Measurement of gains in computer-assisted ASP-mode[1] translation process within EC-funded applied research project CATI[2]

How to prove the efficiency of Controlled Language principles applied to Xplanation's new-generation Translation Technology and the underlying translation market?

Manu Herbigniaux, Project Manager

Xplanation Language Services
Research Park Haasrode, Technologielaan 21/2, B-3001 Haasrode

Manu.Herbigniaux@Xplanation.com
http://www.xplanation.com/
http://www.hcr.pt/cati/

Preamble

In the month of March 2004, the EC-funded applied research project CATI will come to an end. The CATI project aims at applying controlled language (CL) and automated translation technology to the technical documentation production process of manufacturing and software development companies. CL is being successfully applied in the aeronautics sector, because of the existence of a sector-wide CL standard "AECMA Simplified English". Current checking and translation technology, however, is mostly based on client-server infrastructures, requiring substantial investments, that make it hardly accessible to SMEs.

CATI aims at applying checking and translation technology:

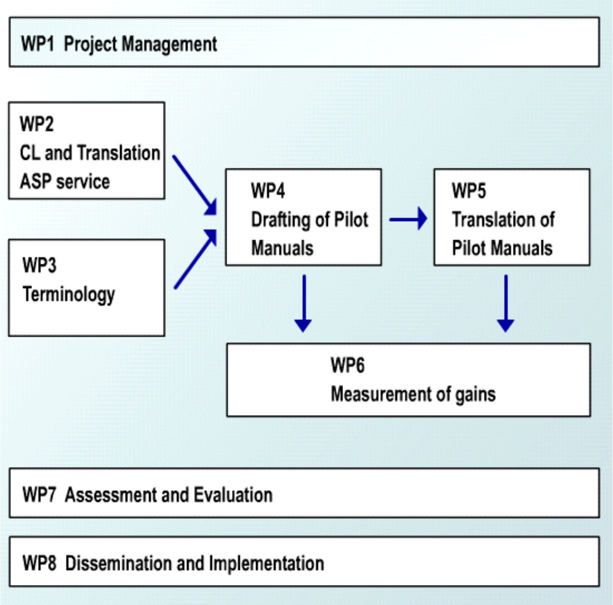
1) using a completely different model, that is ASP based, making the technology available over the web (and avoiding the usage of expensive client-server models);
2) demonstrating the benefits of the technology in other sectors than aeronautics. Three technical sectors have been identified: software development, medical equipment construction and mechanical engineering;
3) setting up a framework to collect objective data to measure the benefits realised through this technology in the documentation production chain;
4) disseminating the technology through the economical and technical evidence collected during the project.

CATI therefore aims at achieving the following project objectives:

- To expose new industrial sectors to language technology
- To develop new and improved services towards SMEs
- To customise the language technology to new industrial sectors
- To set up a framework for the measurements of gains
- To disseminate the results
- To reinforce international links

[1] Application Service Provider
[2] Controlled Authoring and Translation over the Internet

The main project components and their relationship to the project workpackages are depicted in the diagram below.



Introduction

In the present article, we will concentrate on the above mentioned third checking and translation technology layer, i.e. "*setting up a framework for collecting objective data in order to measure the benefits realised through the use of technology in the documentation production chain*".

First, the **operational framework** will be scheduled. Then, a brief description will be given of the different **operational** (authoring and translation) **phases** in the project. Main sections will then be devoted to (i) the different types of **relevant data** that are being collated during the different project phases, including the selection of the most relevant pieces of information towards achievement of the CATI objectives (Measurement of gains, Assessment and Evaluation, Dissemination and Implementation) and (ii) the **analysis** of the selected relevant data. In the conclusion, we hope to be able to convince you that Xplanation has now prepared, through the CATI project, the way for a new approach of the translation market, by **integrating Controlled Language technology into the new-generation translation process**. This will also underline the concrete impact of **EC-funded applied research projects** on the market reality behind computational linguistics and translation technology.

Operational framework

At the beginning of the project, 3 CATI partners[3] selected several technical manuals as the starting point for the CATI research. Those were called the **Plain English docu-ments**. Then, using Xplanation's Controlled Language tool LMiS, the Plain English documents were matched against Controlled Language rules, mainly at grammatical and terminological level. As a result of this check, a new version of the original documents was procuced, which we called the **Controlled English documents**. In parallel, both the Plain English documents and the Controlled English documents were pre-translated, using translation technology such as **Machine Translation** (MT) and **Translation Memory** (TM), through Xplanation's integrated translation environment, Tstream Workflow System. Finally, the pre-translated files were post-edited and assessed by human translators, using Xplanation's specific text editor for translation purposes, Tstream Editor Studio.

Project phases

The CATI project was structured along the following phases:

- Preparatory actions

Selection of original documents, definition of a set of Controlled Language rules, integration of customer- and (sub)domain-specific terminology into the Controlled Language tool, the Machine Translation engine and the terminology lookup window of the Tstream Editor

[3] B-K Medical (medical equipment construction, Denmark), IAITI (software development, Portugal) and LMS (mechanical engineering, Belgium)

Studio (context-sensitive term database), and setup of the initial Translation Memory.

- Phase 1

Translation of the first set of Plain English documents, CL check of the first set of Plain English documents; Translation of the Controlled English documents; Comparison of a.o. the translation quality obtained for CL vs. PE documents.

- Intermediary actions

Selection of new original documents, definition of a new subset of Controlled Language rules, integration of supplementary customer- and (sub)domain-specific terminology into the Controlled Language tool, the Machine Translation engine and the terminology lookup window of the Tstream Editor Studio.

- Phase 2 A

New translation of the initial set of Plain English documents, new CL check of the initial set of Plain English documents; Translation of the new Controlled English documents; Comparison of a.o. the translation quality obtained for new CL vs. PE documents.

- Translation Memory update with the materials processed during Phase 2 A

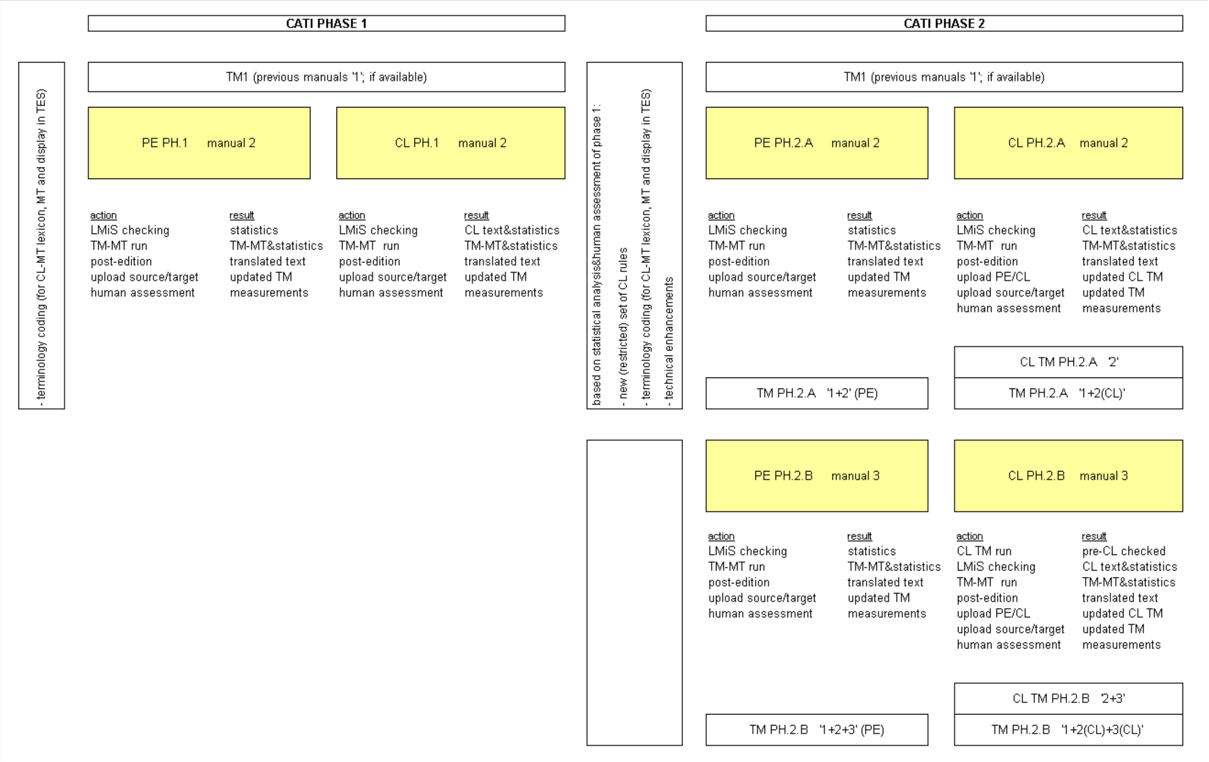
- Phase 2 B

Translation of the second set of Plain English documents, CL check of the second set of Plain English documents; Translation of the resulting Controlled English documents; Comparison of a.o. the translation quality obtained for the second set of CL vs. PE documents

- Final actions

Comparison of the gains obtained in phase 1 (CL vs. PE), vs. those obtained in phase 2A and 2B (CL vs. PE), expecting that linear - or even decreasing - efforts on the side of Controlled Language checking, result in exponential gains w.r.t. translation quality, duration and costs.

Below is a schematic overview of the CATI project phases 1 and 2:



Available data

At each stage of the project, various types of data are being collated, namely:

- Standard translation quality and readability indices, such as **LISA QA Model**, **J2450 Metric** and **Flesh [Kincaid] Indexes** (the latter for English source documents only);

- LMiS statistical data w.r.t. **controlled language** aspects, such as proportion of conformant vs. **non-conformant segments**, proportion of **overruling**[4] for each CL rule, propostion of **passive** sentences, **translatability**, as well as import, machine CL-check, human revision and export **duration**;

- TES statistical data w.r.t. **translation** aspects, such as correlation between **human quality assesssment** and **automatic MT/TM score**, correlation between human quality assesssment and sentence length, correlation between automatic MT/TM score and sentence length, **editing distance/similarity** between human post-edition and MT/TM proposals, correlation between automatic MT/TM score and editing distance/similarity, proportion of MT/TM proposals requiring human post-edition, as well as human processing time;

- Usability indices, such as **System Usability Scale-matrix** (SUS)[5] **and System Usability Measurement Inventory-questionnaire** (SUMI)[6].

The collated data will enable the CATI partners (i) to measure **gains** between processing of the **Plain English** and **Controlled Language** documents, and (ii) to compare the gains measured during phase 1 and phase 2 of the project, i.e. after **enhancement of several coltrolled language and translation technology components**.

Data presentation and analysis

Phase 1 of the CATI project started by 15 November, **whereas Phase 2** started by 1 December. Both phases should be finished by 31 December 2003. **Analysis** of the data, incl. **comparison of the CL-gains** between phase 1 and phase 2, will be performed in the month of Jannuary 2004. The project's **final review** at the European Commission will take place by the end of March.

However, the partial data already collated by 3 December[7] enable us to make several assumptions, regarding the following aspects of the CL and translation process:

For a subset of the sample documents, an average CL **non-conformancy reduction** of approximately 30.0 %, using the full set of available CL rules (64), results in **slight augmentation of the Flesh readability index**, although several CL manuals got a lower readibility score than their original Plain English version. Given the following observations, this might show that **readability** indices are not very reliable in terms of **translatability**;

For the documents already processed so far, the CL-checking and rewriting efforts result in an average **increased usability of the MT proposals** by approximately 32.0 % which, in Xplanation's business model, is very significant;

Interestingly, for a subset of the documents, an average **conformancy increase** of 43.0 % resulted in an **increased Machine Translation score** by 1.5 %, and in a **human quality**

[4] i.e. when a LMiS user decides NOT to correct or adapt a segment which has been marked as non-conformant by a controlled language rule

[5] John Brooke User information Architecture A/D Group Digital Equipment Co. Ltd.

[6] Developed by J. Kirakowski & M. Porteous, University College Cork, Ireland.

[7] Deadline for CLUK paper submission

assessment of the same MT proposals increased by even 9.0 % !
This might in turn prove to be interesting also for the human translators who do the post-edition of the pre-translated documents, since the **editing distance** between the **original proposal** and the **final translation** also tends to **decrease** (to be confirmed), which means less post-editing effort required.

Average non-conformancy reduction of approximately 13.0 %, using a **restricted set of CL rules** (45), took **less than half of the time** spent CL-checking the same documents with the full set of CL rules, although it might result in **comparable translation quality increase and cost savings** (to be confirmed);

Conclusion

At the end of the second phase of the project, we hope to be able to confirm the hypothesis we already have experienced so far, namely that using a **restricted set of highly effective CL rules**, and investing in **terminology mining** preparatory actions is the key to enhanced pre-translation performances, and translation cost reduction.

The phase 1 experiences of the different partners resulted not only in some changes in the measurement set-up. The different ideas on the set-up of the checker and the exploitation got also more focussed. Xplanation started, based on user feedback, with the development of a new version of the LMiS for integration in its translation workflow. This will facilitate ease-of-use, measurability and clearly enlarge the potential market of the checker and still offers potential customers full flexibility and choice, either a full translation offering including controlled language or controlled language only.

Given the impact terminology mining on both controlled language and translation, Xplanation started also a separate terminology mining project.

The final evaluation of the CATI project will of course influence the exploitation plans as well as the commercial contacts on controlled language that Xplanation is starting already.

In its current CL exploitation plan, Xplanation has identified 3 ways of implementing CL components:

1) CL checking that focuses on both readability and lowering the translation costs and is integrated in the translation workflow
2) An automated pre-processing for lowering translation costs
3) A pure CL application mainly focussed on higher readability

In Xplanation's commercial full service translation solution offerings it is very important to have CL as an integrated part of the Tstream translation workflow, which is easy accessible by both customers and sub-contractors. This does not mean that a potential customer can't use controlled language in an ASP configuration only. It does however mean that regardless of what parts of Xplanation's offerings the customer is interested in, it will access and work within the same easy-to-use interface. By integrating the controlled language editor in the Tstream workflow, Xplanation can offer the potential customers full flexibility and choice, either a full translation offering including controlled language or controlled language only.

Another possibility would be to make a much "lighter", more or less fully automatic, version of CL checking to be used as a simpler pre-processing step in the usual translation process. In such an automated tool all the more complex controlled language rules wouldn't be checked and most of the terminology control wouldn't be included. It would simply clean up linguistic "noise" like punctuation, date format, numerals and other simple errors to facilitate MT and TM efficiency. But it wouldn't at all make any major improvements to readability and certainly not give the translation cost-savings a more comprehensive CL checking tool is expected to give.

Still another possibility is to follow the path of the aeronautics controlled language projects mainly focussing on manual instruction readability to be implemented in potential customers' documentation and authoring processes, and not so much focus on lowering translation costs. For a lot of reasons that does not make sense to Xplanation. Improving readability is a very complex task involving user test groups and changing the full documentation strategy of a company. It's also very hard to measure these potential readability improvements. On top of that it comes at a cost that very few, if any, SMEs are willing to accept.

So, the conclusion of the above discussion is that Xplanation will focus on a CL checker that is integrated in the translation workflow. This will facilitate ease-of-use, measurability and the clearly enlarge the potential market for controlled language.

At the end of CATI project phases 1 and 2, we hope to be able to convince you that Xplanation has now prepared, through the CATI project, the way for a new approach of the translation market, by integrating Controlled Language technology into the new-generation translation process. This will also underline the concrete impact of EC-funded applied research projects on the market reality behind computational linguistics and translation technology.

Commented References

We have found relatively few relevant publications related to the measurement of gains resulting from the application of controlled language and translation technology to the professional translation process in technical domains.

As far as we know only 3 experimental studies have been conducted to measure the effects of Controlled English on comprehensibility and translatability. These studies are:

Shubert e.a. (1995)
Shubert, S.K., J.H. Spyridakis, H.K. Holmback and M.B. Coney, The Comprehensibility of Simplified English in Procedures. Journal Technical Writing and Communication, 25(4), 1995, 347-369.

Chervak e.a. (1996)
Chervak, S., C.G. Drury and J.P. Ouellette, Field Evaluation of Simplified English for Aircraft Workcards.

Spyridakis e.a. (1997)
Spyridakis, H., H. Holmback and S.K. Shubert, Measuring the Translatability of Simplified English in Procedural Documents. IEEE Transactions on Professional Communications, 40(1), 1997, 4-12.

In addition to these experimental studies there is a lot of anecdotal evidence in the form of user testimonies (e.g., Kimble 1996-1997).

(Kimble 1996-1997)
Kimble, J., Writing for Dollars, Writing to Please. The Scribes Journal of Legal Writing, 6, 1996-1997.

These testimonies show that Controlled Language helps companies to save money, either directly or indirectly, by:

- Reducing help desk calls
- Reducing customer complaints
- Reducing information retrieval time
- Reducing reading time
- Improving reading accuracy
- Increasing marketing value
- Reducing translation costs
- Reducing translation cycle time

In general, the problem with these testimonies is that they do not specify how the results were obtained and measured. It is therefore unclear whether these can be repeated in different environments and in different circumstances.

Recent publications, also providing interesting references are:

- Sharon O'Brien, Controlling Controlled English - An Analysis of Several Controlled Language Rule Sets;
- Margrethe H. Moller, Grammatical Metaphor, Controlled Language and Machine Translation;

Both are part of the Proceedings of the Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop.

Moreover, mainly for user-oriented issues, the CATI project makes use of information and techniques made available by by VNET5 (http://www.vnet5.org/), where the System Usability Scale-matrix (SUS) and System Usability Measurement Inventory-questionnaire (SUMI) were found.

More information concerning the CATI project is available on the CATI project website (http://www.hcr.pt/cati/), as well as on the the EC Cordis website (http://etip.cordis.lu).