ACCEPT

SEVENTH FRAMEWORK PROGRAMME

THEME ICT-2011.4.2(a)

Language Technologies

ACCEPT

Automated Community Content Editing PorTal

www.accept-project.eu

Starting date of the project: 1 January 2012

Overall duration of the project: 36 months

Seminar Material on Pre-Editing – Edition 1

Workpackage n°6	Name: Community Development
Deliverable n°6.1.1	Name: Seminar Material on Pr e-Editing – Edition 1
Due date: 30 June 2012	Submission date: 30 June 2012
Dissemination level: PU	

Organisation name of lead contractor for this deliverable: Symantec Limited

The research leading to these results has received funding from the European Community's Seventh Framework Programme (*FP7/2007-2013*) under *grant agreement n°288769.*

SEVENTH FRAMEWORK PROGRAMME

Pre-Editing Seminar Brief

Introduction

Our intention is to make the writing process more fruitful for the author, by changing their current practices so that machine translation makes their material accessible to a much larger audience. Many of the authors in online communities have limited formal training in the writing process. The language used by communities is characterised by its colloquial nature and technical focus. Grammar, spelling and style are often sacrificed for speed. In addition, some of the contributors are not native speakers of their chosen forum or are unfamiliar with grammatical rules.

The project intends to provide technology and services to a wide range of users. These users fall into three general classes:

- Third parties as yet unknown to the project,
- European citizens and their information providers,
- The ACCEPT partners themselves.

Third parties would include a range of individuals including but not limited to self-help groups with translation needs, through non-government organisations and charities, to commercial enterprises, who often have large amounts of information available in the source language(s) they use, but cannot readily provide translated materials in other languages.

Ordinary European citizens will act as the end recipients of these services, in that an increased amount of information will be available in a greater range of languages. Information providers in general, including the Commission itself, will have an opportunity to leverage the technology as it develops.

Target Groups, Potential Partners and Other Stakeholders

Symantec is an active member of the Centre for Next Generation Localisation (CNGL) based in Dublin (Ireland). The association with other industry partners in this organization gives us a platform to attract forum members to participate in the ACCEPT project.

We are also building a Special Interest Group (SIG) who will help us test the portal and its component software elements. The membership of the SIG will consist of technically savvy institutions, both commercial and non-profit, who wish to test the technology and deploy it on their own social software stack, as well as smaller groups or companies who opt to use the portal to test the technology. The feedback from these groups will serve to guide the later stages of development of the project. We expect to grow the use of the portal from a few members in the first year to a substantial group of informed and supportive members in the final year.

Challenge of pre-Editing

Pre-Editing Rules

Description

The project will develop pre-editing rules for English and French text that will be processed using Machine Translation. These rules will be based on a standard corpus analysis. The linguistic analysis itself is not within the scope of this work package. This work package deliverable begins by addressing how these complex language writing rules can be presented to our audience of potential authors. In this process, we face two central challenges.

- Rule prioritisation
- Rule presentation

Rule prioritisation

The isolation of the most significant rules for pre-editing in each source language is an ongoing activity and will examine a list of candidates. These rules can be examined for effectiveness. In the first instance we are taking three rules from the lists supplied in Appendix 1, rule_documentation_MT_EN.DOCX and we are composing them as they would be seen in the Symantec forum.

1) Correction of spelling-errors and typographical errors

2) Grammatical error: subject- verb agreements

3) Shorten sentences

These rules are addressed in more detail in the Appendix of the slide deck entitled Pre-editing Briefing for Forum Coordination. The detailed information contained in the Appendix 1 grounds those answers provided by less traditional means in forum interactions.

Rule Presentation

We attempt to present the rules in a format natural for a forum, so as to keep to an absolute minimum any disruption of the practitioner's core activity.

In the accompanying deck we show how the error classes are to be presented to the community members.

Linguistic Information for SMT

Forum feedback is a common practice and is encouraged as it serves to inform following analysis by computational linguists of subsequent rounds of MT adaptation within the ACCEPT project.

Community Dissemination

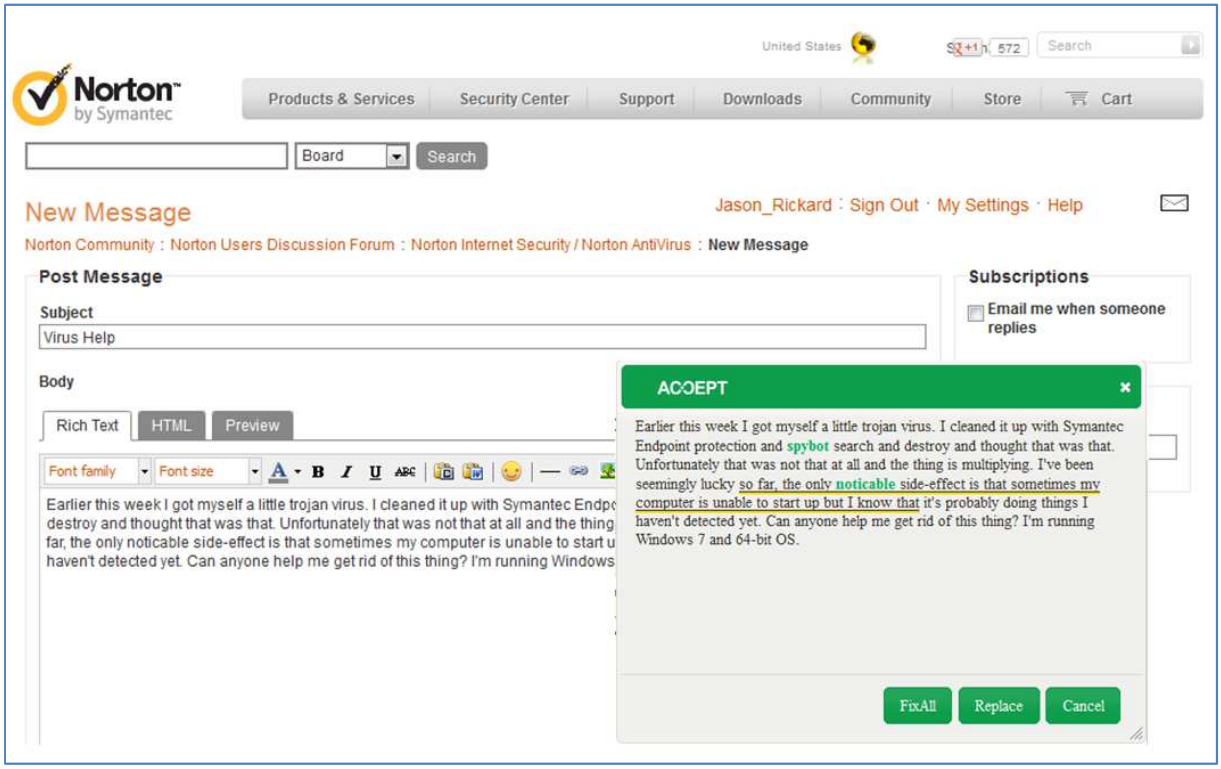
The seminar material will be distributed to the community via the existing discussion platforms, videos, and tips within the ACCEPT portal. It will be presented in a way that is approachable by the average users which will be critical to its use.

Existing Discussion Platforms – The seminar material will be split into smaller segments and posted separately for comment. These will include, but will not be limited to, an introduction to the concept and rules and the benefits of using those rules. The discussion platform for Symantec is the Norton

Forums (community.norton.com); the platform for TWB has not yet been decided. Seminar material will also be available on an ACCEPT Facebook page.

Video – Videos illustrating the seminar material will be made available to the community. These will be made to be approachable as discussed above and will be limited to less than 2 minutes per video to ensure they are viewed and utilized by the community.

Tips – Community feedback to date combined with best practice indicates that the best chance of getting the community to consume help and seminar material is to present it in an approachable way at the time and place it is needed. To meet this requirement, small segments of the seminar material will be presented to the community during authorship via the ACCEPT portal and associated plug-ins.



ACCEPT plug-in