localisation with machine generated pseudo localised strings will allow for a fast check of the localisability of resources and can in general find most types of localisability errors up front. In this context pseudo localisation is often used to postpone the real localisation effort until RTM or RTW (Release to manufacturing or web) or at least shorten the parallel effort, which reduces resource churn and drives down localisation cost. Used in this way it also saves on development costs as the faster an issue is found the cheaper it is to fix it.

There are several goals behind the systemic validation of localisation across all languages. This article uses practical work that has been carried out in Microsoft over the years to map out how a methodology can be built around using systemic validation that can achieve higher savings and better turnaround times than the aforementioned approaches can deliver. The article will start by looking at the single resource approach, where Microsoft's rules based approach is explained, and over the course of this section it will show the kind of issues that one can systematically fix. It will then generalise the approach to a wider pool of resources. We will look at other methods for bug avoidance, and finish up by analysing how testing can systemically be reduced while quality is maintained, or improved, through the introduction of the outlined methods. In this way we will also look at how the traditional linear dependency between the cost of the test effort and the number of languages localised can be broken.

[1] Pseudo localisation is localising strings by replacing the typically US characters with characters from other code pages, and adding tagging before and after the string. Open could for example become become \?p??$@#. Typically the pseudo localisation process is fully automated so it is fast and cheap.

[2] Risk based through the use of orthogonal arrays for example.

The single resource

There can be many reasons why the localisation of a string can cause a bug, be it user interface or functional. In the functional space bugs can be caused by:

- Over-localisation: The string should not have been translated.
- Buffer limitation: The translation of the resource should not be more than a given amount of characters, generally referred to as a string length limitation.
- Illegal characters: Certain characters may not be allowed in the string
- Dependency: Two resources may have to be translated as one, in effect one resource is dependent on the other, references the other.
- Backward compatibility: This is a special case of dependency, basically where changing a string from one version to another could cause a loss of backward compatibility.
- Uniqueness: The string belongs to a group of strings that all have to have unique names (translations), for example, a list of commands.
- Placeholder over-localised: Some localisable strings have placeholders in them. If the placeholder gets localised the program cannot drop the information into the placeholder and display it.
- Required string decoration: Some strings may have control characters in the beginning or end of the string that should not be localised

There are other more special cases, but the above list captures the majority. In many instances strings that can break because of any of the above causes could have been bullet-proofed by the developer, but it is not always the case.

Below are a few examples of strings that would fall into these categories:

Rule	US string	Example loc	Issue description
Over-localisation	Common Files		Might refer to a registry string. Rather than localising the string the program will look up the localised name in the registry
Placeholder	The file %1 could not be opened because %2		%1 and %2 are placeholders
Decoration	\n\nOpen\n\n		\n is a new-line character, sometimes used in command line applications
Placeholder	The file %s was last opened on %d %d	On %d%d the file %s was last opened	%s and %d are positional placeholders, their position has to be maintained, changing them as shown will cause an intermittent memory protection fault

In Microsoft the original approach we had to systematically fix issues once we identified them was LocVer, short for Localisation Verification. For us, LocVer is still an essential part of our strategy. With LocVer, we can create rules to describe the limitations for a given string, and we can then run the rule engine against all languages. By doing so, we can ensure that the issue, once found, is validated and if needed fixed for all languages.

3 LocVer is Microsoft patented and patent pending

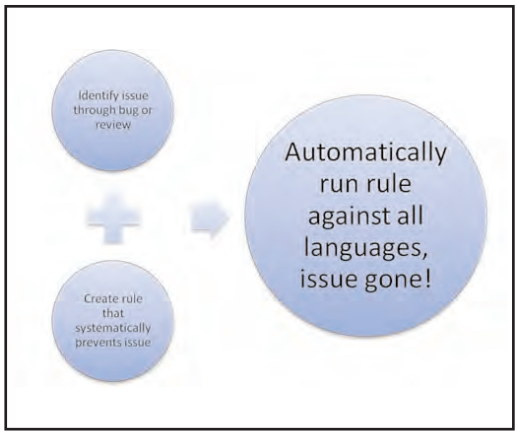

FIGURE 1: A GRAPHICAL REPRESENTATION OF LOCVER

Figure 1 shows a graphical representation of the idea. So as a hypothetical, practical example, the string **Current Accounts** could be used in Microsoft Excel, perhaps it becomes part of a Pivot table, and let us assume that we have identified, through trial and error, that the string cannot contain more than 30 characters. At this point in time an engineer would often have the choice of either transferring the bug to the developer to increase the buffer or accepting the limitation. We choose to accept the limitation and we create a LocVer rule: **MaxLength=30** (meaning that the translation is not allowed to contain more than 30 characters). Since we associate instructions with each string, the rule becomes an instruction for that particular resource ID.

We have a master repository for instructions across all languages, and that is where our new instruction will be added. The system is designed so that the localisation vendors frequently receive the latest instructions and has been designed to run this type of validation at each handoff, so in effect given a certain lapse time the rule will have migrated across to all languages, and for any language where the rule has been broken an error will be returned and logged, so that the issue can be resolved.

LocVer has been in use in Microsoft for many years now and has developed from supporting simple rules like the one above to more complex scripting rules. As we progressed with this we realised that there was a need for added functionality, such as the ability to conditionally apply a LocVer rule, perhaps the rule should only apply to a subset of languages, or maybe a subset of languages should be excluded. You can have strings that accept ANSI 1252 characters, but where Cyrillic characters may cause an issue, and hence you may lock translation for those. Or you may have a program where some advanced functionality is available for a few main languages (speech recognition for example), but since it is not available for other languages the items should not be localised.
There is a cost involved in this per resource based instruction approach. Whenever a resource changes, the associated instruction will have to be updated, so the adding of rules and their maintenance can become a serious effort. Measuring the return on investment can also be a little difficult, in effect how do you account for the bugs prevented?

Still, in many instances the individual instructions are often the only viable way of dealing with per string limitations. However this is not always the case and that is what the next section deals with.

Fixing systematically across a wider pool of string resources

The previous section dealt with a per resource approach to the systematic validation of resources. It works, it is proven and we use it a lot. However, the cost involved means that we have had to consider whether we could further develop the approach in such a way that all the benefits of the above system can be retained but without the management overhead of dealing with the individual resources.

There are several different approaches that can be considered to reduce the management overhead. At Microsoft we have, in practical terms, at least three concurrent systems in place, each serving different needs. One way of approaching it is to see if we can create generic rules. Where this is possible we can then remove many specific rules and rely on a few generic rules instead. This turns out to be very applicable for certain placeholders. If for example %1, %2 and %3 always denote placeholders, then we can remove the specific LocVer placeholder rules from the individual resources and create a generic rule that stipulates that %1, %2 and %3 are always placeholders and that the translation has to mirror the source in their usage. This is an approach that we use frequently to avoid functional issues.

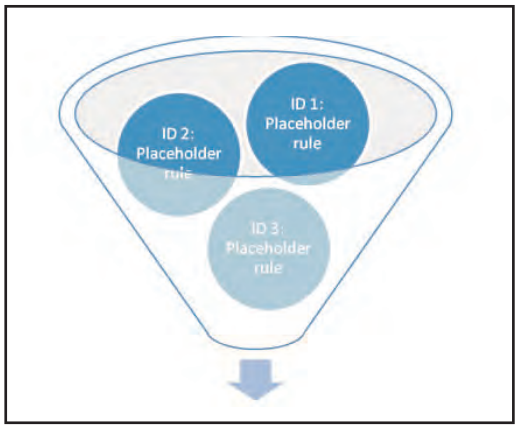

FIGURE 2: A GENERIC PLACEHOLDER RULE WITH NO ASSOCIATED ID

In terms of ensuring that certain keywords and copyright text are correct it is possible to do something quite similar. As a case in point, in the old days we used placeholder rules to ensure that certain legal text was kept across languages, but this is inappropriate for several reasons, not least of which is the maintenance associated with this method. In parallel with the generic rules we therefore keep a list of SQL queries that we can run to ensure that copyright text and application names are treated consistently.

This approach can be used, for example, where a product is developed with a code name up to a very late stage when the official product release name is decided upon. So, for example, InfoPath was referred to as XDocs during the development phase, and it was only very late in the process that it became InfoPath. In this case a simple SQL query could be run against all languages to identify places where localisers had "forgotten" to change the name.

We have built up a list of checks, either single language checks or relative to the US English source that we can run in these situations. For example, in the last release of Microsoft Office we might have run checks to find strings where the US source had "Copyright" in the text but the localised text didn't, or the US version had "2007" in the source but the translation did not contain it, or the US version had "Microsoft Word" in the source but the translation didn't.

There are many reasons why a more explicit per resource rule-based approach is not acceptable here. Firstly, it would involve far too many rules and drive up cost. Secondly, there is an advantage to having clear separation between the functional quality assurance and the legal/linguistic assurance. The functional quality assurance process is ongoing and needs a continuous focus. In general you expect very few false positives, in effect rules misfiring. As opposed to those for linguistic/legal material the rules can misfire frequently, and the linguistic quality issues are not as time critical as the functional ones; the functional quality has to be high continuously. The reason for this will be made apparent below.

A functional issue may block the testing of an area, and consequently the fixing of the functional issue may in turn uncover more issues. As opposed to that a linguistic issue may often be benign, not an issue at all. The string **"Microsoft Office Word could not open the file %1"** could, for example, in some languages be translated to the equivalent of **"The file %1 could not be opened"** as a space saving measure, and that would in many instances be quite OK. To complicate matters more, we would allow the use of the Cyrillic "i" instead of the Latin "i" for Cyrillic languages, hence the word **Microsoft** would not even match up for Cyrillic languages in a comparison between the language and US.

In terms of the legal/linguistic searches we find therefore that there is a trade off point, after which it is not worthwhile. We would tend to run our queries a couple of times during production and pay the localisation vendor to review the results, calling out the ones that need fixing and getting them fixed. Geopolitical issues can be dealt with in a similar way. We maintain lists of words or phrases that are geopolitical on a per language basis, and we can run them through the same system that we use for legal quality.

Adding the language dimension

Over the last 10 years the amount of languages we localise into at Microsoft has seen a dramatic increase. When I started as a localiser in 1996 we probably localised into around 15 languages, now we are getting closer to 100, if not exceeding this number. This pattern seems to be repeating itself within the industry as a whole. The way we approach localisation changes with the addition of more and more languages. Approaches that would previously have been too costly start to become viable. Likewise, certain approaches become possible that before would have been impossible. That is the topic of this section. We also, conveniently, enter the newer and most exciting or promising areas of localisation innovation here. Whereas there are innumerable examples relat-

ing to the last two sections, here there are fewer examples and, of the examples that there are, some may be rather theoretical.

In the previous section I touched upon running queries against a target language and source language (US), to find product names and copyright issues. With the addition of the language dimension smarter queries can be run that look across multiple languages to find patterns. If nine out of ten languages turn out to start or end with a certain sequence of characters, or if for example the word **Microsoft** appears in nine out of ten, then there is good reason to assume that it should be in the last language as well, it becomes a pattern that triggers an exception for evaluation. The return on investment (ROI) on creating various different rules obviously goes up with more languages added, so this is a space that is open for creativity.

One thing you can systematically look for is true repeated strings. Quite often, just because two US strings are identical one cannot assume they carry the same meaning, **Open** can be the imperative, as in the command **Open the door**. It can also be the infinitive, to **Open**, and the two are translated differently for most languages. But if the strings are the same for US and nine out of ten localised languages, then the deviation for the tenth language is probably an error. Going further, one could for all languages after, for example, the tenth language simply remove identified repeats and only have them translated once per language.

Another effort that we have invested in is tweaking our pseudo localisation engine, so that it understands and adheres to our LocVer rules. That means we only ever find the same issue once, since the pseudo localised strings won't break the rules we have already added. Pseudo localisation on various different languages is therefore a key part of our strategy. The question becomes "what further testing needs to be carried out on fully localised languages?" Figuring this out involves analysing exactly what kind of testing needs to be performed on the actual languages themselves, and that means analysing the localisation model.

The localisation model in this context is the various processes that are applied to get from the US English source files to the localised files. Each process needs to be analysed to figure out what error sources the process can introduce and what error sources are systematically prevented. For each error source identified that requires testing to revisit the various languages, the challenge is to identify a solution that can systematically fix the root cause of the issue so that the need to test the various languages is kept to a minimum or eliminated in certain instances. DAL (Dynamic Auto Layout of dialogs) for example may, implemented correctly, mean that it won't be necessary to review each localised dialog, but rather a subset, or just the pseudo dialog depending on confidence levels. These confidence levels are subjective and based on experience.

Similarly, we have introduced a systematic way of assigning hotkeys, so that the assignment of hotkeys per language is really a matter of running a set of scripts at a chosen moment. This is accomplished by building up a list of all resources with hotkeys and where they appear. Building up that list is complicated, it is partially populated through the use of automated trawler tools that identify which resources belong to which dialogs, but it can also be populated or improved upon through manual entries. So the process is a bit costly. With the lists populated we can then run a tool on a per language basis that knows which characters can be assigned hotkeys per language, and that can resolve hotkey conflicts including dealing with resources that appear in multiple dialogs or menus.

Other concerns that need to be addressed are the behaviour of the product with various code pages. So one thing that is possible is to group languages under ANSI code pages and test a representative from each code page exhaustively. Again, getting full Unicode from scratch would be an advantage of course, but failing that this approach can drive down cost.

With all of the above implemented, or parts of it, one can evaluate the approach to testing. Engineering are in a position to guarantee that testing only ever needs to find an issue once, and engineering can guarantee that it will systematically be fixed across all languages with no further test need for verification. That in turn can facilitate moving away from very specific test cases to higher level Test Design Specifications. This is helped by the fact that the testers, assuming you use the same testers for all languages, gradually build up a better understanding of the tested product and the type of localisation bugs that appear. So rather than executing very specific test cases step by step, overall quality can be improved at a reduced cost by having the testers test features with some high level guidelines that lead them through the features but are still much more

abstract than specific test cases. The result will be that the tester for each language will go through a feature based on his or her high level of understanding of this feature, and since the tester is not following a strict script we can count on a degree of variation or randomness to be introduced in the ways the various languages are tested. The introduced randomness that this brings adds value to the process, because through the testing of various languages, pseudo or real, the randomness introduced will ensure better overall test coverage as compared to strictly exercising the same test paths per language.

We saw an example of this in a recent Office release cycle. We had a test case that stipulated the comparison of two files, and included the two files as part of the test case. One tester chose to compare two files that were already on disk rather than take the two supplied for this test case and as a result uncovered a bug that for several languages had not been found.
All of the above efforts should mean that the test and engineering cost per language can be reduced. In an ideal world the cost of adding an extra language should get close to the cost of the pure translation; realistically in our case we have not achieved that, but we have definitely seen dramatic cost reductions. It is difficult to give exact savings numbers, due to a number of factors such as the difficulty in calculating the savings of a bug that has been avoided and also because we have introduced these efficiencies gradually, and in some cases are still working on fully introducing them. But to give an idea of the importance of this, the group that I work in is requested to make serious savings version on version, and this is one of our favourite hunting areas for those savings.

Controlled English and Machine Translation
Another benefit of localising into many languages is that more structured approaches within the area of controlled English and Machine Translation (MT) become feasible. So at some stage it makes sense to use controlled English or elements of controlled English, starting with simple checks on sentence length and verbs in passive tense and moving on from there.

Machine translation is trickier. There is no standard emerging in the MT space to automate translation in an intelligent way. Also, most MT engines go from English to another language, but much more can be gained with an effort that translates between close language pairs, for example Iberian Portuguese and Brazilian Portuguese, or Norwegian Bokmål and Norwegian Nynorsk. The localisation verification models can ensure that rules are not broken and cost can be reduced, although an initial investment is needed. As an alternative to full MT, automated transliteration can be considered from some languages for language pairs that are closely related.

Conclusion
There is no substitute for a well engineered product in the first place. Any bullet-proofing that can be done within the code is of course preferred. In the real world there will however always be limitations to the upstream efforts, and that is what we are looking at here.

Figure 3: Evolution and return on investment of various approaches

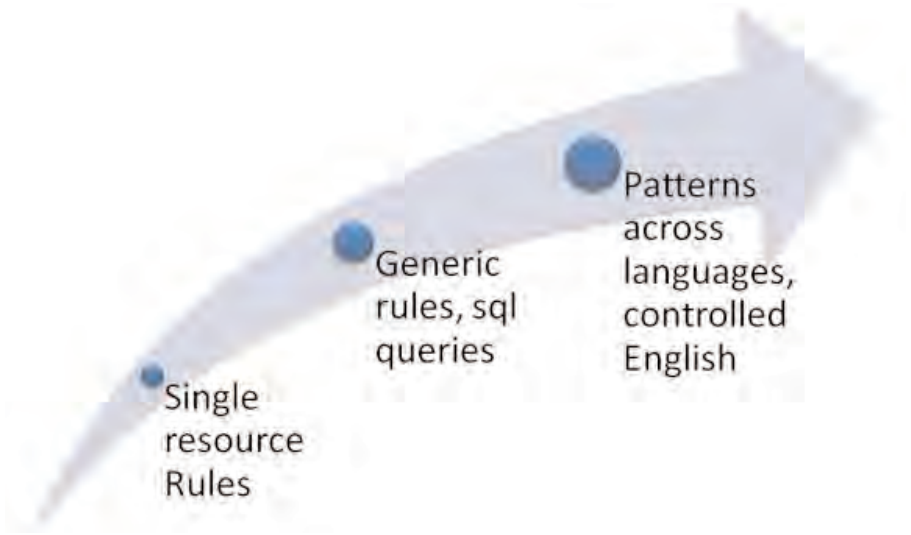

Figure 3 illustrates the evolution and the return on investment of the various approaches. The further to the right the greater the impact, but this doesn't mean that single resource rules are not valuable, just that they are rather costly in comparison to the other approaches.

The above sections have outlined a graduated approach to systemically capture and fix issues when a product is being localised into an increasing number of languages. We began by looking at approaches to individual strings. This is often a necessary approach in functional bug fixing and prevention, and on the positive side it means that we only ever have to catch a specific issue once, and we can then systematically ensure that, should the issue occur in other languages, an error will be raised and we can deal with the issue manually or automatically. The downside to this approach is that it necessitates either an inspection of all resources, where certain kinds of issues cannot be found, or that the issue is identified as a bug at some stage in a language. The other downside is the actual cost of running a system like this; depending on the thoroughness applied the cost can be quite severe.

A system where the rules are generic is therefore preferred, but will never be able to cover everything. The advantages of the generic rules are that they look for identifiable patterns and automatically apply when a pattern is identified. Therefore, new strings that conform to the same patterns, for example %1 as a placeholder, are automatically covered as soon as they are added. String changes, the addition or removal of placeholders, will automatically be covered, and the management overhead is dramatically reduced in comparison to the single resource rules. Similar benefits of scale can be achieved in the legal and linguistic space through the use of SQL type queries.

Finally, with the addition of multiple languages pattern recognition across languages becomes interesting. Certain types of errors become much easier to detect, and things like controlled English make it possible to ensure a higher end localisation quality (because the localisable text is less ambiguous).

It is therefore possible to use the fact that a product is localised into many languages to systematically deal with some issues, to apply learning across the languages that can help raise the overall quality of the product, and to drive down the cost of testing and bug fixing. This approach has the potential to break, what is often, otherwise, a linear dependency between the number of languages you localise into and the total cost of bug fixing and testing.

Productivity and quality in the post-editing of outputs from translation memories and machine translation

Ana Guerberof Arenas
PhD programme in Translation and Intercultural Studies
Universitat Rovira i Virgili, Tarragona, Spain
Ana.Guerberof@gmail.com

Abstract

Machine-translated segments are increasingly included as fuzzy matches within the translation-memory systems in the localisation workflow. This study presents preliminary results on the correlation between these two types of segments in terms of productivity and final quality. In order to test these variables, we set up an experiment with a group of eight professional translators using an on-line post-editing tool and a statistical-based machine translation engine. The translators were asked to translate new, machine-translated and translation-memory segments from the 80-90 percent value range using a post-editing tool without actually knowing the origin of each segment, and to complete a questionnaire. The findings suggest that translators have higher productivity and quality when using machine-translated output than when processing fuzzy matches from translation memories. Furthermore, translators' technical experience seems to have an impact on productivity but not on quality.

Keywords: *Translation memory, machine translation, post-editing, revision, productivity, quality, errors, editing, professional translators, experience, fuzzy match, processing speed, localisation*

Introduction

New technologies are creating new translation processes in the localisation industry, as well as changing the way in which translation is paid for. In the past, translation involved precisely that, the translation of entire software, documentation and help materials into new target texts for the local markets. As localisation matured, translation memories (TM) were created and texts were recycled in different but rather similar projects. Productivity increased and consequently prices of translations decreased. Since the 1980s, machine translation (MT) technology has improved significantly and has been incorporated into the localisation workflow as another type of translation aid, rather than attempting to have a fully automatic high-quality translation. It remains to be seen what effect this technological development will have on pricing structures.

Major software development companies now pre-translate source text using existing translation memories and then automatically translate the remaining text using a machine-translation engine. This "hybrid" pre-translated text is then given to translators to post-edit. Following guidelines, the translators correct the output from translation memories and machine translation to produce different levels of quality. Gradually this activity, post-editing, is becoming a more frequent activity in localisation, as opposed to the full translation of new texts.

In an industry that moves so rapidly, there is more focus on finalising projects than on the process itself. Therefore these translation aids are used in the localisation workflow with limited data to quantify the actual translation effort and the resulting quality after post-editing. Since productivity and quality have a direct impact on pricing, it is of capital importance to explore that relationship in terms of productivity and quality of the post-editing of texts, coming from translation-memory systems and machine-translated outputs, in relation to translating texts without any aid.

In this context, it seems logical to think that if prices, quality and times are already established for TMs according to different level of fuzzy matches then we only need to compare MT segments with TM segments, rather than comparing MT output to human translation. Therefore, once the correlation is established, the same set of standards for time, quality and price can be used for the two types of translation aid.

Preliminary premises

After a study by Sharon O'Brien (2006) where she establishes a correlation between MT segments and TM segments from the 80-90 percent category of

fuzzy match, we formulated our initial hypothesis. This one was that *the time invested in post-editing one string of machine translated text will correspond to the same time invested in editing a fuzzy matched string located in the 80-90 percent range.* This hypothesis is predicted on the assumption that the raw MT output is of reasonable quality according to the Bleu Score (Papineni et al 2002, p. 311).

Measuring productivity on its own, as in our first hypothesis does not make sense if it is not done in relation to an equal level of final quality. If the time necessary to review MT segments is greater than the time necessary to review New or TM segments, the productivity gain made during the translation and post-editing phase would be offset by the review phase. Therefore, we claimed that *the final quality of the target segments translated using MT is not different to the final quality of New or TM segments.*

Localisation has a very strong technical component because of the content as well as the tools required. On many occasions we associate technical competence with speed, that is, the more tools we use the more automated the process becomes and the less time we spend completing a project. Therefore, our third hypothesis claimed *that the greater the technical experience of the translator, the greater the productivity in post-editing MT and TM segments.*

Methodology

In order to prove our hypotheses we carried out an experiment with nine subjects. One subject carried out the preliminary test and the remaining eight performed the actual pilot experiment. The translators used a web-based post-editing tool to post-edit and translate a text from English into Spanish. The text had 791 words; 265 words of new segments (new text to translate), 264 words of translation-memory segments (Trados was used to create the fuzzy matches) and 262 words of machine-translated segments (Language Weaver's statistical-base engine was used to create the output). We selected a supply-chain software product for the corpus as we wanted to use typical content from the localisation industry. At the end of their assignment, the subjects filled in a questionnaire with information related to the pilot experiment and their own experience in the field. The final output was then revised, errors were counted and conclusions drawn.

Experiment design

Translators

We contacted a group of nine professional translators, five women and four men, with ages ranging from 22 to 46 years. They all have first degrees or Master's Degrees in Translation. Their experience ranges from 1 year to more than 10 years in the translation industry and most have specific experience in localisation. They were contacted by email in all cases and they received no training to carry out the pilot experiment, only a set of instructions. The translators were not paid for the work that they carried out and although they knew the work was for research, and they might have inferred from the tool that the research dealt with machine translation, they were not given any specific information on the topic. Due to the fact that they were professional translators working for a short period of time and that they knew their work would be part of a research project, we would assume they maintained their usual working standards.

Training the engine

We provided Language Weaver with a translation memory containing 1.1 million words and a core glossary. They then created a customized engine using the relevant translation memories and a validated terminology list. Finally, they uploaded these segments into the post-editing tool.

Creating the translation memory segments

For our research we needed to create a file containing segments in the 80-90 percent category to feed these lower fuzzy matches into the tool. To prepare the file, we pre-translated existing html files from a help project of the supply-chain software with a previous memory in order to obtain fuzzy matches using the option Pre-translate in SDL Trados (version 7.1). We created ttxs files with different fuzzy match values. We then exported all segment pairs together with their corresponding fuzzy level (54, 75, 86 and so on) to Excel. This was done with a small tool created specifically for this purpose called Slicer.

Since we only needed a small number of words and not all of the segments, we randomly selected a number of segments from each category using the function *Random.between* in Excel. This gave us the desired number of segments in a random selection.

Post-editing tool

The translators were able to connect to the post-editing tool online. They could then translate/post-edit

the proposed segments of text without knowing their origin (MT, TM or New segments) and the tool measured the time taken in seconds for each task. The post-editing tool required the translator to log on with a specific user name and password, so each translator could only see the text assigned to them. Once they opened the task, they were presented with a screen containing the actual task as seen in Figure 1.

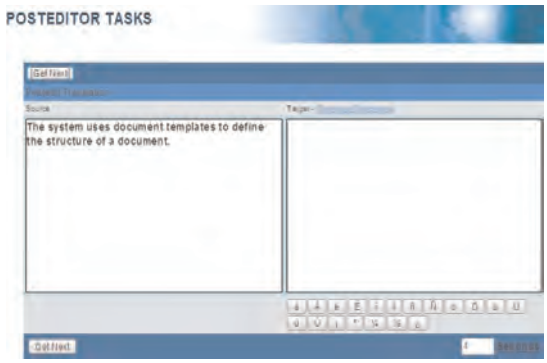

FIGURE 1: WEB-BASED TOOL FOR POST-EDITING TASKS

The Source window contained the source text in English, and the Target window contained either a blank screen or a proposed text in Spanish. The Spanish text was either a MT or TM segment. Once finished with a string, the translator had to click on the Get Next button and proceed with the following segment until they had reached the end of the assignment.

Questionnaire
The aim of the questionnaire was to define the translators' experience in localisation, tools, subject matter and post-editing MT. The questionnaire consisted of 17 questions that addressed these areas. It contained dichotomous questions, questions based on level of measurement and filter questions. The main aim of the questionnaire was to describe the group of translators and establish their experience in localisation, supply chain, knowledge of tools, and post-editing MT, as well as gather their views on MT. We matched the answers from the translators to the processing speed from the tool and the number of errors in the final sample.

Quality of the samples
The final target texts were checked to see the number of final errors in each sample. This could give us an indication of productivity versus quality. If translating with an aid was faster than the human translation, but there were more errors, then more time would be needed in a final review phase, thus altering the real translators' productivity.

We used LISA standards to measure and classify the number of errors. We classified the errors according to their source (New, MT or TM segments) to see if each category had similar number of errors. We classified errors according to type to see their frequency in each type of segment. Lastly, we matched the errors with the productivity by means of a coefficient of error based on the average revised word per minute.

Results

Productivity

Processing speed
Processing speed is the processing time in relation to the words processed in that time, that is, words divided by time. The number of words was almost identical in the three categories, New (265 words), MT (262 words) and TM (264 words) consequently our processing times and processing speeds were not notably different. The results are given in Table 1. We have highlighted in bold the maximum and minimum values per segment category.

Translator	New	MT	TM
TR 1	12.12	18.69	14.52
TR 2	10.76	10.28	10.75
TR 3	**22.08**	**21.21**	16.40
TR 4	8.55	9.79	10.22
TR 5	**5.85**	12.04	8.18
TR 6	8.11	9.12	**8.08**
TR 7	20.03	20.77	**18.48**
TR 8	7.42	**8.96**	10.47

TABLE 1: TRANSLATORS' PROCESSING SPEED IN WORDS PER MINUTE PER SEGMENT CATEGORY

This table shows that four out of eight translators performed faster using MT (TR 1, TR 5, TR 6, and TR 7), two were faster translating New segments (TR 2 and TR 3), and two were faster processing TM segments (TR 4 and TR 8). In total, six were faster using a translation aid than translating without any aid. Only TR 2 shows the slowest processing speed when using MT by quite a small margin in comparison to New or TM segments.

Let us have a look at the statistical summary:

Translator	New	MT	TM
Mean	**11.87**	**13.86**	**12.14**
Median	**9.66**	**11.16**	**10.61**
Std. Deviation	**6.02**	**5.40**	**3.87**
Max	22.08	21.21	18.48
Min	5.85	8.96	8.08
Range	16.23	12.25	10.41
1st Quartile	7.94	9.62	9.71
3rd Quartile	14.10	19.21	14.99
Diff quartiles	6.16	9.59	5.28

TABLE 2: STATISTICAL SUMMARY OF PROCESSING SPEED

Table 2 shows, in bold, that translators process, on average, more words per minute in MT than in TM or New segments and that they process, in turn, more words in TM than in New segments. All the same, the standard deviation is extremely high, 6.02 for New segments, 5.4 for MT and 3.87 for TM. For example, the range of variation (seventh row) between the maximum and minimum values is 16.23 words in New segments, 12.25 in MT segments and 10.41 in TM segments. Hence the mean, as a unique value is, not a fully representative number for the data shown here. The median for all the values, in bold, tells us that MT continues to be faster than human translation (approximately 16 percent) and faster than using TM (approximately 5 percent). The first quartile (eighth row) shows that processing TM segments is faster than processing New or MT segments, only 1 percent higher than MT, and in turn MT is faster than processing New segments, by approximately 21 percent. In this case, the quartile analysis shows that the translators that process fewer words per minute have a higher correlation between TM and MT than the group that processes more words. The second quartile, equivalent to the median, shows that MT is faster than New and TM segments, although the difference between MT and TM values is not very pronounced. In the third quartile, ninth row, we see that the speed for New segments and TM is extremely close, while MT is definitely faster. The difference between the first and third quartile, tenth row, shows us that there are pronounced differences, especially in MT with 9.59 words difference, then in New with 6.16 and in TM with 5.28 words.

Productivity gain

The productivity gain is the relationship between the number of words translated per minute per single translator without any aid and the number of words translated per minute by the same translator with the aid of a tool, TM or MT. This value is expressed as a percentage value.

In Table 3 we see the statistical summary regarding productivity gain:

Translator	MT vs. New	TM vs. New
Mean	**25%**	**11%**
Median	**13%**	**10%**
Std. Deviation	37%	23%
Max	106%	41%
Min	-4%	-26%
Range	110%	67%
1st Quartile	2%	-2%
3rd Quartile	29%	25%
Diff quartiles	27%	27%

TABLE 3: STATISTICAL SUMMARY OF PRODUCTIVITY GAIN

The mean values in MT and TM in relation to New segments show us that translators have a higher productivity gain if they use a translation aid. The gain was higher in MT segments than in TM segments, with 25 and 11 percent respectively. Nonetheless, the standard deviation is extremely high and the range of variation is very pronounced. The median value, in bold, shows that MT has a higher productivity gain (13 percent) but that the difference with TM is not very pronounced (10 percent). In the first quartile, eighth row, the productivity gain provided by the translation aid, MT or TM, is not very pronounced, and relatively similar (4 percent variance). Still the productivity gain for TM is negative, indicating a decrease in productivity. This quartile includes TR 2, TR 3, TR 7 and TR 6. In the third quartile, the productivity gain for both MT and TM is higher (29 and 25 percent respectively). This quartile includes TR 4, TR 5, TR 8 and TR 1. The highest productivity gain, if we take the statistical values, never goes over 29 percent (third quartile using MT). We should remark that the values in the quartiles correspond partly to the faster and slower translators and this seems to indicate that faster translators take less advantage of translation aids than do slower translators.

Quality

Existing errors and changes in MT and TM

Before we looked at the errors found after the assignment was completed, we needed to look at the number of errors and corrections existing in the MT and TM segments before the pilot took place. Otherwise,

if we found that one category, MT or TM, contained more errors than the other, it would have been logical, although not necessarily true, to assume that there would be more errors after the assignment was completed in that same category. Similarly, we classified the errors found using the LISA standard and we had identified the number of changes that were necessary to perform in the TM segments.

The TM segments contained 1 Mistranslation, 1 Accuracy, 1 Terminology and 2 Language errors. These five errors came from the legacy material used to build the translation memory and were therefore made by human translators. There were 17 changes needed in the text. These changes were text modifications, insertions and deletions between the original source text and the new source text. This meant that there were 5 existing errors and 17 changes to make in the TM segments.

On the other hand, the MT segments contained 25 Language and 2 Terminology errors, a total of 27 existing errors in the MT segments. The typical errors found in MT output were wrong word order, grammar mistakes (concordance of verb and subject, concordance of genre) and inconsistent use of upper and lower cases. There were also a couple of cases where the MT engine chose the wrong term for the cotext given.

A priori, the number of existing errors and changes in TM versus the ones in the MT segments was very similar: 22 in the TM segments versus 27 in the MT segments, and this meant that the source text should not necessarily condition the final target text. The actual process needed to correct the texts was different in our view. This was due to the fact that the TM segments, on the one hand, needed insertions, changes and deletions where it was necessary to constantly refer to the source text, as well as 5 "standard" errors where the main reference was the target text. On the other hand, MT errors involved mainly language changes that were quite distinct and where a constant reference to the target text was necessary because they involved changing the word order, use of verb tenses, use of upper and lower cases and concordance of number. This difference in the required post-edit approach could mean different results in the final text depending on where the focus was when translators were working on the target text. It is important to mention at this point that translators did not know the origin of the segments (MT or TM) and obviously if these segments were full (100 percent) or fuzzy matches (54-99 percent).

Error analysis

We used the LISA form in the eight samples and we counted the errors according to its classification and according to the type of segment in order to compare the results. The classification of errors was carried out by the researcher mainly due to time and budget limitations and also because the researcher had extensive experience in reviewing these type of texts in this language combination. The texts were corrected and then compared against each other to assure that the same classification criteria were followed in all texts.

Table 4 shows the final number of errors per translator according to the type of segment, and the total number of errors. The table is sorted according to ascending total errors. Totals are highlighted in bold.

Translator	New	MT	TM	Totals
TR 3	1	1	4	**6**
TR 2	2	3	6	**11**
TR 4	2	5	6	**13**
TR 1	2	3	10	**15**
TR 6	4	5	8	**17**
TR 8	6	3	9	**18**
TR 7	7	5	9	**21**
TR 5	3	9	13	**25**
Totals	27	34	65	**126**

TABLE 4: NUMBER OF ERRORS PER TYPE OF SEGMENT AND TRANSLATOR

Table 4 shows that all segment categories contain errors, and all translators have errors in all categories. There are a total of 126 errors in the final texts. A total of 27 errors are found in the New segments and 99 in the combination of TM and MT segments. Translators did not have the possibility, when using the tool, to go back and correct their own work and the segments have not been reviewed by a third party. We nevertheless see that in all eight cases there are more errors in TM segments than in any other category. In five out of eight cases, there are more errors in MT than in New segments (TR 1, TR 2, TR 4, TR 5 and TR 6); in two cases (TR 7 and TR 8) there are more errors in New than in MT segments; and in one case there is an equal number of errors in both New and MT (TR 3).

The first striking result is that the number of errors in TM segments (65) is 141 percent higher than that of the New segments (27) and 91 percent higher than that of the MT segments (34). MT segments, on the other hand, contain 26 percent more errors than New

segments. We find that the number of errors in TM segments is consistently higher in all eight cases while the errors for New and MT segments vary among the subjects.

Errors per type

We have analysed how errors are distributed according to the LISA standard to see if the typology of errors varies depending on the type of source text, in order to understand if the type of text has an effect on the number of errors. We can see this analysis in Table 5:

Type of error	New	MT	TM	Totals	% New	% MT	% TM	% Total
Mistranslation	10	2	8	20	**8%**	**2%**	**6%**	**16%**
Accuracy	9	14	34	57	**6%**	**11%**	**27%**	**44%**
Terminology	2	9	9	20	**2%**	**7%**	**7%**	**16%**
Language	6	8	14	28	**6%**	**6%**	**11%**	**23%**
Consistency		1		1	**0%**	**1%**	**0%**	**1%**
Totals	**27**	**34**	**65**	**126**	**21%**	**27%**	**52%**	**100%**

TABLE 5: NUMBER AND PERCENTAGE OF ERRORS PER TYPE OF ERROR

There are 57 Accuracy errors that represent 44 percent of the total number of errors (almost half of the errors), and 34 of them, that is 27 percent of all the errors, are found in the TM segments. There are 9 Accuracy errors in New segments and 14 in MT, representing 6 and 11 percent respectively. One possible explanation for this number of errors in the TM segments could be that when translators are presented with a text that flows "naturally" like a human translation they seem to pay less attention to how accurate that sentence is. On the other hand, because errors in MT segments are so obviously wrong, the mistakes seem to be easier to detect. As we explained above, most of the changes in TM required the translator to look at the source text and not just focus on the proposed target. The fact that the TM segments have so many errors could be explained by the fact that translators possibly consulted the source text less than they would have if they had been translating a new text with no aid. We have seen in previous studies that monolingual revision is less efficient than bilingual revision (Brunette et al. 2005), that there is a trend towards error propagation in the use of TMs (Ribas 2007), and that using TMs increased productivity, but "translators using TMs may not be critical enough of the proposals offered by the system" (Bowker 2005, p.138) and they left many errors unchanged.

In our study there are 29 Language errors that represent 23 percent of the total number of errors: 14 of them, that is 11 percent, are found in TM segments while 6 and 8 (6 percent) are found in New and MT segments respectively. We see again in this case that the TM contains the most errors and, again, this could be due to the reasons explained above: when translators are provided with a text that flows naturally they seem to accept the segments as they are without questioning the text correctness. It is true that some errors could have been spotted on a second review, but we can say that errors in TM were not as frequently spotted as the ones in the MT segments.

From the 20 mistranslation errors, 10 are found in the New segments, representing 8 percent of the total, 8 errors are found in TM and only 2 mistranslation errors are found in MT representing 6 and 2 percent respectively. The fact that there are so few mistranslation errors in MT segments might indicate that using MT helps translators clarify possibly difficult aspects of the source texts thus improving general comprehension of the text.

From the 20 Terminology errors, only 2 are found in the New segments as opposed to 9 in both MT and TM segments. This seems to indicate that translators tend to consult the existing glossaries more when they are presented with new texts, rather than questioning the proposed terminology used in MT and TM. It might be logical not to check terminology in a pre-translated text, but terminology is not always correct in TM and MT outputs due to updates and changes in existing terminology. This indicates that instructions should be provided to reviewers or translators to specifically check glossaries or, alternatively, terminological changes need to be made directly to the TM or MT before the translation process begins.

The consistency error found in the MT segments that represent 1 percent of the total is related to the inconsistent use of upper and lower cases and it is a reflection of a known issue in MT output. We would venture that if the translators had received specific instructions on output error typology, this error would have been corrected.

Errors vs. productivity

We have established that an increase in productivity cannot be considered in isolation from the quality of the samples. So how does the number of errors found in the samples affect the overall productivity of the translators? Can we say that using MT or TM decreases or increases the productivity of a translator taking into account the final errors? To find an answer to these questions, we decided to penalise translators in their processing speed according to the number of errors made. To do this, we calculated a general coefficient of error to be used as a form of penalty (or correction) in words per minute and then we applied this coefficient to the processing speed of the eight subjects in order to see the impact of errors on the productivity gain.

Calculation of the error coefficient

We realised that the best way to determine the error coefficient would be to measure the reviewing time of these segments in a standard revision process by a third party. In this case, because the review is not part of the scope of this study, we took the metrics used for reviewers of localisation texts; approximately 7500 words per day (this figure may be higher or lower depending on the metric used by each individual localisation agency). With this figure in mind, we established that a reviewer reviews 0.26 words per minute (if we took a higher figure the value would be of course higher). We took the number of errors per translator and we applied the coefficient of error for each source of error and then recalculated their processing speeds, thus obtaining a final figure that reflected the impact of errors on their processing speed.

Once we had the new processing speeds for all translators, we recalculated the productivity gain comparing the different categories in order to see the impact on productivity that the errors might have had in a working environment. Negative values are highlighted in bold.

Translator	Total processing speed	MT vs. New	TM vs. New
TR 1	41.43	54%	3%
TR 2	28.93	**-7%**	**-10%**
TR 3	58.53	**-4%**	**-30%**
TR 4	25.18	6%	8%
TR 5	19.57	91%	**-5%**
TR 6	20.89	11%	**-15%**
TR 7	53.82	7%	**-11%**
TR 8	22.17	40%	39%

TABLE 6: TRANSLATORS' PRODUCTIVITY GAIN MINUS COEFFICIENT OF ERROR

In Table 6, MT is still faster than translating with no aid in six out of eight subjects (TR 1, TR 4, TR 5, TR 6, TR 7 and TR 8). The other two subjects (TR 2 and TR 3) have a negative value. This value has increased for TR 2 and remained stable for TR 3 (who made an equal number of errors in MT and TM categories), and in both cases the negative value is never below 7 percent. TR 4, TR 6 and TR 7 show a positive value of around 10 percent. On the other hand, TR 1, TR 5 and TR 8 show a positive value above 40 percent. Even if errors are considered, using MT is still more productive than no aid at all.

If we look now at the productivity gain of TM, the changes are more pronounced. Five out of eight cases have a negative productivity when compared to New segments (TR 2, TR 3, TR 5, TR 6 and TR 7), and in four cases the negative value is equal to or below minus 10 percent. In the case of TR 3, the value goes down to minus 30 percent. In two other cases (TR 1 and TR 4) TM brings a slight productivity increase with 3 and 8 percent respectively. Only the remaining case (TR 8) seems to have a pronounced productivity increase, with 39 percent. If errors are considered, using TM fuzzy matches (80-90 percent) does not appear to be productive when compared with translating without any aid.

In brief, if we consider errors when calculating the productivity gain, we see that although MT seems to play an important role in increasing productivity in most cases, TM has the opposite effect. It is important to remark here that we are referring to segments that belong to the 80-90 percent category of fuzzy match and not TM segments that include all levels of matches. It could well be that this translation memory as a whole provides a productivity increase for translators. But the 80 to 90 percent category of fuzzy matches does not appear to do so, and this is remarkable if we consider that these segments tend to be paid at 60 percent of their value (the global price including review), thus assuming a 40 percent productivity gain, and that this productivity was not achieved by any of our translators when errors are considered.

Table 7 shows the statistical summary of the new productivity gain. Mean and Median values are highlighted in bold.

Translator	MT vs. New	TM vs. New
Mean	**25%**	**-3%**
Median	**9%**	**-8%**
Std. Deviation	34%	20%
Max	91%	39%
Min	-7%	-30%
Range	98%	68%
1st Quartile	3%	-12%
3rd Quartile	43%	4%
Diff quartiles	40%	16%

TABLE 7: STATISTICAL SUMMARY OF PRODUCTIVITY GAIN MINUS COEFFICIENT OF ERROR

The correlation between MT and TM in relation to New segments shows that translators have a higher productivity gain if they use MT but a negative productivity gain if they use TM (80-90 percent matches). The range of variation is very pronounced (TR 5 has a value of 91 percent as opposed to TR 2 who has -7 percent). If we take the mean values, in bold, we see that MT has a productivity gain of 25 percent while TM presents a negative value of minus 3 percent in comparison to the previous positive value of 11 percent. The median values for both MT and TM have changed from 13 to 9 percent in MT and from 10 to minus 8 percent in TM. The first quartile shows that the productivity gain provided by MT is small with just 3 percent and negative in the TM with minus 12 percent. In the third quartile, the productivity gain for both MT and TM is positive (43 and 4 percent respectively).

Technical experience

Our third hypothesis claimed that the greater the technical experience of the translator, the greater the productivity in post-editing MT and TM segments. The first question that comes to mind is "What does technical experience mean?" We are aware that the term embraces several aspects of a translator's competence. For the purpose of this study we have defined technical experience as a combination of experience in localisation, in knowledge of tools, in subject matter (in this case supply chain), and in post-editing of machine translated output.

We obtained this data from the questionnaire that was provided to the translators at the end of the assignment. This data was then contrasted with the translators processing speed and number of errors to see if there was a correlation between technical experience, processing speed and errors. We took the processing speed as a result of the experiment without including the coefficient of error because we analyzed the errors separately. We took the mean in the processing speed as the number of subjects was smaller than in the productivity section, in the sense that all subjects were grouped according to experience thus decreasing the number of subjects per group, and the mean and median obtained were in most cases the same value.

The fact that the group was small and that the data obtained in terms of processing speed was dispersed made drawing final and general conclusions on any correlation between technical experience and productivity difficult. Nevertheless, we think it was necessary to correlate the processing speed obtained from the post-editing tool, errors and the questionnaire, even if it served only to test our methodology.

Summary data on translators' experience

To summarize: data that includes experience in localisation, knowledge of tools, supply chain and post-editing, we singled out the translators that showed more experience in all of the above sections. The translators that declared having more experience in the four areas were TR 3, TR 4, TR 5 and TR 7. The translators with less experience were TR 1, TR 2, TR 6 and TR 8. We took the mean value for each group of translators in relation to the processing speed and number of errors. Table 8 shows these results:

	Processing speed			Number of errors		
Experience	**New**	**MT**	**TM**	**New**	**MT**	**TM**
More	14.13	15.95	13.32	3.25	5.00	8.00
Less	9.60	11.76	10.95	3.50	3.50	8.25

TABLE 8: OVERALL EXPERIENCE VS. PROCESSING SPEED AND NUMBER OF ERRORS

The table shows that experience has a clear effect on the processing speed. The experienced group is faster than the group with less experience. We can see that the faster group is faster when working with MT than with New segments and TM (in this order). The slower group is also faster when working with MT segments than with TM and finally with New segments. The translators with less experience seem to make better use of both translation aids than the ones with more experience. Additionally, we see that the translators with no experience have very similar processing speeds for MT and TM segments (as we claimed in our first hypothesis).

The total number of errors is slightly higher in the experienced group than in the one with little experience, by 1 error. The number of errors in MT is high-

er in the experienced group by a small margin, 1.5 errors when compared to New and TM segments. This could be due to the fact that translators with more experience are more accustomed to MT output and this familiarity prevents them from seeing very visible errors precisely due to this familiarisation.

Final conclusions

Conclusions on productivity
Considering the mean value, the processing speed for post-editing MT segments is higher than that for TM and New segments. And post-editing TM segments, in turn, is faster than translating New segments. The data dispersion is nevertheless quite pronounced, with very high standard deviations and great differences between maximum and minimum values. The standard deviation is higher for processing New segments than for processing MT or TM segments which might indicate that using pre-translated segments slightly standardizes processing speed.

The fastest overall processing time results from translating New segments without any aid, while the translator with slowest processing time took advantage of MT and TM. This low productivity is more pronounced for TM than for MT. If we look at the productivity gains, the translators with lower processing speeds seem to take more advantage of the translation aids than the translators with higher processing speeds. We would need further research to confirm this trend.

The productivity gain, when compared to New segments, for translation aids is between 13 and 25 percent for MT segments, which is higher than the percentage reported by Krings (2001) and lower than the figures reported by Allen (2005) and Guerra (2003), and from 10 to 18 percent for TM segments.
Our first hypothesis is thus not validated in our experiment since MT processing speed appears to be higher when compared to the processing speed in TM fuzzy matches. The correlation between MT and TM is quite close in the groups that processed fewer words per minute. There exists, however, a pronounced difference in the groups that processed more words per minute, where MT ranks higher. The deviation is high, nevertheless, and we cannot draw concrete conclusions as productivity seems to be subject dependant. Krings (2001) also found that in measuring processing speeds, the variance ranged from 1.55 to 8.67 words per minute. Although O'Brien (2006) offers an average processing speed across four subjects without mentioning any deviation values she highlights (2007) that there can be significant individual differences in post-editing processing speed in-line with these findings.

Conclusions on quality
Overall we can say that there are errors in all translators' texts and errors are present in all three categories: New, MT and TM. This seems to be logical, considering that the tool did not allow the translators to go back and revise their work, and that no revision work was done afterwards by a third party.

More than half the amount of total errors, 52 percent, can be found in the TM segments, 27 percent in MT segments and 21 percent in New segments. The high number of errors in TM could be explained by the fact that the text flows more "naturally" and translators do not go back and check the source text, they just focus on the target text, while the MT errors are rather obvious and easier to spot without having to check the source text.

The number of errors in TM is higher than in any other category for all translators. On the other hand, the number of errors in MT is greater than in New segments in five out of eight cases. In two cases, there are more errors in the New than in the MT segments and in one case there is equal number of errors.

Accuracy errors represent the highest number of errors, 44 percent, and they represent the highest value in TM and MT. This seems to indicate that translators do not question the TM or MT proposal and do not check the source text sufficiently to avoid this type of error. Mistranslation errors had the highest value in New segments, but it is very low in MT segments. This could indicate that MT clarifies difficult aspects of the source texts, although more data is needed to explore this trend. Terminology errors are lower in New than in MT and TM segments, indicating that translators tend to accept the proposed terminology in MT and TM without necessarily checking the terms in the glossaries. This might lead to a recommendation that terminological changes or updates be made before starting the translation process or that the translators be instructed to check the glossary often.

The four fastest translators account for 53 errors while the four slowest translators account for 73 errors, which might indicate that the fastest translators tend to make fewer errors and vice-versa, although this is not true for all cases. The reason behind this difference could be that some translators

found the assignment more difficult than others, but at any rate this difference does not indicate an improved quality.

When a coefficient of error is applied, based on an average review speed per minute, to the processing speed, productivity decreases for all segments and in particular for TM segments. This is only applicable to matches from the 80-90 percent category. MT, on the other hand, presents a productivity increase in relation to translating New segments. The increase is higher than 7 percent as was presented in Krings' study (2001), and it seems to be located between 9 and 25 percent. Krings finds that when comparing existing errors in the output with actual errors found after post-editing, the translators are rated at 3.38 (in a range from 1 to 5) covering almost 80 percent of all the errors in MT. In our case the difference in errors between New and MT segments is not very pronounced, but the errors are quite high in TM segments. As far as we know, other research such as O'Brien (2006), Guerra (2003) and Allen (2003 and 2005) does not offer a matrix of final errors and consequently we do not really know how increases in productivity related to the final quality of their samples. O'Brien (2007) mentions the issue of quality and promises to address the topic in a follow-up study. The forthcoming article will be published in the Journal of Specialised Translation (2009).

The pilot study thus indicates that using a TM with 80 to 90 fuzzy matches produces more errors than using MT segments or human translation. The reason behind this could be that translators trust the content that flows naturally without necessarily critically checking accuracy against the source text.

Finally, our second hypothesis is not proven true by the pilot study as our results show that the quality produced by the translators is notably different when they use no aid, MT or TM, although the number of errors found in MT segments is closer to those found in New segments.

Conclusions on translators' experience

If we consider the results obtained we can say that experience has an incidence on the processing speed. Translators with experience perform faster if the average is considered. Similar to the findings by Dragsted (2004) when comparing the processing speed between students and professionals, translators with less experience in our pilot are slower than the ones with more experience.

The data on errors is not conclusive, as the difference between experienced and less experienced translators is none or very small. In the summary data on translators' experience, experienced translators have a higher number of errors in MT and in New segments when compared to the group with less experience. This could be explained by the small number of subjects, or the possibility that translators with more experience grow accustomed to MT type of errors and they do not detect them as easily as a "newcomer" to the field. The translators with less experience have more errors in TM but less in MT and New.

We could say that our third hypothesis is partially proven because translators with greater technical experience do have higher processing speeds in both MT and TM overall. It is important to point out as well that experience does not seem to have an impact on the total number of errors.

There is a strong need to further explore how new technologies are shaping translation processes and how these technologies are affecting productivity, quality and hence pricing. If translators and the translation community as a whole acquire more knowledge about the actual benefits of the tools in real terms, we can be prepared to come into the negotiating arena with the knowledge necessary to reach common ground with translation buyers.

References

Allen, J. 2003. "Post-editing". In Computers and Translation: A Translator's Guide. Harold Somers, ed. Amsterdam & Philadelphia: Benjamins. pp. 297-317.

Allen, J. (2005). "An introduction to using MT software". The Guide from Multilingual Computing & Technology. 69. pp. 8-12

Allen, J. (2005). "What is post-editing?" Translation Automation. 4: 1-5. Available from www.geocities.com/mtpostediting/.[Accessed June 2008].

Bowker, L. (2005). "Productivity vs Quality? A pilot study on the impact of translation memory systems". Localisation Reader 2005-2006: pp. 133-140.

Brunette, L. Gagnon, C. Hine, J. (2005). "The Grevis Project. Revise or Court Calamity". Across Languages and Cultures 6 (1). pp. 29-45

Dragsted, B. (2004). Segmentation in Translation and Translation Memory Systems. PhD Thesis. Copenhagen. Copenhagen Business School.

Gow, Francie. (2003). "Extracting useful information from TM databases." Localisation Reader 2004-2005. pp.41-44.

Guerra Martínez, L. (2003). Human Translation versus Machine Translation and Full Post-Editing of Raw Machine Translation Output. Minor Dissertation. Dublin. Dublin City University.

Krings, H. (2001). Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes. G. S. Koby, ed. Ohio. Kent State University Press.

Language Weaver. 2008. Homepage for the language automation provider. www.languageweaver.com/home.asp. [Accessed June 2008].

LISA. (2008). Homepage of the Localisation Industry Standards Association. www.lisa.org/products/qamodel/. [Accessed June 2008].

O'Brien, S. (2006). "Eye-tracking and Translation Memory Matches" Perspectives: Studies in Translatology. 14 (3). pp. 185-205.

O'Brien, S. (2007). "An Empirical Investigation of Temporal and Technical Post-Editing Effort". Translation and Interpreting Studies (tis). II, I

O'Brien, S. Fiederer, R. (2009). "Quality and Machine Translation: A Realistic Objective?". Journal of Specialised Translation, 11.

Papineni, K. Roukos, S. Ward, T. Zhu, W.J. (2002). "BLEU: A method for automatic evaluation of machine translation". In Proceedings of Association for Computational Linguistic. Philadelphia: 311-318. Also available from http://acl.ldc.upenn.edu/P/P02/P02-1040.pdf. [Accessed June 2008].

Ribas, C. (2007). Translation Memories as vehicles for error propagation. A pilot study. Minor Dissertation. Tarragona. Universitat Rovira i Virgili.

SDL. (2008). Homepage of SDL Trados 2007. www.sdl.com/en/products/products-index/sdl-trados/default.asp. [Accessed June 2008].

A Comparison of Statistical Post-Editing on Chinese and Japanese

Midori Tatsumi **Yanli Sun**
School of Applied Languages and Intercultural Studies
Dublin City University
Dublin 9, Ireland
midori.tatsumi2@mail.dcu.ie yanli.sun2@mail.dcu.ie

Abstract - *This paper analyses both quantitatively and qualitatively the results of a recent Statistical Post-editing (SPE) experiment on English to Chinese and English to Japanese translations. Quantitatively, it compares the number of changes resulting from SPE between the two languages; qualitatively, a linguistic analysis of the changes is conducted. It also investigates the effect of SPE on the fluency and adequacy of the translation as well as the potential impact on human post-editing effort. Our study indicates that, in general SPE results in more improvements than degradations in both languages although the linguistic changes are different between the two languages. In addition, SPE could improve the fluency and adequacy of MT outputs and shorten human post-editing time in both languages.*

Keywords: Statistical Post-Editing, RBMT, SMT, Chinese, Japanese

1 Introduction

None of the Machine Translation (MT) systems that are currently available are good enough to produce error-free outputs, and as Allen & Hogan (2000) point out, MT errors are likely to recur throughout or across documents. Therefore, post-editors are often dispirited by the need to make the same correction over and over again (Isabelle et al 2007: p 255). In order to ease the burden placed on human post-editors, Allen & Hogan (ibid) proposed the development of an automatic post-editing (APE) module that would automatically repair mistakes in raw MT output by utilising the information on the changes that were made during the post-editing process from "parallel tri-text (source texts, MT output, post edited texts)" (Allen & Hogan 2000: p 62). Elming (2006) presented the first results of the use of an APE module to correct the output of a rule-based machine translation (RBMT) system and it was noted that translation quality increased noticeably in terms of BLEU scores (an automatic machine translation evaluation metric) (Papineni et al 2002).

The advent of statistical machine translation (SMT) not only presented an entirely new method of machine translation, but also opened the door to the possibilities of combining two different MT systems to benefit from the advantages of both. Knight & Chander (1994) proposed to use SMT techniques to learn the mapping between a large corpus of "pre-postedited" (ibid, p 779) texts with aligned corresponding post-edited text. Simard et al. (2007a, 2007b) tested and extended this proposal by using a statistical phrase-based MT system to post-edit the output of an RBMT system. The basic mechanism of this kind of system, which is now often referred to as a statistical post-editing (SPE) module, is as follows: an SMT system is trained using a set of raw RBMT output and its corresponding reference text, which is either human post-edited or human translated (training corpora). In this way, SMT learns how to "translate" raw RBMT output to better quality text. Their experiments showed that this SPE module could improve the quality of the RBMT output. However, a detailed analysis of the improvements and degradations of SPE in the previous experiments had not been presented until Dugast et al. (2007) described their experiment on a combination of Systran and SPE. They evaluated, qualitatively, the changes made by SPE modules on the output of Systran, including some linguistic analysis such as improvements, degradations and

equivalent effects. However, as with most of the previous studies, their study was only conducted on European language pairs. Until recently, little research has been done on the effect of SPE on Asian languages such as Chinese and Japanese.

One such instance of this research is an experiment conducted in 2008 by Systran and Symantec (Senellart & Roturier forthcoming) to investigate the potential of SPE when used in combination with Systran, an RBMT system. The general procedure of the SPE process used in the experiment was as follows: Systran 5.05 was used as the RBMT system, and Moses (an open-source toolkit for statistical machine translation) (Dugast et al 2007, Koehn 2004) was used as the SPE tool. All of the training and test resources were provided by Symantec, which included translation memories (TM) and user dictionaries (UD), in the following language pairs: English to French, German, Chinese, and Japanese. Four parallel corpora have been produced for each language: translation by Systran without UD (referred to as *Systran – Raw* for the purposes of this paper), Systran translation with UD (*Systran - Customised*), Systran translation without UD, combined with SPE (*Systran - Raw & SPE*), and Systran with UD, combined with SPE (*Systran – Customised & SPE*).

The current paper analyses and compares, both quantitatively and qualitatively, the Japanese and Chinese output of *Systran – Customized* and *Systran – Customized & SPE*. The experimental setting for these two languages is as follows: Chinese (ZH) TM consisted of 529,822 translation units of English source texts and corresponding target text, while Japanese (JA) TM consisted of 143,742 units. Both TMs were created based on human translation, instead of human post-edited MT output due to insufficient post-edited data. The UD included Symantec-specific user interface terms as well as general terms to which certain target language words had been assigned. The UD contained 8,832 entries for Chinese and 6,363 for Japanese.

A preliminary classification and evaluation of the changes made by SPE on Chinese and Japanese is conducted in Section 2. Sentence level human evaluation results are presented in Section 3. Finally, Section 4 concludes this study and points out future work.

2 Classification and Evaluation of Changes

2.1 Evaluation setup

As mentioned earlier, in this study, the authors (one Japanese and one Chinese) have decided to carry out detailed linguistic comparisons of the results from *Systran - Customised* and *Systran - Customised & SPE*, since *Systran - Customised* is the standard MT translation method currently employed by Symantec, and we are interested in what would happen if the SPE process was added to the current standard operating procedure. The aforementioned experiment, by Senellart & Roturier (ibid), has shown that *Systran - Customised & SPE* outperformed *Systran - Customised* for both Chinese and Japanese in terms of BLEU and GTM (Turian et al 2003) scores. The BLEU score rose by about 6 points and 12 points, and the GTM score by about 10 points and 7 points for Chinese and Japanese respectively. To reveal the detailed linguistic changes that have caused these improvements in performance, the authors randomly selected a sample of 100 translation segments from each of the Chinese and Japanese test sets, and conducted a quantitative and a qualitative evaluation of the results, comparing the source text, Systran output, SPE output, and the reference human translation to see how many and what types of improvements and degradations had been made during the SPE process.

The quantitative evaluation was conducted using evaluation categories defined by the authors based on the Error Classification suggested by Vilar et al. (2006). The classification was modified to make it more suitable for categorising changes rather than errors, and simplified to ensure applicability to both Chinese and Japanese. The changes made during the SPE process were categorised into Words/Phrases Alteration/Deletion/Addition, Forms (Tense or Voice, Formality, and Imperative), Translation of Fixed Expression, Word or Phrase Reordering, and Punctuation. The number of improvements, degradations, and equivalent changes in each category was counted. It was decided to adhere strictly to the reference translation when assessing each change in order to avoid the

subjectivity of the authors and standardise the evaluation process. Therefore, when the translation of a word in either MT output or SPE output did not match with the one in reference translation, it was regarded as an "equivalent change" even if the change made during the SPE process seemed to have improved the quality of the translation.

The qualitative analysis was performed after the quantitative evaluation in an effort to explain some of the most common changes made during the SPE process. Firstly, similarities and differences between the language pairs were identified, and the factors responsible for these similarities and differences were studied.

2.2 Results and discussion

Below is the result of the quantitative evaluation of the changes that resulted in improvements, degradations, and equivalent effects for Chinese (ZH) and Japanese (JA) respectively.

Change Categories		Improvement		Degradation		Equivalent	
		ZH	JA	ZH	JA	ZH	JA
Word/Phrase Alteration	Content Words	137	45	19	40	28	25
	Function Words	38	45	6	9	17	30
Word/Phrase Deletion	Content Words	0	9	0	2	0	1
	Function Words	51	57	4	5	12	16
Word/Phrase Addition	Content Words	4	0	3	2	2	0
	Function Words	12	1	8	2	15	1
Forms	Tense or Voice	6	3	0	0	3	5
	Formality	0	1	1	0	0	0
	Imperative	0	8	0	0	0	2
Fixed Expression		8	0	0	0	0	1
Word/Phrase Reordering		9	1	3	3	0	1
Punctuation		31	47	4	9	0	4
Total		296	217	48	72	77	85

Table 1. Number of Improvements, Degradations and Equivalents in ZH and JA

As can be seen from the table, the number of improvements for Chinese text is noticeably higher than Japanese, and the number of degradations for Japanese is noticeably higher than Chinese. Based on the evaluation that we conducted in this research, it can clearly be seen that the SPE process has had a more positive impact on the Chinese text than on the Japanese text.

Most notably, there have been numerous improvements in the choices of content words/phrases for Chinese, which happened three times more often than in the Japanese text. For Japanese, the changes that were made to content words/phrases have done as much harm as good. Another thing that is worth mentioning is the changes that were made to punctuation, which have had an equally beneficial impact on both Chinese and Japanese.

Based on the quantitative analysis, we find that the most frequently changed categories are similar in Japanese and Chinese, be they improvements or degradations, such as function words/phrases alteration and function words/phrases deletion. On the other hand, there are also great differences between the two languages, for instance, there have been no content word deletions in Chinese while there have been some in Japanese. A detailed investigation on what constitutes those changes and whether the same category contains the same linguistic changes in Japanese and Chinese has also been conducted.

2.2.1 Similar effects of SPE between the two languages

By "similar effects", we mean those categories that share almost the same level of changes after SPE in Japanese and Chinese. SPE had a similar effect on the following categories in Japanese and Chinese: function words/phrases (alteration or deletion) and punctuation.

Improvements in function words/phrases alteration

One of the most prominent similarities observed in this study is found in the changes made by SPE on function words/phrases. The Improvement/Degradation rates for the Alteration of Function Words were 6.3:1 and 5:1 for Chinese and Japanese respectively, and the rates for the Deletion of Function Words/Phrases were 12.7:1 and 11.4:1 respectively.

Some of the function words/phrases alterations have been made in a very similar manner for both Japanese and Chinese. One example of this would be the changes to more appropriate translations for certain prepositions, such as "to" and "about". Another example of common types of function words/phrases alteration is the correction of modal verb translations such as "can" or "must". In Table 2 below, MT output refers to the output of *Systran – Customised* while SPE output refers to the output of *Systran – Customised & SPE* as we mentioned in section 1. Glosses are omitted due to the fact that SPE mostly makes subtle changes by using more appropriate or desired words in the specific context, and the basic meaning often remains the same.

Source	MT output	SPE output
To maintain …	JA: 保守するため…	維持するには…
Reverts to	ZH: 恢 复 对	恢 复 到
must configure	JA: 設定しなければなりません	設定する必要があります
You can …	ZH: 您 能	您 可 以

Table 2. Example of Function Words/Phrases Alteration

However, within the same categories, there have also been some differences in the types of alteration, presumably mostly due to the language differences. A couple of examples for Japanese cases are shown in the table 3. The first one is a stylistic change of character types from Kanji (ideograms) to Hiragana (phonetic characters), which could also be handled by a simple global search and replace operation in any text editor. However, the second one may be a good example of SPE-specific abilities, where the subjective postposition has been changed to one that is more appropriate in the specific context.

Source	MT output	SPE output
(Imperative sentence ending)	JA: して下さい	してください
Messages are deleted	JA:メッセージは削除されます	メッセージが削除されます

Table 3. Example of unique alteration in JA

Specific changes in Chinese include translations for some of the relative pronouns, demonstrative pronouns, and quantifiers being changed to more appropriate ones during the SPE process. For example, the translation of "this" was changed from "此" to "该". Although these two Chinese characters share the same meaning and their back translation is probably the same, the second one is the more commonly used word in the reference translations.

Improvements in function words/phrases deletion

One common improvement as a result of the Deletion of Function Words/Phrases among Chinese and Japanese was the desirable omission of personal pronouns, such as "you" and "they", which are commonly dropped both in Chinese and Japanese. Yoshimi (2001) has suggested a method of eliminating or substituting unwanted pronouns

in English to Japanese machine translation without human intervention using a decision-tree learning method. In his method, however, the corpora for statistical learning must be created by a human for generic purposes, whereas the training corpora in the current research have been compiled automatically from the very specific domain text, and have been proven to be effective. In Table 4, the underlined translation was omitted during SPE.

Source	MT output	SPE output
the actions that you specify for that rule	JA: あなたがその規則のために指定する処理	そのルールに指定する処理
After you configure your	ZH: 在 您 配 置 您 的	配 置

Table 4. Function Words/Phrases Deletion

Other than personal pronouns, there are no notable similarities in the types of deletions made to the two languages. For Japanese, a number of improvements were made by the positive deletion of unnecessary prepositions, such as "for" (ための), and unnecessary sentence endings caused by wrong part-of-speech parsing. For instance, "definition files" was originally translated as a sentence "定義はファイルします" [The definition files (something)], which has been correctly changed to a noun phrase "ファイル定義" [definition file] during SPE. In Chinese, the deletion of unnecessary translations for quantifiers is quite common, for example, the translation of "Provides a more detailed explanation" [MT output: 提-供 一 个 详 细 说 明] is modified in SPE by deleting the translation of "a' [SPE output: 提-供 详 细 说 明].

Improvements in punctuation

Another notable similarity is found in the changes made to punctuation. The Improvement/Degradation rates were 7.5:1 and 5.1:1 for Chinese and Japanese respectively. In the case of Japanese, one of the major reasons for improvement was the successful deletion of unnecessary hyphens that had been inserted during the RBMT process. Another major positive impact was due to the alteration of the type of full stops to ones that are preferred in the specific context of Symantec. In the case of Chinese, one improvement is the deletion of incorrectly generated commas in front of sentences as the first Chinese example in Table 5 shows. Another improvement is the correct alteration of regular commas into special Chinese enumeration commas when separating items constituting a list, see the second Chinese example in Table 5.

Source	MT output	SPE output
MPE provides an option …	JA: オプションを提供 します 。	オプションがあります 、
Control Centre performance may be diminished while the synchronization is in progress.	ZH: ，当 同 步 进 展 中 时…	同 步 处 理….
You can add, edit, copy, delete …	ZH: 您 能 添 加 、编 辑 、复 制 、删除	您 可 以 添 加 、编 辑 、复 制 、删 除

Table 5. Punctuation changes in JA and ZH

2.2.2 Different effects of SPE between the two languages

Difference here refers to the fact that the influence of SPE is not universal for all categories within the two languages; certain categories in one language are influenced much more than those in the other.

Improvements and degradations in content words/phrases

The most notable difference between the results of the two languages might be the changes made to content words and phrases. Caution must be exercised not to overestimate the number of improvements made by the content

words/phrases alteration for Chinese since there have been a large number of repeated identical changes in the analysed Chinese segments. Nevertheless, the Improvement/Degradation rate of 7.2:1 is worth investigating especially considering the significantly low rate of 1.1:1 for Japanese.

For Chinese, there are several kinds of improvement. One improvement is the alteration of nouns, be it general terms or domain-specific terms. Firstly, terms that were originally not translated have been translated after the SPE module, for example "sub domains" which is not translated by RBMT, is translated into "子 域" as in the reference translation; secondly, terms have been translated more appropriately. For example "scanner" whose translation is "扫 描 设 备" by RBMT, was changed to "扫 描 器", the same as the reference translation. The appropriate adaptation of verb translation is another major reason for improvement, for example, "recommend" was translated into "推 荐" while the SPE module changed it into another, more desirable, translation "建 议". More appropriate choices of adverbs and adjectives are the other reasons for the improvement, such as changing the translation of "unchecked" from "未 经 检 查 的" to "未 经 选 中 的" etc.

Most of the alterations of content words/phrases have had a positive effect, however, there are cases where the changes have had a negative effect, for example, "extra" is correctly translated by RBMT as "额 外 的", however, SPE incorrectly changed it into "无 关 的 [unrelated]".

For Japanese, almost the same number of improvements and degradations have occurred. One of the notable reasons for improvement was the correction of the part-of-speech parsing. For example, noun phrases, such as "filtering rules" and "console commands", which had originally been translated by RBMT as sentences, such as "フィルタは弋配します [The filtering rules (something)]" and "コンソールは命じます [The console commands (something)]", were properly converted back to noun phrases in Japanese as a result of the correction of mistranslations of plural nouns as third person verbs by RBMT. Another reason for improvement was the achievement of better collocation. For example, the translations of certain words, such as "grant" or "unwanted" need to be carefully selected depending on their collocations, and some of the incorrectly selected terms in RBMT output were changed to words that were more appropriate in each circumstance. In addition, some of the translations of domain specific terms, such as "rule" and "run" have been found to be translated into more appropriate Japanese words.

However, many of the degradations also resulted from the mistranslation of domain specific terms. For example, the words "document", "store", and "alert", which had originally been translated properly, conforming to the user dictionaries, were incorrectly changed to different terms. Misinterpretation of general terms has also occurred. For example, the word "number", which had correctly been translated as "番号 [sequential number]", was changed to "数 [quantity]". In addition, there has been an instance of case confusion, where a correctly translated instance of "it" was changed to "IT (Information Technology)". Also, some of the correctly translated words have been replaced with different words; for example, "バックアップ弋書(backup document)" changed to "バックアップデータ (backup data)" or "バックアップファイル (backup file)". Such degradations might be attributed to using human translation as the training data of SPE, which may have included more variations in translating the same English words than human post-edited MT text.

Function words/phrases addition

Another major difference was the addition of Function Words/Phrases. Since there are no inflections and derivations in Chinese, which means that Chinese is not a morphologically rich language compared to most European languages, the Chinese language uses additional function words to express tense or voice. For example, "A black dash indicates that it is disabled" is translated as "黑 色 破 折 号 表 明 它 禁 用", the SPE correctly modifies this into "黑 色 线 表 明 它 已 禁 用" by adding a word expressing the tense and voice. Also, some English prepositions should be translated into circumpositions in Chinese, which require an additional character placed after the phrase. For example, "on the Spim tab" is originally translated into "在 Spim 选 项 卡" and later changed into "在 Spim 选 项 卡 上" with the underlined word added to express the full meaning of the preposition "on".

Fixed expressions

Another difference between Chinese and Japanese is that fixed phrases in English have been translated into more appropriate Chinese phrases as in the reference translation. For example, "In general" has been modified to "通 常 情 况 下" from "一 般 情 况 下". This type of change has not happened at all in the Japanese text.

Words/phrases reordering

Word order is one of the most important factors for determining the meaning of a sentence in Chinese. Correct order is vital in adequacy and fluency. There are cases where SPE has corrected some incorrect word order, such as, "These threats are then…" is translated literally as "这 些 威 胁 然 后", while SPE modified it into the correct order as in the reference translation "然 后, 这 些 威 胁" by putting "then" in font of the Chinese sentence. While words/phrases reordering happened frequently, and often with a positive impact in Chinese, it occurred only a few times in the Japanese text and three occurrences resulted in ungrammatical sentences, for example sentences that begin with an adverbial particle.

Form of Imperatives

One of the interesting differences may be the changes made to the sentence pattern, which happened relatively frequently in Japanese but not at all in Chinese. Changes from the polite imperative form [して下さい] to the polite basic form [します] occurred 10 times in the Japanese text, eight of which had positive effects. This change conforms to the sentence pattern commonly preferred in user manuals.

2.2.3 Errors that have not been corrected by SPE

There have been similar errors produced by Systran in both languages outputs that were not corrected by SPE. Firstly, long range word reordering rarely happened, thus when the MT system produced improper sentence structures, mainly due to misplacement of clauses or incorrect parsing of prepositional phrase attachments, they were very rarely corrected. Secondly, intelligibility was rarely improved, even when local changes such as terms and function words were altered. These examples may suggest that it may not be appropriate to expect that sentence level correction can be achieved by SPE processes.

One error produced frequently by Systran in Chinese is the mistranslation of "and" conjunction phrases, which were not corrected during the SPE process, whereas such a mistranslation was rarely observed for Japanese. Secondly, some of the terms that should remain in English have been undesirably translated into Japanese, for example, the translation of "OLE" to "オーレ", which was not observed in the Chinese text and was not corrected in the SPE process. Finally, some of the user interface terms have been unnecessarily translated in the Japanese text and remained incorrect during the SPE process. On the other hand, in Chinese, some of the user interface terms that Systran failed to translate were successfully translated into the correct Chinese terms during the SPE process. This may have been due to differences in the types of user dictionary entries or the training data provided between the two languages; in any case, further investigation may reveal useful information for the effective use of SPE.

3 Sentence Level Evaluation

3.1 Evaluation setup

In addition to the aforementioned evaluation, we also conducted a pilot experiment evaluating the effect of SPE at the sentence level using three criteria: Fluency, Adequacy, and Post-Editing (PE) time. Based on the definition set by the Linguistics Data Consortium (LDC), fluency refers to the well-formedness based on the target language grammar, and adequacy refers to how much of the information and meaning of the original source text has been expressed in the target text (LDC 2005). These two criteria have been widely used (Papineni et al 2002, Turian et al 2003, Callison-Burch et al 2007, Doddington 2002, Owczarzak 2008, Boitet et al 2006), etc.

PE time here refers to the time needed to edit the output in order to raise the quality of the text so that it is appropriate for publishing purposes. PE time is important as PE is one of the major elements of human efforts in MT workflows and therefore reducing the PE time can have a significant impact on optimising the MT workflows.

Four evaluators for each language were recruited by Symantec. All four evaluators are native speakers of Chinese and Japanese respectively, and all are professional translators. Their experience in translation varies from three to twenty-two years, and the average is seven years. For each of the hundred segments, the English source text, RBMT output, and SPE processed text were presented to the evaluators, and they were asked to decide which target text is better than the other in terms of a) Fluency and b) Adequacy, and then decide which text would need less time to post-edit (Less-PE time). They were asked to put their answers in the table columns similar to Table 6 below. For each of the Fluency, Adequacy, and Less-PE columns, they were given three choices, 1 stands for the first output, 2 stands for the second output, while E stands for the equivalent quality for the two outputs. To avoid any bias on the part of the evaluators, the two target texts were presented in a mixed order, that is, for a random half of the hundred segments, RBMT output texts were presented as Output 1, and the SPE processed texts as Output 2, and the other way round for the rest of the segments. The four evaluators conducted the evaluation individually, and had no discussions or any form of information exchange during the evaluation.

Source_EN	Output 1	Output 2	Fluency	Adequacy	Less PE Time
Turns on or off the special meaning of metacharacters	オン/オフ回転メタ文字の特別な意味。	有効または無効にメタ文字の特別な意味します.	1 or 2 or E	1 or 2 or E	1 or 2 or E

Table 6. Human Evaluation Sample with Japanese output

We have used this simple "choice between three" method for the evaluation mainly due to time restrictions. Fluency and adequacy are normally evaluated in a scaled manner; for fluency, for instance, it is common that the evaluators are given five choices: 1: Incomprehensible, 2: Disfluent English, 3: Non-native English, 4: Good English, 5: Flawless English (Callison-Burch et al 2007, Boitet et al 2006), while we have given evaluators only a relative choice between MT, SPE, and Equal. The same is true for PE time. A number of attempts have been made to measure the post-editing effort using different methods (O'Brien 2007, Krings 2001), which have proved that measuring the post-editing effort is not a straightforward task. We are aware that the method we have employed here may put a restriction on supporting our findings.

3.2 Results and discussion

Table 7 shows the average results of the four evaluators for Chinese and Japanese. From the Chinese results, it can be seen that fluency and adequacy are regarded as having been improved during the SPE process in fewer than 40 cases on average, while PE time is thought to be shortened in nearly 50 cases. In contrast, for Japanese, fluency, adequacy, and PE time are considered to have been almost evenly improved in around 60 cases.

Language	Chinese			Japanese		
Criteria	Fluency	Adequacy	Less PE Time	Fluency	Adequacy	Less PE Time
MT	12.75	15.50	15.00	14.50	8.00	9.75
SPE	37.75	38.00	48.25	59.25	61.50	62.50
Equal	49.50	46.50	36.75	26.05	30.50	27.75
Total	100.00	100.00	100.00	100.00	100.00	100.00

Table 7. Average results for each criterion

After aggregating the results, we applied the Kappa coefficient equation (Carletta 1996) to the results to ensure inter-evaluator agreement, which showed another noticeable difference between the two languages. The following table shows the result of Kappa coefficients, which are widely used for determining the level of agreement among multiple evaluators (Callison-Burch et al 2007). According to the definition set by Landis & Koch (1977), 0.0 - 0.20 is regarded as having slight agreement, 0.21 - 0.40 fair agreement, 0.41 - 0.60 moderate agreement, 0.61 -

- 0.80 substantial agreement, and 0.81 - 1.00 almost perfect agreement. Based on this definition, Japanese inter-evaluator agreement is either at the higher level of moderate or the lower level of substantial agreement, while Chinese inter-evaluator agreement is all at the middle level of fair agreement. The high score for Japanese might be explained by the fact that the evaluation was conducted in a simple way by giving only three relative choices, but it does not explain the rather large difference in values between the languages. It must also be noted that low level of agreement might affect the generalisability of the findings.

Evaluation Criteria	Chinese	Japanese
Fluency	0.276	0.598
Adequacy	0.288	0.582
Less PE Time	0.284	0.624

Table 8. Kappa coefficient values for inter-evaluator agreement

3.2.1 Discussion on the Chinese results

The Chinese evaluators vary on their opinions; only 9 segments out of 100 gain unanimous judgements among the four evaluators and 36 gain unanimous judgements among three of the four evaluators. For the rest of the segments, different evaluators share different opinions on which are better in adequacy and fluency as well as which require less PE time.

Overall, the effect of SPE is obvious if we are looking for an incremental improvement in the translation quality. Many more credits were assigned to the SPE output than to the MT output. This may correlate with the improvements in different categories that we analyzed in section 2. Those improvements help to improve the adequacy, fluency of the translation and reduce post-editing time. For example, among the 7 sentences in which the four evaluators agree that SPE output is better in terms of Fluency and Adequacy and needs less PE time, 6 of them have at least one positive content words/phrases alteration.

However, degradation in those changes might also have a negative effect on adequacy and fluency and hence need more post-editing time. For 15% of the sentences, the evaluators think that the original MT output is better in adequacy and fluency and need less post-editing time. For example, for the three sentences which receive unanimous agreement that MT is better in fluency and adequacy and need less PE time, the degradations within them are inappropriate content or function words/phrases addition and inappropriate function words/phrases deletion.

3.2.2 Discussion on the Japanese results

One of the most striking outcomes found in the Japanese results is that, on average, the evaluators have estimated that SPE output should require less PE time in over 60% of the cases. In fact, in 42 segments out of 100, all four evaluators unanimously concluded that fluency, adequacy, and PE time have all been improved during the SPE process. The evaluators' opinions have varied in other cases where SPE time is considered to have been shortened. Therefore, it is not easy to conclude whether fluency improvement or adequacy improvement is likely to result in shorter PE time. In any case, it might be fair to say that, in general, SPE had a considerably positive impact on improving fluency, adequacy, and PE time.

On the other hand, there are eight segments where at least three evaluators have agreed that MT output should require less PE time than SPE output. Having investigated the reasons for this by revisiting the analysis conducted in section 3, it was found that one or more content word alterations had caused degradation during the SPE process in six out of eight cases. The remaining two cases consisted of one case where function word degradation occurred and another case where Words/Phrases Reordering caused degradation. This might suggest that controlling the content word alteration may, to some extent, help prevent adverse effects of SPE.

4 Conclusion and Future Research

In this study, we conducted a detailed investigation into the Chinese and Japanese results of a prior experiment carried out by Systran and Symantec (Senellart & Roturier forthcoming). The first objective of the research was to find out what linguistic changes SPE can and cannot make, and what their consequences are. The second objective was to research the effect of SPE in terms of fluency, adequacy, and reducing the subsequent human PE effort.

One of the notable findings from the linguistic analysis was that the most frequent changes made during the SPE process for both Chinese and Japanese were content words/phrases alterations, function words/phrases alterations, function words/phrases deletions, and punctuation changes. While content word alterations have resulted both in improvements and degradations in both languages, function word alterations, function word deletions, and punctuation changes have mostly resulted in improvements in both languages. It is an interesting finding that Chinese and Japanese share the same categories of changes, although the exact types of changes made within the same category differ partly due to the language differences. It may also be worth pointing out that the changes made during the SPE process are largely limited to the word level, and changes in the sentence structure or reordering of the words or phrases in a long range seemed difficult to achieve with the current SPE system.

Sentence level evaluation was also conducted to shed light on the effect of SPE in terms of fluency, adequacy, and PE time reduction. One of the most important findings from this evaluation is that the evaluators, on average, think the text after the SPE process requires less time for post-editing in around 50% and 60% of the cases in Chinese and Japanese respectively. This may suggest the potential of SPE in reducing the human effort in MT workflows, which could result in productivity gains, although the results are not clear-cut considering the simple form of evaluation method that was applied. Another curious finding is that the results of the sentence level evaluation contradict the results from the evaluation of changes conducted in Section 2. While SPE has a greater positive impact on Chinese than Japanese in the evaluation in Section 2, the sentence level evaluation has contradicted this and a noticeably better effect has been observed on the Japanese text. The result may be different if fluency and adequacy had been evaluated on a scale rather than with the "choice between three" method, and if the post-editing had been precisely timed, rather than subjectively assessed.

There are several limitations to the current study. Firstly, the conditions for two languages are not identical. Using the same RBMT system for both languages does not necessarily mean that the MT output quality and error types for the two languages are the same. By the same token, human translation in two different languages used for training SPE could not be guaranteed to have the same level quality. Moreover, the training and test materials used in this experiment for Japanese and Chinese were not identical. Secondly, the resources were limited. A detailed investigation was only carried out on a hundred sample segments, and only one native speaker of Japanese and Chinese (the authors) respectively worked on the quantitative and qualitative evaluation of the changes made by SPE in each language. In addition, although four evaluators participated in the sentence level evaluation in each language, because we used "a choice between three" method, rather than the scaled evaluation metrics, the data obtained is impressionistic. The evaluation for post-editing time is estimation rather than a strictly measured duration of editing time.

Nevertheless, this work may have revealed a number of possibilities and limitations of current SPE from a linguistic point of view, especially on less investigated languages such as Chinese and Japanese. A similar but larger scale research project could be conducted in the future using larger corpora with identical source text as well as using more finely scaled evaluation metrics for fluency and adequacy, and actual timing of post-editing. Also, comparison of the SPE text with the human post-edited text using some metrics to measure the textual differences, such as Translation Edit Rate (TER) (Snover et al 2006), may provide us with an interesting opportunity for further investigation of the correlation between the changes made during the SPE and their effects on PE effort.

References

Allen, J. & Hogan, C. (2000) Toward the Development of a Postediting Module for Raw Machine Translation Output: a Controlled Language Perspective. In: *Proceedings of The Third International Workshop on Controlled Language Applications (CLAW 2000)*, Seattle, Washington, pp. 62-71.

Boitet, C., Bey, Y., Tomokio, M., Cao, W. & Blanchon, H. (2006) IWSLT-06: Experiments with commercial MT systems and lessons from subjective evaluations. In: *Proceedings of International Workshop on Spoken Language Translation: Evaluation Campaign on Spoken Language Translation [IWSLT 2006]*, Kyoto, Japan, pp. 8-15.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. & Schroeder, J. (2007) (Meta-) Evaluation of Machine Translation. In: *Proceedings of The Second Workshop on Statistical Machine Translation* Association for Computational Linguistics, , pp. 136-158.

Carletta, J. (1996) Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics,* 22, 2, 249-254.

Doddington, G. (2002) Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics. In: *Proceedings of The Second International Conference on Human Language Technology*, San Diego, CA, pp. 138-145.

Dugast, L., Senellart, J. & Koehn, P. (2007) Statistical Post-Editing on SYSTRAN's Rule-Based Translation System. In: *Proceedings of The Second Workshop on Statistical Machine Translation*, Prague, pp. 220-223.

Elming, J. (2006) Transformation-based correction of rule-based MT. In: *Proceedings of EAMT-2006: 11th Annual Conference of the European Association for Machine Translation*, Oslo, Norway, pp. 219-226.

Isabelle, P., Goutte, C. & Simard, M. (2007) Domain adaptation of MT systems through automatic post-editing. In: *Proceedings of MT Summit XI*, Copenhagen, Denmark, pp. 255-261.

J. R. Landis, J. R. & Koch, G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics,* 33, 159-174.

Knight, K. & Chander, I. (1994) Automated Postediting of Documents. In: *Proceedings of 12th National conference of the American Association for Artificial Intelligence (AAAI 1994)*, Seattle, Washington, USA.

Koehn, P. (2004) Statistical Significance Tests for Machine Translation Evaluation. In: *Proceedings of Conference on Empirical Methods in Natural G128 Language Processing (EMNLP)*, pp. 388-395.

Krings, H.P. (2001) *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes.* The Kent State University Press, Kent, Ohio.

LDC (2005) *Linguistic Data Annotation Specification: Assessment of fluency and adequacy in translations.* Report number: Revision 1.5.

O'Brien, S. (2007) An Empirical Investigation of Temporal and Technical Post-Editing Effort. *Translation and Interpreting Studies,* 2, 1, 83-136.

Owczarzak, K. (2008) *A novel dependency-based evaluation metric for machine translation*, PhD edn, DCU.

Papineni, K., Roukos, S., Ward, T. & Zhu, W. (2002) BLEU: a Method for Automatic Evaluation of Machine Translation. In: *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, pp. 311-318.

Senellart, J. & Roturier, J. (forthcoming) Automation of Post-Editing in Localization Workflows. Presented at LISA Forum Europe 2008.

Simard, M., Goutte, C. & Isabelle, P. (2007a) Statistical Phrase-based Post-editing. In: *Proceedings of NAACL-HLT-2007 Human Language Technology: the conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, NY, pp. 508-515.

Simard, M., Ueffing, N., Isabelle, P. & Kuhn, R. (2007b) Rule-based translation with statistical phrase-based post-editing. In: *Proceedings of ACL 2007: The Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, pp. 203-206.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. & Makhoul, J. (2006) A Study of Translation Edit Rate with Targeted Human Annotation. In: *Proceedings of 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA, pp. 223-231.

Turian, J.P., Shen, L. & Melamed, I.D. (2003) Evaluation of Machine Translation and its Evaluation. In: *Proceedings of MT Summit IX*, New Orleans, USA, pp. 386-393.

Vilar, D., Xu, J., D'Haro, L.F. & Ney, H. (2006) Error Analysis of Statistical Machine Translation Output. In: *Proceedings of LREC-2006: Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy, pp. 697.

Yoshimi, T. (2001) Improvement of Translation Quality of Pronouns in an English-to-Japanese MT System. *自然言語処理,* 8, 3, 87-106.

Readability: Examining its usefulness in the field of controlled language

Patrick Cadwell
School of Applied Languages and Intercultural Studies
Dublin City University
Dublin 9, Ireland
patrick.cadwell2@mail.dcu.ie

Abstract

This article is based on a survey of text-user attitudes. The survey aimed to discover: whether the concept of readability has merit in the field of controlled language; and whether readability is increased by applying controlled-language rules to a sample of technical texts. The article attempts to provide much-needed empirical data to a neglected area of controlled-language research, and to examine the concept of readability that appears to be misunderstood, undervalued and misused. In particular, the paper examines issues concerning reader preference, the predictability of readability formulas, and the variables that impact on readability as a whole. Moreover, though the participant samples in the study were too small to be generalised to larger populations, the trends identified here indicate useful directions for future research.

Keywords: *readability; controlled language; reader preference; empirical data.*

1 Introduction

In this article, the results of a survey carried out to examine whether readability has merit in the field of controlled language (CL), and whether it is increased by applying CL rules to texts is described. The paper is based on a Masters-level dissertation completed earlier this year by the author: the full dissertation can be viewed at http://www.localisation.ie /resources/Awards/Theses/Theses.htm.

This article focuses foremost on the theoretical issues that informed the study; only a very brief outline of the methodology is given, while placing focus on how questions raised in the review of the literature were answered by empirical data.

The literature, in this case, is made up of three main groups: Group 1 consists of papers submitted to the various International Workshops on Controlled Language Applications (CLAW), workshops that have been running since 1998; Group 2 is made up of scholarly works on the concept of readability, in particular Klare (1963, 1974, 2000), Flesch (1948), Dale and Chall (1948), Fry (1958), Gunning (1952); Group 3 is comprised of articles from journals that specialise in technical writing, for example Journal of Technical Writing and Communication or ACM Journal of Computer Documentation.

The three groups speak to very different audiences and represent different trends in research into document production: Group 1 sees readability as interesting, but focuses its efforts on matters more directly linked to machine translation; Group 2 defines and helps us understand the concept of readability, but is weak at exemplifying its practical applications; Group 3 tends to be highly critical of readability as a concept and focuses research on other areas of document analysis. In short, the literature is broad and does not generally examine a link between a clear understanding of what readability is and how it can be practically and beneficially applied to the field of CL. This is the gap that my study hoped to fill. In particular, I had three main motivations for carrying out my work:

1.1 Motivation: need for empirical research

Eight years ago, Knops (2000, p.134) called out for more empirical data in the field of CL. He said:

> *Generally speaking, there is an urgent need for facts and figures obtained in experimental situations and real-life production environments and relating to the effects of particular CL standards, rules and rule sets on readability and translatability.*

Since that time, researchers have answered that call and several empirical studies have been published in the field of CL rules. However, these studies have focused largely on more machine-oriented topics, such as translatability, comprehensibility, proof-read-

ing effort, etc. More human-oriented analyses - such as readability, usability, etc. - have been neglected. In fact, since Knops call I have found no major work that outlines how readability and controlled language interact in a contemporary, commercial situation. According to Hayes, Maxwell and Schmandt (1996, pp.84-85), this may be because readability advantages are harder to quantify than those of translatability, comprehensibility, etc. I believe, however, that this may also be because few people have an understanding of how readability is correctly defined; without such an understanding, how can anyone determine the advantages that applying readability guidelines may bring?

1.2 Motivation: assumptions held about readability

Several under-tested and unchallenged assumptions about readability now hold firm. For example, "reducing the complexity of syntactic structures of a text increases its readability" (Spaggiiari, Beaujard and Canneson 2003, p.152). Similarly, Reuther (1998, p.174) claims that:

> *It is a well known and indisputable fact within the CL community that the use of a Controlled Language (CL) in technical documentation leads to quality improvement with respect to readability, consistency and translatability.*

In the literature, it is difficult enough to find a consistent definition for readability, let alone empirical evidence for such strong statements. My study intended to show the ultimate value of readability in the field of CL: regardless of how easy a CL makes a document to translate or comprehend, these benefits will be for nothing if the text is written in a way that causes the reader to discard it.

1.3 Motivation: lack of terminological rigour

In preparation for this study, I came across an abundance of terms used to describe how CL texts can be analysed and evaluated (hereafter referred to as metrics). Metrics in the field of CL include: readability; comprehensibility; translatability; usability; post-editing effort; consistency, legibility; acceptability; accessibility; learnability. Depending on the author, these terms can be treated as sharing many, all or no characteristics. Clearly, there is a need for a more systematic and rigorous definition and treatment of these concepts.

2 Brief outline of the survey

For this study, I was given access to various naturally-occurring technical texts by the software publisher, Symantec. An internal training document was deemed to be the most appropriate for use in the survey. This type of document was chosen as it could be claimed that the Symantec participants would all be expert in, or at least familiar with, its contents. Moreover, it seemed long and varied enough to provide different examples of writing styles and readability.

This training text was divided up into short, similar passages and popular readability formulas were applied to the passages. Over 50 readability formulas have been developed over the years, but, of these, six formulas are particularly influential: Flesch; Gunning Fog; Dale-Chall; Fry Graph; SMOG; and Automated Readability Index (ARI). The formulas used in this study were Flesch Reading Ease, Flesch Kincaid Grade Level and Gunning Fog; they were chosen for their speed and ease of calculation, and for the large number of times that they are cited in works in the field of CL.

Of the short, similar passages, three individual passages were finally selected to be experimented on. One received an extremely unfavourable readability prediction from all three formulas. It will henceforth be referred to as the 'Norton' passage. One received an extremely favourable readability prediction from all three formulas. It will be referred to as the 'Shared' passage. The last passage received a prediction score midway between these two from all the formulas and will be referred to as the 'HijackThis' passage. Collectively these three naturally-occurring passages with varied readability scores were termed 'NCL' passages in the dissertation.

Next, these three passages were controlled for readability. Identifying the rules to apply to control for readability was a long and difficult task, and will be described in great detail in 3.3. Here, it will suffice to say that once the author had edited the passages (termed "CL" passages in the study) so that they conformed to the readability rules, the same readability formulas were applied to them: all three scores improved marginally. However, the relationship between the three passages remained the same (i.e. the most difficult to read remained the most difficult, etc.).

With NCL and CL versions prepared, it was time to carry out the survey. The survey used two samples of respondents: one sample was made up of Symantec staff that had knowledge and practical experience of

the technical domain from which the texts were taken (Symantec); the other sample was made up of participants without such domain knowledge and experience (Control). The two groups, with 12 participants in each group, were largely balanced for English ability, gender, and educational profile.

The survey was done on a test / retest basis. There was a survey in Stage 1 and another similar survey in Stage 2 to examine attitude variance depending on whether a CL or NCL version was read. To counter any inherent bias, participants did not know when they received a CL version: for both groups, a table of random numbers was used to randomly distribute a CL version to half the participants in Stage 1 and half in Stage 2. At both stages, the subjects read the passages and then filled out a questionnaire that examined their opinions as to the readability of the texts.

3 Theoretical overview

The survey outlined in Section 2 set out to discover whether readability has merit in the field of CL, and whether it is increased by applying CL rules to the training texts under experimentation. However, to decide how to move from such broad goals to specific survey questions required a lot of theoretical preparation. This preparation involved examining readability in isolation - how we define, measure, predict and produce readability - and examining readability in relation to other metrics used in the field of CL.

3.1 Defining and measuring readability

Readability is regularly mentioned in the CL literature, but is rarely defined by the authors that use it. Perhaps this is because it is an idea that is prevalent in general language, and authors assume that readers understand the concept as "the ease with which written language can be read with understanding" (Crystal 1992, p.326). For this experiment, a more detailed definition was needed to show how readability differed from other metrics, such as legibility, comprehensibility or clarity. At first, it was necessary to consult some fairly dated sources: theoretical work on readability began in the US in the 1940s when literacy levels of the general population were still low, but when the government needed to disseminate increasingly complex written documents in the medical, legal and financial fields. Key works by influential scholars at the time include Rudolf Flesch (1948), Edgar Dale and Jeanne Chall (1948), and Robert Gunning (1952). Dale and Chall (1949 in DuBay 2004, p.3) provided the detail lacking in more general definitions. For them, readability is:

> *The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting.*

Klare (1977 in Harkins and Plung 1982, p.149) concurs with Dale and Chall in their definition, and states that when we talk of readable writing, "...we mean that the intended readers are able to read it quickly, understand it clearly, and accept it readily (i.e. persevere in reading it)". In other words, it is the combination of these three elements that differentiates readability from the other metrics. Flesch (1948), Dale and Chall (1948), Gunning (1952), Fry (1968) and Klare (1963) note that we make documents readable to help readers understand them better, and to help them avoid making mistakes that they might otherwise have made. Crucially, though, they emphasise that we also make them readable to save the readers time and effort, and to ensure that they do not give up on reading the document.

These definitions tell us how to go about measuring readability. However, they do not explain how to predict whether one text is more readable than another. Nor do they instruct us how to produce readable text.

3.2 Predicting readability: readability formulas

As was shown in 3.1, measuring what makes a document readable involves the detailed analysis of complex concepts such as the reader's understanding, reading speed, and perseverance. However, such complex analysis may not always be possible. Thus, scholars have tried to develop formulas which use variables in the text to predict how difficult that text may be for a particular audience. Over 50 procedures claiming to compute how difficult a text is to read have been devised over the last 80 years. Of these, six are particularly influential: Flesch; Dale-Chall; Fog; Fry Graph; SMOG; and Automated Readability Index (ARI). Others that are often utilized include FORCAST, Lorge and Spache (Klare 1974, p.68).

Formulas do not define or explain readability; they do not point to all the areas of a text that make it readable or comprehensible (Davison and Kantor 1982, pp.189-190). The formulas are merely intended as indices or predictors of how difficult a text is likely to be for an intended reader. To construct a formula,

the researcher assembles large numbers of 'criterion' passages; these are usually texts taken from the US educational system. Most formulas in use today originated in research projects in the US and are based on the study of American-English data. Moreover, the formulas are compiled with a view to making the English spoken in that part of the world more readable. This is a matter not usually considered by those who use these formulas in other geographic and linguistic settings. Readability formulas are constructed for other languages aside from English, but their features were not considered in this study. Nonetheless, comparing and contrasting the formulas used in different languages would surely be an interesting research theme in the future.

Once the 'criterion passages' have been chosen, language variables from these passages - typically word difficulty and sentence length - are selected. The researcher then sees how these vary with the scores that readers have given the passages in terms of reading speed, reader preference, and comprehension, to name the three most common values. If a language variable and the readers' scores correlate closely, the variable is said to be a characteristic of readable writing and is combined statistically into a formula. These results are then further validated with other scores for reliability (Klare 1977 in Harkins and Plung 1982, p.149).

To use a formula, a passage of at least 100 words is selected; such a length is necessary for the statistical regressions used in most formulas to be valid. Then, a count is made of the language variables that have been identified as being characteristic of readability. These counts are entered into the formula, and an overall score for the passage is given. This score will typically be expressed in different ways: some formulas place the score on a graph (Fry); some express the score as the US grade-school level the reader needs to have completed to be able to read the passage (Flesch); some express the score on a simple scale from 0 to 100, with 0 being the most difficult, and 100 being the easiest (Flesch Reading Ease); while others express it as the number of years of formal education a reader needs to be able to read the passage (Fog).

Crystal (1997, p.254) and DuBay (2004, p.54) both emphasise the increasing significance and popularity of readability formulas in the field of educational research. However, other authors criticise the formulas for being unsophisticated and unsuited to use on any other texts than those intended for children in the US school system: this is because the criterion passages on which they are based have been selected and validated with schoolchildren in mind (Hargis 2000, p.105; Giles and Still 2005, p.66).

Now let us look at how readable writing can be produced.

3.3 Producing readability: rules to conform to

This section describes the CL created by the author for this experiment: the rules outlined here are shown in the literature to have a positive impact on readability. These rules will be divided into four major categories: textual / pragmatic; syntactic; grammatical; lexical.

Textual rules

Have no more than one idea per paragraph.
According to Davison and Kantor (1982, pp.189-191), readability must take into account elements that contribute to a coherent and well-formed text. They emphasise that the inference and cognitive load placed on readers should not be too great.

Each paragraph should start with a topic sentence.
Davison and Kantor (1982, pp.196-197) also found that reading time shortened and readability increased if closely relevant context information was placed at the beginning of each paragraph.

Give old information before new (theme-rheme progression).
Both Farrington (1996, p.16) and Reuther (2003, p.128) assert that human readers process texts better when new and complex information is presented slowly, in a logical progression, and without too many new chunks at one time.

Use headings for paragraphs and leave sufficient 'white space'.
Dayananda (1986 in Crystal 1997, p.383) advises writing for the eye as well as the mind: using white space, combined with headings, subheadings, etc., makes the organisation of ideas in the text clearer

Put long lists in bullet points.
Hargis (2000, p.129) concurs with many CL authors in recommending that long lists should be presented in the form of bullet points.

Syntactic rules

In general, Klare (1977 in Harkins and Plung 1982,

pp.150-151) reminds us that correctly-punctuated 'Simple Active Affirmative Declarative' sentences are the most readable. However, to be more specific, as far as syntax is concerned:

Sentences should not exceed 25 words.
O'Brien (2003, p.110) explains that in CL the maximum number of words allowed in a sentence varies from somewhere between 20 and 25 words. Generally, the lower limits are applied for procedural texts, and the higher limits for descriptive texts.

Have variety in the length of sentences within this 25-word limit.
As was shown in 3.1, perseverance is a key pillar of readability. Klare (1977 in Harkins and Plung 1982, pp.150-151) underlines that if each sentence is uniformly similar, the reader will become bored and give up reading.

Have a maximum of two clauses per sentence.
Bram (1978 in Harkins and Plung 1982, p.146) showed that to increase readability, there should only be one or two statements per sentence, with no additional qualifying or explanatory information. In general, difficult texts have a longer more complex structure and impose a greater cognitive load on the reader.

Grammatical rules

Avoid using ambiguous constructs.
Some linguistic constructs - for example the connectors 'like' and 'or', or the 'slash' - are ambiguous and require resolution by the reader. These increase reading time and complexity, and should be avoided (Nyberg and Mitamura 1996, p.80).

Avoid using the passive voice.
Dayananda (1986 in Crystal 1997, p.383) states that the passive voice is less readable than the active voice as it generates greater cognitive load.

Avoid ellipsis and pronominal reference.
Klare (1977 in Harkins and Plung 1982, pp.150-151) states that leaving out parts of sentences and using pronouns - even when the meaning can be understood without the original noun or ellipted item - creates more difficulty for the reader and should be avoided.

Lexical rules

According to Nyberg and Mitamura (1996, p.77), a pre-approved vocabulary that is consistently used by authors is vital to the success of a CL. However, this predefined word list will vary depending on the domain in which the texts are used: the list for writing a school textbook will be very different to the one used in writing an airplane maintenance manual. No scholar has yet found a better multi-purpose lexical rule for readability than 'the simple word should be favoured over the complex'. This, however, is not very instructive. Without a vocabulary list specific to this experiment, only one other lexical rule was identified by the author. This was:

Ensure that all words are spelt correctly.
Hargis (2000, p.129) reminds us that poor spelling can increase processing time, misunderstanding and frustration. Its impact on readability should not be underestimated.

It must be stressed that the above rules should not be accepted without challenge: many are severely criticised in the literature. For example, though bulleting is intuitively held by many to be easier to read, research done by Garrod (1998 in Grover et al. 2000, p.91) contests whether doing so actually works. Similarly, Davison and Kantor (1982, pp.192-195) claim that shortening sentences can just lead to the dilution of logical relations between clauses and sentences, which in turn leads to mistaken inferences being made by the reader. Moreover, Hargis (2000, p.126) asserts that the break-up of sentences not only interferes with understanding in this way, but also produces a choppy, monotonous style that will bore and frustrate the reader. Despite these criticisms, however, the weight of evidence in the literature at present points to the above rules positively impacting on readability.

So far, only the linguistic variables impacting on readability have been dealt with. However, several extra-linguistic variables also have a strong influence on readability and must not be neglected.

3.4 Extra-linguistic variables
Many variables outside the linguistic realm help or hinder readers in understanding, in reading more quickly, and in persevering with their reading. These include: motivation; reading ability; interest in the topic; relevance of the topic; familiarity; prior knowledge; and testing conditions. DuBay (2004, p.39) points out that many experiments in the field of CL do not achieve the expected results because they fail to control for such variables. It is not difficult to create illustrative examples. Imagine the number of readers that neglect to sign a simple form, even

though the instructions to do so are easy to understand and clearly presented: in such a case, it is likely that motivation or interest is lacking. Similarly, we can think of a document that would be completely unreadable or incomprehensible to the average person, but that would be smoothly read and easily comprehended by an expert with prior knowledge of the topic and familiarity with the text type.

Clearly, then, these variables can have an impact on readability. For example, Klare (1977 in Harkins and Plung 1982, p.150) shows that:

> *Someone who is very highly motivated can read very difficult materials, where the mismatch between reading ability and readability is considerable.*

He gives some examples: low-ability readers are able to successfully complete a tax return or decipher a complex medical chart when failure to do so would result in serious negative consequences for these readers. Even though motivation and the other extralinguistic variables are known to be critical, their subjectivity means that most experiments are unable to control for them.

3.5 Other metrics related to readability

An abundance of metrics in the CL literature has led to conceptual confusion. Many of the metrics do not, in fact, deal with monolingual document production, as is the case with readability; they deal instead with the work of translating a text from a source language into a target language. This is not to say that readability is of no interest to the field of translation; I would argue that an understanding of how to produce readable documents in English is vital to any translator working into or out of English. However, the fact that many people in the field confuse essentially monolingual metrics like readability with translation-oriented metrics makes the need for terminological rigour all the more urgent. Some of the main metrics with which readability can be confused are detailed below.

3.5.1 Comprehensibility

As pointed out by Roturier (2006, p.3), comprehensibility should be defined as the ease with which a translation can be understood by its reader. However, in the literature, comprehensibility is often used synonymously with comprehension, understandability and understanding. Is comprehensibility the same as understanding? Few CL works explicitly define the metrics they use in their research and often use terms interchangeably. The conceptual map of the field becomes still more confused when translatability is introduced.

3.5.2 Translatability

This concept is generally taken to be the extent to which a document is amenable to processing by either a human translator or, more often, a machine translation system. However, Reuther (2003) declares readability to be a subset of translatability, while authors like Hargis (2000) reverse this position entirely and see translatability as being just one level of readability.

3.5.3 Usability

Authors like Redish (2000) and Schriver (2000) see usability as a metric which completely excludes the need for readability and comprehension. For them, the aim is to read to do, or to read to carry out a procedure successfully. They see the reader's level of understanding as unimportant, once the document has been 'used' effectively and the desired result has been achieved.

3.5.4 Others

Aside from these major concepts, other authors introduce even more 'similar-yet-different' ways to look at CL texts. For example, Puurtinen (1995, p.230) defines 'acceptability' as the readability and speakability of a text as well as how well a text receiver accepts a translated text as cohesive, coherent and capable of utilization. She defines 'accessibility' as the ease of comprehension due to the style of writing. Furthermore, Hargis (2000, p.123) sees concepts such as 'learnability' and 'doability' as being merely different levels of readability.

Regardless of whether the above definitions are accurate, it is certain that conceptual organisation is required. How might we introduce such terminological rigour to the field?

3.6 Introducing terminological rigour

The various metrics are highly interrelated: as much as some concepts can appear to contradict each other, others can be shown to be highly complementary. For example, O'Brien and Roturier (2007) were able to show that many of the CL rules used in their separate studies had a high impact on both comprehensibility and post-editing effort, suggesting that these concepts complement each other.

Thus, a way of mapping CL metrics is required that accounts for such interrelations, complementarities

and contrasts. This study proposed the use of a Venn diagram to better understand the conceptual map. Perhaps it will never be possible to draw clear distinctions between what is readable, comprehensible, translatable, etc. Rather than looking to make these concepts entirely distinct, it might be more useful to look at where they have their focus. The metrics in 3.5 focus to a greater or lesser degree on the text itself, the reader of the text, and the outputs of the text. These three elements then become the three circles of a Venn diagram (see Figure 1).

To illustrate with examples from the diagram: take 'interest'. Figure 1 illustrates that this metric tells us more about the reader of the text than anything else. In contrast, 'legibility' tells us more about the text than anything else. It is by no means the author's assertion that this is a perfect mapping of the concepts. It is simply intended to convey the opinion that readability can be shown to have a much lesser focus on the reader than comprehension, or that readability can be shown to have a much lesser focus on the results of the text than usability, and so on.

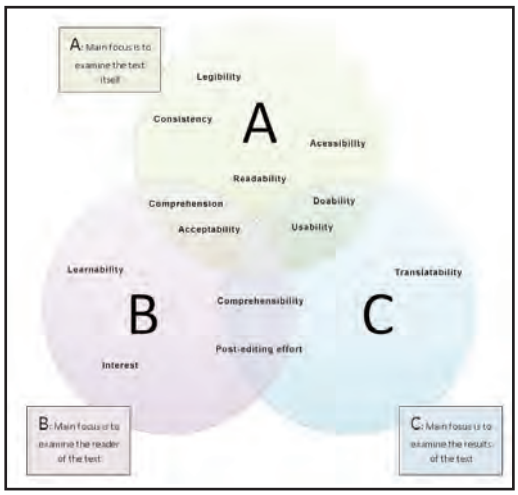

FIGURE 1: VENN DIAGRAM PLOTTING CL METRICS

To summarise, the theory of readability discussed here in Section 2 shows us that the concept is much richer than that which is currently prevalent in the CL literature. When we think of readability, we must consider more than just comprehension and must focus on three key pillars: speed, perseverance and understanding. We must recognise, too, other important variables outside of the linguistic realm that impact on readability. Moreover, it is important to differentiate between the prediction and production of readability when it comes to analysing texts, and especially how this relates to the many popular formulas now in widespread use. With these theoretical issues in mind, let us look at the data produced by this experiment.

4 Summary of the empirical data

The survey carried out for this dissertation produced a large amount of raw data that can be interpreted in many ways. However, the theoretical issues discussed in Section 3 of this article lead us to ask four main questions. Here is how the data seemed to answer these questions:

4.1 Would the CL version be preferred by readers?

At Stage 2, after having read both CL and NCL versions of the text, participants were asked which one they found easier to read. A majority of participants (see Figure 2) in both groups said that they found the CL version of the texts easier to read. To this extent, it can be claimed that the CL versions were preferred by these readers.

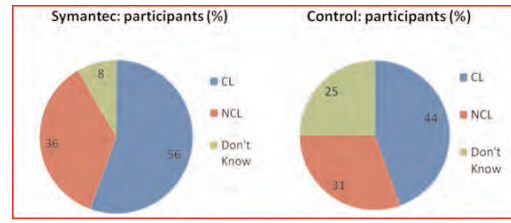

FIGURE 2: WHICH TEXTS WERE EASIER TO READ - WHAT % OF PARTICIPANTS CHOSE WHAT

If a version is preferred by a reader, they are more likely to persevere with reading it than another version. So, on this level, we can say that CL versions appear to be more readable. But what do the questionnaires tell us about the other two pillars of readability?

4.2 Would the other two pillars of readability be altered in the CL version?

According to Klare (1977 in Harkins and Plung 1982, p.149) the key elements of readability - efficiency and understanding - can be tested by analysing reading speed and retention of key vocabulary respectively. The first point we will examine is the retention of key vocabulary. After reading all the passages, participants were asked to identify keywords from a list including synonyms that did not appear in the passages. Figure 3 breaks down the number of correctly retained keywords - less incorrectly selected synonyms (noise) - for each group at each stage.

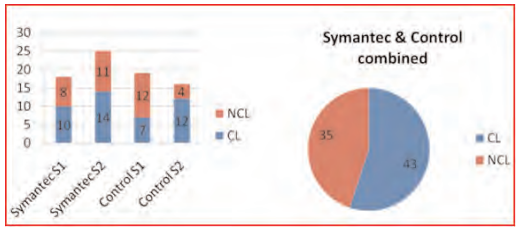

FIGURE 3: NUMBER OF KEYWORDS CORRECTLY RETAINED MINUS NOISE

Overall, this shows that more key vocabulary was correctly retained with less noise when the CL version was read. This result indicates that retention is better when a CL version is used, and that, on this level, the CL texts were more readable. However, as we know from 3.1, readability is about more than just retention and perseverance. The last element to be tested was reading speed.

Figures 4 and 5 show that in the majority of texts, for both groups and at both stages, the CL versions actually took longer to read. This was an unexpected result and was probably due to the fact that the method of timing used was crude and inaccurate; in the study, participants were asked to use a regular wall clock and note starting and finishing times on the questionnaire sheet. This might be sufficient for very long time periods, but in this study, with such short passages, accurate counting of seconds and even milliseconds would have provided much richer data. In this way, the use of eye-tracking software in future work, with its accurate time measurement and complex reading-pattern display, would bring great benefits.

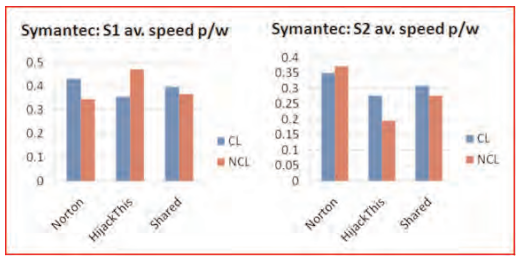

FIGURE 4: BREAKDOWN OF AVERAGE READING SPEED PER WORD PER TEXT FOR SYMANTEC (SECONDS)

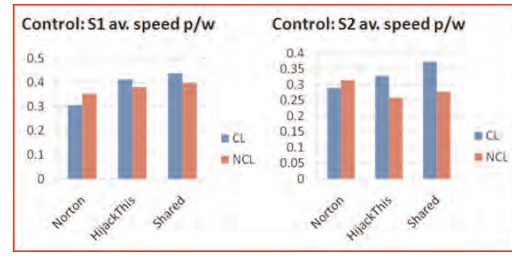

FIGURE 5: BREAKDOWN OF AVERAGE READING SPEED PER WORD PER TEXT FOR CONTROL (SECONDS)

In summary, speed was not positively impacted by applying CL rules, though retention and perseverance were: therefore, it would seem that applying CL rules to technical documents does increase readability.

4.3 Would the formulas' predictions correspond to readers' opinions?

Overall, it seemed that the predictions made by the formulas did not correspond to what readers thought. Figure 6 illustrates that for both groups the majority of readers' ratings did not correctly correspond to the readability formula predictions.

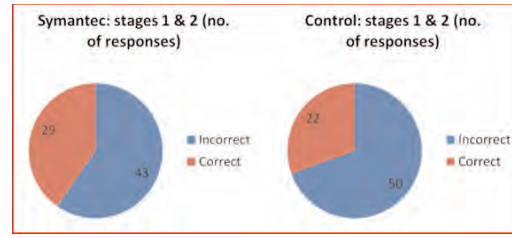

FIGURE 6: NO. OF TIMES READERS' RATINGS CORRECTLY CORRESPONDED TO FORMULA PREDICTIONS

The formulas chosen do not appear to be useful predictive tools. However, there was another group of data that told a different story: participants were asked in the questionnaire to underline any parts of the text that they found difficult to read. By very crudely counting the number of texts in which participants underlined something for reading difficulty, it can be shown that the formulas actually predicted the relationship between the three passages correctly. Thus, in Table 1 we see that in the naturally-occurring NCL versions, 'Norton' had the most responses with underlined sections (17), 'Shared' had the least (11), and 'HijackThis' came in the middle (13). This corresponds exactly to the ranking for difficulty that the formulas predicted in Section 2.

Total no. of texts with underlining for difficulty				
		Symantec	Control	TOTAL
NCL	Norton	11	6	17
	HijackThis	6	7	13
	Shared	5	6	11
CL	Norton	6	5	11
	HijackThis	6	7	13
	Shared	6	7	13

TABLE 1: NO. OF TEXTS WITH UNDERLINING FOR DIFFICULTY

Of course there is a 'black-box' problem with the validity of this data. That is, the underlined sections can only be said to represent areas that participants were less satisfied with; we do not know whether people were truly underlining for problems of readability, comprehension, or some other objection (type-face, legibility, subject matter, etc.). Despite this validity issue, the underlined sections still raise important issues about the way of testing reading difficulty: asking an opinion or rating is subjective and value-laden. Perhaps task-related testing, such as the underlining task, generates more objective data.

4.4 Would extra-linguistic variables impact on readability?

In 3.4, prior knowledge of a domain (or technical expertise) was identified as an extra-linguistic variable that is shown to increase readability in the minds of readers. By this hypothesis, then, the Symantec group should have given more favourable ratings than the Control group. However, this was not the case. Table 2 shows that Symantec underlined more for difficulty than Control, again pointing to the fact that prior knowledge of the domain was not positively impacting on readability.

Symantec: breakdown of underlining (no. of texts)					
	Stage 1		Stage 2		TOTAL
NCL	Norton	6	Norton	5	11
	HijackThis	4	HijackThis	2	6
	Shared	4	Shared	1	5
					22
CL	Norton	1	Norton	5	6
	HijackThis	4	HijackThis	2	6
	Shared	3	Shared	3	6
					18

Control: breakdown of underlining (no. of texts)					
	Stage 1		Stage 2		TOTAL
NCL	Norton	4	Norton	2	6
	HijackThis	4	HijackThis	3	7
	Shared	3	Shared	3	6
					19
CL	Norton	2	Norton	3	5
	HijackThis	5	HijackThis	2	7
	Shared	3	Shared	4	7
					19

TABLE 2: NO. OF TEXTS UNDERLINED FOR DIFFICULTY - SYMANTEC VS. CONTROL

In contrast, the opposite effect was shown for familiarity. Table 3 illustrates that, in both groups, from Stage 1 to Stage 2 underlining decreased, favourability improved, and speed decreased. This would seem to suggest that - because it occurred equally in Control and Symantec - just becoming familiar with a text, even if it was not comprehended or used effectively, makes that text seem more readable.

Familiarity had an impact on readability					
		Symantec		Control	
		Stage 1	Stage 2	Stage 1	Stage 2
Underlining decreased		22	18	21	17
Favourability scores improved		54	49	53	47
Reading speed improved	CL	0.39395	0.31128	0.38513	0.33023
	NCL	0.39483	0.28092	0.37756	0.28336

TABLE 3: FAMILIARITY APPEARED TO HAVE A STRONGLY POSITIVE IMPACT ON READABILITY

The final extra-linguistic variable that was tested by this study was participant profile. In particular, native-English ability was shown to have a strong impact on views of readability. By comparing Figures 7 and 8, it can be seen that non-native speakers found the CL versions easier to read, while native speakers tended to find the NCL versions easier to read. This is not a criticism of the non-native speakers' English ability: most rated themselves 8, 9 or 10, where 10 represented native-level fluency. This difference most likely comes about because native speakers are much more familiar with, and tolerant of, the eccentricities and exceptions of naturally-occurring language.

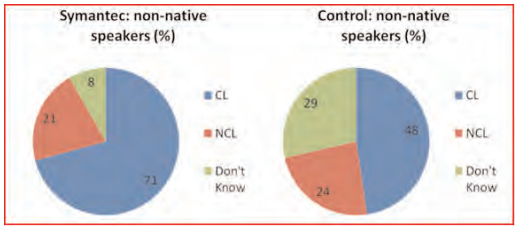

FIGURE 7: NON-NATIVE SPEAKERS FOUND CL MORE READABLE

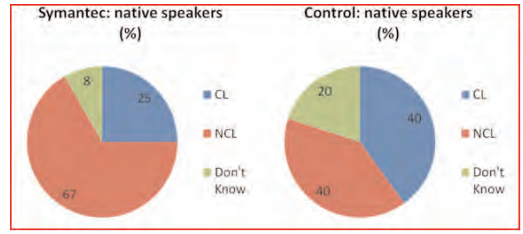

FIGURE 8: NATIVE SPEAKERS FOUND NCL MORE READABLE

5 Conclusions

In conclusion, therefore, this work has provided new empirical data, albeit limited in scope, to show that CL versions are thought to be easier to read; are viewed more favourably; and encourage better retention of keywords. In short, this data seems to suggest that the application of CL rules increases readability. However, these are not the only conclusions that should be drawn from this experiment.

One such additional conclusion concerns readability formulas. Overall, this study appeared to show that the formulas made inaccurate predictions. Most criticisms of formulas probably arise because of people incorrectly using them in ways for which they were not intended: as guidelines for writing or as tools for correction. If the formulas are used within their recognised limitations - as rough predictors of difficulty to prompt more detailed textual analysis - then

they may have more merit.

An important conclusion that needs to be made, too, is that much work remains to be done on clarifying the many metrics that are used in the field of CL. The ideas presented in this study are a tentative first step at disentangling the web. Rather than thinking that one form of analysis is better or worse, any new approach should encourage people just to consider the appropriateness of each metric to their situation.

The final conclusion concerns recommendations for how future studies might proceed. Clearly, more widely-differing texts will produce more noticeable trends in the data: the documents presented in this study were all from the same original document and were selected only based on the predictions of readability formulas. Future studies should incorporate not only formulas, but also semantic assessments, expert advice from the users or authors of the texts, and other criteria to decide the difficulty of the passages to be experimented on. Similarly, such tests should try to incorporate extra-linguistic variables into their methodologies.

Hopefully this article has shown that analysing texts for readability is a useful exercise in the field of CL. By making us consider key elements like reading speed, reader perseverance and reader understanding, as well as influential external factors like motivation and familiarity, the study of readability can promote a comprehensive approach to the theory of document production.

References

Crystal, D. 1992. An encyclopedic dictionary of language and languages. Oxford, UK; Cambridge, Mass., USA: Blackwell.

Crystal, D. 1997. The Cambridge encyclopedia of language. 2nd ed. Cambridge; New York: Cambridge University Press.

Dale, E. and Chall, J.S. 1948. A formula for predicting readability: Instructions. Educational research bulletin. 27 (2). pp.37-54.

Davison, A. and Kantor, R.N. 1982. On the failure of readability formulas to define readable texts. Reading research quarterly. 17 (2), pp.187-209.

DuBay, W.H. 2004. The principles of readability [Online]. Available from: http://www.impact-information.com/impactinfo/readability02.pdf [Accessed 2 September 2008].

Farrington, G. 1996. AECMA Simplified English: an overview of the international aircraft maintenance language. IN: Proceedings of the first controlled language application workshop, March 1996. Leuven: Centre for Computational Linguistics, pp.1-21.

Flesch, R. 1948. A new readability yardstick. Journal of applied psychology. 32 (3), pp.221-233.

Fry, E.B. 1968. A readability formula that saves time. Journal of reading. 11, pp.513-516 and pp.575-578.

Giles, T.D. and Still, B. 2005. A syntactic approach to readability. Journal of technical writing and communication. 35 (1), pp.47-70.

Grover, C., Holt, A., Klein, E. and Moens, M. 2000. Designing a controlled language for interactive model checking. IN: Proceedings of the third international workshop on controlled language applications. April 2000. Seattle: ANLP/NACLA, pp.90-104.

Gunning, R. 1952. The technique of clear writing. New York: McGraw-Hill.

Hargis, G. 2000. Readability and computer documentation. ACM journal of computer documentation. 24 (3), pp.122-131.

Harkins, C. and Plung, D.L. 1982. A guide for writing better technical papers. New York: IEEE Press.

Hayes, P., Maxwell, S. and Schmandt, L. 1996. Controlled English advantages for translated and original English documents. IN: Proceedings of the first controlled language application workshop, March 1996. Leuven: Centre for Computational Linguistics, pp.84-92.

Klare, G.R. 1963. The measurement of readability. Ames, Iowa: Iowa State University Press.

Klare, G.R. 1974. Assessing readability. Reading research quarterly. 10 (1), pp.62-102.

Klare, G.R. 2000. Readable computer documentation. ACM journal of computer documentation. 24 (3), pp.148-168.

Knops, U. 2000. Efficient roll-in and roll-out of controlled language applications. IN: Proceedings of the third international workshop on controlled language

applications. April 2000. Seattle: ANLP/NACLA, pp.134-135.

Nyberg, E.H. and Mitamura, T. 1996. Controlled language and knowledge-based machine translation: principles and practice. IN: Proceedings of the first controlled language application workshop, March 1996. Leuven: Centre for Computational Linguistics, pp.74-83.

O'Brien, S. 2003. Controlling controlled English: an analysis of several controlled language rule sets. IN: European association for machine translation: international workshop, May 2003. Dublin: Dublin City University, pp.105-114.

O'Brien, S. and Roturier, J. 2007. How portable are controlled languages rules: a comparison of two empirical MT studies [Online]. Available from: http://www.mt-archive.info/MTS-2007-OBrien.pdf [Accessed 2 September 2008].

Puurtinen, T. 1995. Linguistic acceptability in translated children's literature. Joensuu: University of Joensuu.

Redish, J. 2000. Usability testing reveals more than readability formulas reveal. ACM journal of computer documentation. 24 (3), pp.132-137.

Reuther, U. 1998. Controlling language in an industrial application. IN: Proceedings of the second international workshop on controlled language applications. May 1998. Pennsylvania: Carnegie Mellon University, pp.174-184.

Reuther, U. 2003. Two in one - can it work? Readability and translatability by means of controlled language. IN: European association for machine translation: international workshop, May 2003. Dublin: Dublin City University, pp.124-132.

Roturier, J. 2006. An investigation into the impact of controlled English rules on the comprehensibility, usefulness and acceptability of machine-translated technical documentation for French and German users. PhD Thesis. Dublin City University.

Schriver, K. 2000. Readability formulas in the new millennium: what's the use? ACM journal of computer documentation. 24 (3), pp.138-140.

Spaggiiari, L., Beaujard, F. and Canneson, E. 2003. A controlled language at airbus. IN: European association for machine translation: international workshop, May 2003. Dublin: Dublin City University, pp.151-159.

Guidelines for Authors

Localisation Focus
The International Journal of Localisation
Deadline for submissions for VOL 8 Issue 1 is 15 June 2009

Localisation Focus -The International Journal of Localisation provides a forum for localisation professionals and researchers to discuss and present their localisation-related work, covering all aspects of this multi-disciplinary field, including software engineering and HCI, tools and technology development, cultural aspects, translation studies, human language technologies (including machine and machine assisted translation), project management, workflow and process automation, education and training, and details of new developments in the localisation industry.

Proposed contributions are peer-reviewed thereby ensuring a high standard of published material.

If you wish to submit an article to Localisation Focus-The international Journal of Localisation, please adhere to these guidelines:

- Citations and references should conform to the University of Limerick guide to the Harvard Referencing Style
- Articles should have a meaningful title
- Articles should have an abstract. The abstract should be a minimum of 120 words and be autonomous and self-explanatory, not requiring reference to the paper itself
- Articles should include keywords listed after the abstract
- Articles should be written in U.K. English. If English is not your native language, it is advisable to have your text checked by a native English speaker before submitting it
- Articles should be submitted in .doc or .rtf format, .pdf format is not acceptable
- Article text requires minimal formatting as all content will be formatted later using DTP software
- Headings should be clearly indicated and numbered as follows: 1. Heading 1 text, 2. Heading 2 text etc.
- Subheadings should be numbered using the decimal system (no more than three levels) as follows:
  Heading
  1.1 Subheading (first level)
  1.1.1 Subheading (second level)
  1.1.1.1 Subheading (third level)
- Images/graphics should be submitted in separate files (at least 300dpi) and not embedded in the text document
- All images/graphics (including tables) should be annotated with a fully descriptive caption
- Captions should be numbered in the sequence they are intended to appear in the article e.g. Figure 1, Figure 2, etc. or Table 1, Table 2, etc.

More detailed guidelines are available on request by emailing LRC@ul.ie or visiting www.localisation.ie

Localisation Focus
The International Journal of Localisation
VOL. 7 Issue 1 (2008)

CONTENTS

Editorial
Reinhard Schäler …… 3

Research articles:

Systematic validation of localisation across all languages
Martin Orsted ……...……..... 4

Productivity and quality in the post-editing of outputs from translation memories and machine translation
Ana Guerberof Arenas ……..... 11

A comparison of statistical post-editing on Chinese and Japanese
Midori Tatsumi & Yanli Sun ……..... 22

Readability: Examining its usefulness in the field of controlled language
Patrick Cadwell ……..... 34