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

Combined Activity and Management Report Half-Period 2

Workpackage n° 1	Name: Coordination
Deliverable n° 1.2	Name: Combined Activity and Management Report Half-Period 2
Due date: 30 June 2013	Submission date: 28 June 2013

Dissemination level: PU

Organisation name of lead contractor for this deliverable: UNIGE

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

Contents

Combined Activity and Management Report Half-Period 2

Objectives of the Deliverable

This report summarizes for M13 to M18 1) project planning and status, 2) people working on the project, 3) use of resources at M18, 3) consortium meetings and 4) risks. It also includes at the end a list of planned/submitted deliverables.

Project Planning and Status

Table 1 gives an overview the on-going tasks (Months 13-18) for WP2 to WP10, with their status. Phase 2 of the project (Month 13-24) is on schedule and does not show any deviation from the DOW.

From M13 to M18, we took the accumulated development of the first 12 months (in particular the baseline, the portals and the pre-editing rules for English and French), exposed it to the communities (Symantec and TSF) and provided comparison with the baseline. In particular, we showed that pre-editing rules have a significant impact on the baseline (Gerlach et al., 2013, for French input) and that pre-editing rules which improve the translation quality also have positive impact on post-editing effort (Gerlach et al, submitted). The first evaluation campaign began in M16 (WP9) and results/data will be available soon. Our dissemination activities are continuing as planned, and include new academic publications and an intensified interaction with the relevant communities.

During the period, we have also worked on strengthening the links between work packages. In WP2/WP4, we compared pre-editing rules with weighted graphs derived from a large-scale pronunciation resource, with weights trained from a small bicorpus of domain language (Bouillon et al., 2013). The different requirements of Lexcelera (TSF) and Symantec induced changes in the portal functionality (WP5), as did the various evaluation schemas. To this end, WP5 has circulated a schedule of development work broken down into monthly sprints, each with its list of mini deliverables tied to specific requirements from each work package. This list was presented before the Paris meeting in May and discussed there to clarify planning up until M24. We have also decided to better synchronize the development of new portal functionalities (WP5) with the demonstration of the technology to the Special Interest Group members (WP10).

Table 1. Tasks (Month 13-18) and status

On-going tasks	Status (Month 18)
Tasks 2.1. *Pre-editing rules for MT (1-18)*	A stable set of pre-edition rules is available both for Symantec and TSF data in English and French. The rules were evaluated with Amazon Mechanical Turk (AMT) judges and showed positive impact on the baseline (deliverable D 2.2). We proved that AMT judgements correlate with translator judgements (Gerlach et al., 2013) and that pre-editing rules which improve quality of the translation also have a positive impact on the post-editing time and the quality of post-edited text (Gerlach et al., submitted).
Task 3.1. *Classification of forum content (10-16)*	We have followed two approaches to the forum content classification: first, we have developed an automated topic extraction and clustering algorithm for forum posts, which we will use in *Task 3.4* to find correlations with the translation quality. Second, we have trained a classifier to rate forum posts by usefulness, which helps to filter out low-quality posts that do not contain information that should be targeted by the ACCEPT translation workflow in the first place.
Task 3.2. *Classification of NGO content (16-22)*	We are currently looking at transferring the topic classification task for forum posts to the documents provided by Lexcelera and Translators Without Borders (TSF). We will adapt the algorithm to the fact that the NGO documents provided cover a much broader range of topics.
Task 4.2. *Domain adaptation methods (4-24)*	We added functionality to Moses to support the weighting of training instances (i.e., training sentences) for domain adaptation, but this obviously leads to the problem of how to find the weights. The first approach we tried was using the modified Moore-Lewis (MML) weights for instance weighting, and we used these in the experiments for our WMT shared task submission. However, the gains over MML filtering (documented in the forthcoming Edinburgh WMT system paper) were quite small. We are now investigating whether a method based on pairwise ranked optimisation (PRO) can be used to train the instance weights. This is a continuation of the work presented by Barry Haddow at NAACL 2013.
Task 4.3. *Linguistic back-off (4-24)*	We have been looking at whether the factored backoff method can help with unknown verb-forms in French-English translation, but find that its lack of source-side context can be a problem. We are continuing to investigate other types of factored models to improve French-English translation. In work recently accepted for the ACL 2013 Workshop on Hybrid Approaches to Translation (Bouillon et al., Two Approaches to Correcting Homophone Confusions in a Hybrid Machine Translation System), we also explored pronunciation-based correction of spelling errors and spelling variation in the input. One of the two approaches constructs confusion networks based on the pronunciation of the words in question, considering all possible spellings for each pronunciation. This allows us not only to “glance over” obvious spelling errors but also to consider some alternatives for verb forms in French (which often are spelled differently but pronounced the same).

Task 4.4. *Exploitation of Usage Data (13-36)*	Post-editing of machine translation output by human translators creates a novel type of data resource: corrected machine translation output. We currently use this data to train confidence measures and explore other uses. One of these uses is to dynamically update the phrase table (i.e., the phrase-level translation database) to include recently translated material, so that recent translations can automatically be considered in translating new material. We expect this to be particularly useful in translating document-specific recurring phrases (e.g., terminology and product names). We have a prototype implementation in place and are currently investigating to what degree implementation-specific variation in computing feature function values affects translation quality.
Task 5.1. *Develop a thin browser-based checking client (1-18)*	A pre-editing prototype has been developed as a JQuery plug-in, and made available on www.accept-portal.eu for testing and download. This plug-in has received some feedback from gurus on the Symantec forum and has been amended based on user feedback. We are currently adding new features with a view to improving the overall user experience.
Task 5.2. *Adapt Translation Environment to Editing Scenarios (1-36)*	The post-editing prototype is available on the portal. It has gone through a very useful round of internal feedback and has been updated (e.g., with an administrative project status functionality and the ability to show the source text when needed). The plug-in is being extensively used in user studies and we are currently investigating how it could be used outside of the portal (e.g., on the Norton forum or AMT).
Task 5.3. *Adapt Evaluation Environment (1-36)*	The ACCEPT Evaluation API has been released and is currently being used to collect user ratings from the German Norton Forum. Some investigations are currently underway to determine how to create custom evaluation tasks (e.g., should a third-party system, such as Appraise, be leveraged?)
Task 6.1. *Build and grow Communities around the Symantec forum (1-36)*	We have disseminated the seminar material by posting within the Symantec forum. The activity to date has been limited to the Guru forum which contains 10 of the most prolific and influential community members. We collected feedback, scrubbed it and delivered it to WP5 for consideration. We have deployed updated training material as part of community engagement plans with the entire community.
Task 6.2. *Build and grow Communities around the TBW/NGO translation activity (1-36)*	At month 18, the TSF (TWB) community consists of more than 100 members. Members communicate through Facebook and LinkedIn. The community now contains a mix of volunteers, both translators and non-translators and both English-speaking and French-speaking. In order to raise the community's profile on the Web, information has also been published through other communities. Some mentors have been designated to lead the community with the support of the Lexcelera team. This last task is part of the rewards program that is now starting.
Task 6.3. *Prepare a seminar on pre-editing rules (17-18)*	We have prepared video tutorials and usage documentation for both the pre-editing portal (browser-based checking client, *Task 5.1*) and the integrated version in the Symantec forums. The feedback collected has been incorporated into our updated tutorials.
Task 6.4. *Prepare a seminar on monolingual editing (17-18)*	The first user studies involving translation of Symantec forum and TSF content have been carried out (*Task 7.1*). Guidelines and tutorials were created to guide the target community. The second Symantec study was launched in M17 and updated training material was made available to the communities.

Task 7.2. *Enriching MT output (1-32)*	We have identified a novel type of help that can be provided to translators and are currently implementing it. This help enables the translator to identify part of the machine translation output and request alternate translations and paraphrases that are more likely to be useful.
Task 8.2. *Comparison of bilingual translation performance with monolingual editing performance (13-36)*	Between month 12 and month 18 a new post-editing session took place, covering bilingual and monolingual post-editing and adding one language compared to the previous experiment, so that we now have both French and English. Participation in this second session was much higher than during the first one. We noticed that in order to maintain community motivation, Lexcelera needs to have subjects participating in the project in a continuing manner, changing parameters and source content on a regular basis.
Task 9.1. *Evaluate the impact of pre-editing rules on SMT (12-18)*	A large-scale evaluation campaign is being conducted to evaluate the impact of the stable pre-editing rule sets defined in WP2. The campaign is in full swing. We expect full manual judgments to be obtained, and results to be reported shortly.
Task 9.2. *Evaluate the impact of post-editing rules on translation quality (12-24)*	This task will begin at Month 20, when post-editing rules will be ready.
Task 9.4. *Compare post-edited RBMT with post-edited SMT using automated metrics (12-30)*	This task will be planned later in the project in the combined deliverable D 9.2.1-D 9.2.2 (Month 24).
Task 10.1. *Dissemination plan (1-36)*	The dissemination plan is a live document and serves to record and spur this activity. Among others, the plan was useful to coordinate the communication with the Special Interest Group (SIG) (see *Task 10.2*) and the publication of two press releases on progress in ACCEPT.
Task 10.2. *Exploitation plan (1-36)*	We have further developed our exploitation plan. We developed the idea of a Special Interest Group (SIG). This was launched at TAUS in Paris on June 1st 2012 and we have received strong interest from a number of high-tech companies as well as social groups. In the first instance we have invited three companies to the SIG with the promise of later addition of several more. Using this phased approach we hope to expose the new technology to a few savvy participants, obtain feedback and iron out any major issues before letting the wider audience have access. In months 12 to 18 we have introduced the SIG membership to the technology through a series of one-to-one telephone meetings. The feedback received so far indicates that ease of installation into prospective partners' own social software stacks might be considered as an additional priority.
Task 10.5. *Exploitation by publication and poster (1-36)*	Our list of dissemination activities by publications and poster continues to outstrip our initial list with a total of 10 publications (in AMTA 2012, CNL 2012, the Seventh Workshop on Statistical Machine translation, TALN 2013, NAACL 2013, the Machine Translation SUMMIT 2013 and the ACL 2013 Second Workshop on Hybrid Approaches to Translation) and a poster (EAMT 2012) (cf. http://www.accept.unige.ch/Products.html).

People Working for the ACCEPT Project

Core personnel were stable during the period. Table 2 shows all the people involved in the ACCEPT project between M13-18, their roles and the work packages (WP) they are contributing to. We also indicate if the project collaborators listed in the table are paid by ACCEPT funding or not and at which percentage. New people joined the project: Chris Cook at UEDIN, Torsten Jachmann at ACROLINX and Marialaura Giova at SYMANTEC.

Table 2. ACCEPT project collaborators between M13 and M18

People	Role	Work Package (WP)	Paid by ACCEPT Funding	%
UNIGE				
Pierrette Bouillon	Coordinator	WP1 (WP leader), WP2, WP9	No	
Manny Rayner	Senior researcher	WP4, WP9	No	
Violeta Seretan	Senior researcher	WP1, WP9 (WP leader)	No	
Victoria Porro Rodríguez	PhD student	WP2	Yes	100%
Johanna Gerlach	PhD student	WP2	No	
Silvia Rodríguez Vázquez	PhD student	WP1	No	
UEDIN				
Philipp Koehn	Senior researcher	WP1, WP7 (WP leader)	Yes	10%
Barry Haddow	Senior researcher	WP1, WP10, WP4 (WP leader)	Yes	35%
Herve Saint-Amand	Senior researcher	WP7	Yes	50%
Kenneth Heafield	Senior researcher	WP4	Yes	30%
Ulrich Germann	Senior researcher	WP4	Yes	30%
Chris Cook	Computing Officer	WP4	Yes	12%
ACROLINX				
Robert Grabowski	Senior software engineer and researcher	WP1, WP2 (WP leader), WP3 (WP leader), WP4, WP5, WP9	Yes	66%
Sabine Lehmann	Chief linguist	WP2, WP3, WP5	Yes	25%
Andrew Bredenkamp	Acrolinx CEO	WP10 (WP leader)	Yes	5%
Katja Höhne	Professional services	WP10	Yes	5%
Ben Gottesman	Senior researcher	WP3	No	17%
Torsten Jachmann	Student	WP2, WP9	Yes	25%
SYMANTEC				
Linda Mitchell	Research assistant	WP5, WP7, WP8	Yes	44%
David Silva	Programmer	WP5	Yes	42%
Fred Hollowood	Project Technical Manager	WP1, WP10	Yes	16%
Jason Rickard	Community Manager	WP6 (WP leader)	Yes	6%
Johann Roturier	Principal Researcher	WP2,WP3, WP4 WP5 (WP leader)	Yes	56%
Robert Leyden	Software Architect	WP5	Yes	3%
Marialaura Giova	Research Assistant	WP6	Yes	5%

LEXCELERA				
Noémie Colin	Computational Linguist	WP8	Yes	50%
Lara El Keilany	Community Manager	WP6	Yes	25%
Laurence Roguet	Operation Manager	WP6, WP8 (WP leader), WP10	Yes	8%
Lori Thicke	Lexcelera CEO	WP10	No	

Use of Resources

Table 3 gives an overview of the PM per partner and per work package at M18. No deviation is shown for the period. All the partners used less than 50% of resources, except Symantec who had to make the different portals available at Month 12 (WP5).

Table 3. Overview of PM per partner and per work package

Partner	WP1	WP2	WP3	WP4	WP5	WP6	WP7	WP8	WP9	WP10	Total	Planned	%
UNIGE	1.5	1.5		4					6		13	38	**34%**
UEDIN	1.2			12.5		1.5	4.9	1.5		0.3	21.9	58	**37%**
ACROLINX	0.5	9.5	3.5	3.5	3.5	0.5			1.5	2.5	25	50	**50%**
SYMANTEC	2.9	2.0	1.6	0.55	18.8	2.6	2.52	3.7	0.02	1.2	35.89	62	**57%**
Lexcelera						3.38		8.19		1.77	13.34	30	**44%**
Total	6.1	13	5.1	20.55	22.3	7.98	7.42	13.39	7.52	5.77	109	238	**45%**
Planned	17	21	21	50	28	28	8	26	21	18	238	238	
% of grand total	**35%**	**61%**	**24%**	**41%**	**79%**	**28%**	**92%**	**51%**	**35%**	**32%**	**45%**		

Managing Consortium Meetings

During the period, we had one face-to-face meeting in Paris, four web meetings and the review meeting in Luxembourg (Table 4).

Table 4. List of project meetings

Date	Project Meeting
February 14-15	ACCEPT review meeting (Luxembourg)
May 24	ACCEPT project meeting (face to face, Paris)
Jan 16	ACCEPT project meeting (web meeting)
Jan 28	ACCEPT project meeting (web meeting)
Feb 8	ACCEPT project meeting (web meeting)
April 28	ACCEPT project meeting (web meeting)

Risk Management

Table 5 summarises the risks and planned solutions.

Table 5. List of risks

RISK	PLANNED SOLUTION
Month 6	
Resourcing constraints in UEDIN in respect of WP7 - Monolingual Post-editing.	Symantec will kick off this work and produce the deliverable D 7.1.1 for M12.
Late approval of the strategy for deliverables D 6.1.1 and D 6.2.1	Given approval on May 11 this work is proceeding at an accelerated pace to comply with the deadline on M6.
Month 12	
Synergies between some WPs could be improved.	The strong interdependence between the WPs is now being supported by a series of bi-partite planning sessions. This action will increase alignment early in the research process.
Month 18	
At month 18 we are focussed on assuring that our WPs are tightly aligned.	We are sharing our plans for the period in detail and using our meetings as an opportunity to address any disjoints in expectations.
We agreed to use post-editing time as gold standard in designing a new metric, but we may not get enough post-editing data.	WP8 will have the community perform bilingual post-editing work. We will also try to recruit post-editors on AMT.

List of Deliverables

Five deliverables were submitted during the period (Table 6). Two deliverables (D 5.4 and D 9.1.2) were cancelled following agreement with the Project Officer.

Table 6. List of deliverables (M13-18)

Del. no.	**Deliverable name**	**WP no.**	**Lead beneficiary**	**Nature**	**Dissemi-nation level**[1]	**Delivery date from Annex I** *(proj month)*	**Actual / Forecast delivery date** *(Dd/mm/yyyy)*	**Status** *(No submitted/ submitted)*	**Contractual** *(Yes/No)*
D 1.2	*Combined Activity and management report - Half period*	1	UNIGE	R	PU	M18	28.06.2013	Submitted	Yes
D 2.2	*Definition of pre-editing rules for English and French*	2	ACROLINX	R	PU	M18	28.06.2013	Submitted	Yes

[1] **PU** = Public
PP = Restricted to other programme participants (including the Commission Services).
RE = Restricted to a group specified by the consortium (including the Commission Services).
CO = Confidential, only for members of the consortium (including the Commission Services).
Make sure that you are using the correct following label when your project has classified deliverables.
EU restricted = Classified with the mention of the classification level restricted "EU Restricted".
EU confidential = Classified with the mention of the classification level confidential "EU Confidential".
EU secret = Classified with the mention of the classification level secret "EU Secret".

D 3.1	*Taxonomy of forum content and rules for automatic classification*	3	ACROLINX	P	PU	M16	30.04.2013	Submitted	Yes
D 5.4	*Browser-based client demonstrator used to access acrolinx IQ server*	5	SYMANTEC	D	PU	M18		CANCELLED – to be merged with deliverables D 5.5 and D 5.6	Yes
D 6.1.3	*Seminar Material on Pre-Editing – Edition 3*	6	SYMANTEC	R	PU	M18	28.06.2013	Submitted	Yes
D 6.2.3	*Seminar Material on Post-Editing – Edition 3*	6	SYMANTEC	R	PU	M18	28.06.2013	Submitted	Yes
D 9.2.1	*Survey of evaluation results – version 1*	9	UNIGE	R	PU	M18		CANCELLED – to be merged with deliverable D 9.2.2	Yes

References

- Pierrette Boullion, Johanna Gerlach, Ulrich German, Barry Haddow, and Manny Rayner.
  Two Approaches to Correcting Homophone Confusions in a Hybrid Machine Translation System.
  In *Proceedings of Second Workshop on Hybrid Approaches to Translation (HyTra)*, Nice, France, August 2013.
- Johanna Gerlach, Victoria Porro, Pierrette Bouillon and Sabine Lehmann.
  La préédition avec des règles peu coûteuses, utile pour la TA statistique des forums ?
  In Proceedings of 20ème conférence du Traitement Automatique du Langage Naturel (TALN), Sables d'Ollone, France, June 2013.
- Johanna Gerlach, Victoria Porro, Pierrette Bouillon and Sabine Lehmann (submitted).
  Combining pre-editing and post-editing to improve SMT of user-generated content.
- Barry Haddow.
  Applying pairwise ranked optimisation to improve the interpolation of translation models.
  In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, Georgia, USA, June 2013.