PROJECT FINAL REPORT

Version 1.1 – 15. December 2014

(public parts)

Grant Agreement number:	**296347**
Project acronym:	**QTLP**
Project title:	**QTLaunchPad**
Funding Scheme:	**FP7**
Period covered:	**1 July 2012–30 June 2014**
Scientific representative:	**Hans Uszkoreit, Scientific Director, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) GmBH**
Tel:	**+49 30 238 95-1811**
Fax:	**+49 30 238 95-1810**
E-mail:	**hans.uszkoreit@dfki.de**
Project website address:	**http://qt21.eu/launchpad**

Contents

1 Executive Summary

QTLaunchPad is a Coordination and Support Action dedicated to identifying the barriers to high-quality machine translation and to preparing a larger action to address these barriers. QTLaunchPad and the planned follow-up action (QT21) have focused on a collaborative approach that brings together various stakeholder groups to work together in pursuit of common goals.

The project has focused on three primary areas:

First, the project has developed **shared, analytic translation quality metrics** that apply to both human and machine translation. Known as the Multidimensional Quality Metrics (MQM), these metrics have built upon international and industry standards to provide a flexible method for developing task-specific measures of quality. The first version of them was adopted, with some additions, as the *localization quality issue type* data category in the Internationalization Tag Set (ITS) 2.0 specification from the World Wide Web Consortium. Further development of MQM resulted in significant changes (while still preserving broad compatibility with ITS 2.0). MQM has now been implemented in various tools (the open-source translate5 tool, a simple "scorecard" tool developed at Brigham Young University in the U.S., and in the commercial tool XTM), and has been utilized by a number of companies organizations (Mozilla foundation, a number of language service providers, and planned for use by the Caribbean Regional Information and Translation Institute). In addition, it has demonstrated its utility in academic research on MT, where the analytic approach it emphasizes has enabled research to identify root causes of some translation problems with greater precision that would otherwise be possible. These metrics were used to develop corpora of error-annotated translations and test suites for machine translation that highlight various problems so that MT developers can see how well their systems perform.

Second, the project has worked on extending **quality estimation** techniques. Quality estimation provides a way for users of MT to reliably identify potential problems in MT without the need for existing reference translations and is thus a major step forward in automatic quality determination. Work within the project focused on the QuEst system, developed at the University of Sheffield, but by working together with the WMT workshop series to host the first quality estimation shared task at ACL 2014 in Baltimore, the project was able to support trials of other quality estimation efforts, by using data annotated with MQM and ranked for subjective evaluation of translation quality. The shared task provided a venue for broad testing of Quality Estimation and the MQM annotation described above. The results of this work are encouraging, and there is considerable interest from both research and industry in the techniques pioneered within QTLaunchPad.

Third, the project as worked on **defining interfaces and specifications for MT-oriented services** along with **plans for sharing resources**. These aspects of the project directly address the currently fragmented environment of research and industry, with far too many valuable results remaining locked up in academic silos and never brought to production. By defining shared interfaces, QTLaunchPad has defined the future of collaborative research and provided valuable resources to developers who want to ensure that others can utilize their results. The results have been incorporated into extensions to the QTLaunchPad META-SHARE node, where they will be widely accessible to the public as tools that can be used in research and industry.

Finally, based on these three areas, QTLaunchPad has laid the ground for QT21, a major effort to improve translation quality for Europe. QT21 will build upon the results of QTLaunchPad to apply them in various industry scenarios and to further develop them in a hybrid collaborative/cooperative research model in order to lower language barriers within the European Union in support of the EU's goals for digital democracy and the single digital market. The plan for this major action developed within QT21 extends aspects of META-NET's Strategic Research Agenda to bring MT development to wide-scale deployment.

2 Project Context and Objectives

There are many stakeholders with an interest in machine translation (MT). They include consumers of information, individual translators and language service providers (LSPs), researchers in MT, creators of content that will be translated via MT, and buyers of MT services (who are often, but not always, content creators). The concerns of these various stakeholder groups vary considerably, and one problem in the development of MT has been that efforts are not aligned. Freelance translators, in particular, have a complex relationship, with widespread concern that MT will take away their jobs or lower quality expectations and drive down the prices they can charge their customers. LSPs and buyers of services are primarily concerned with the potential for MT to lower overall prices while expanding language coverage, speed to market, and maintenance of quality. Information consumers (i.e., people who utilize MT for "inbound" content) tend to focus primarily on price and usability of content, with much lower "quality" concerns than other users.

In this environment of competing concerns, one of the problems in MT research has been that the various stakeholders are not in contact with each other. Translators fear that they have no say in the development of MT, researchers often focus on marginal increases in BLEU scores that do not serve the needs of users, and content creators have no way of knowing whether MT meets their needs or not. QTLaunchPad has sought to overcome these barriers by involving the various stakeholder groups within projects. By working closely with the research community, LSPs (via GALA as a major subcontractor), translators (through FIT, the International Federation of Translators), various content creators (e.g. the Mozilla Foundation, automotive manufacturers), and developers of translation technology (e.g., XTM International, mittag-qi), the project has worked to address their needs in a consistent manner.

Although QTLaunchPad was not a research project, it involved many research aspects. These aspects were targeted primarily at achieving practical results to bring the research and industry stakeholders closer together. For example, work on quality estimation and metrics can help MT to meet the needs of all groups and bring them together.

Another urgent need in the MT community has been for common tools that can meet the needs of the various stakeholder groups. Previously much of the research work on MT has remained as tools oriented at researchers that could not be easily scaled for production work, that were too difficult for general users to work with, or that only worked in specific environments. One of the most important needs has been for tools that can assess quality in ways meaningful to users of MT that allow MT quality to be assessed for previously untranslated texts and for MT and human translation (HT) quality to be compared directly, something currently impossible with reference translation-based methods like BLEU.

Similarly, since much of the work in MT is carried out in academic projects in various environments, users have traditionally faced the problem that various systems use their own interfaces and formats. One of the goals of QTLaunchPad has been to support the development of open-source tools usable to all stakeholder groups.

As a Coordination and Support Action (CSA), QTLaunchPad has focused on bringing industry and research stakeholders together and in laying the ground for future research within this unified framework. As such, it had substantial outreach activities integrated with all work packages and has put a high priority on spreading information about the project's goals and activities to communities that have traditionally not been directly served by EU-funded MT research activities.

3 Main Results

3.1 Quality Barriers and Quality Metrics (WP1)

One of QTLaunchPad's central objectives was finding a common way to assess the quality of both machine translation (MT) and human translation (HT). Until now these two kinds of translation have been assessed in completely different ways: MT quality has been assessed on how similar the output is to reference human translations or on how much effort is required to post-edit the translation into an acceptable output; HT quality, by contrast, has been assessed based on a counting of "errors" in the target text, often calculated as a quality score from 0 to 100. The differences in these approaches had led to a situation in which no meaningful comparison of MT and HT quality was possible.

In addition, MT assessment methods have generally relied on the existence of one or more reference translations, meaning that quality can only be assessed for texts that have already been translated. As a result, MT quality cannot be assessed in production environments—which focus on previously untranslated text—using these measures. Since automatic measures such as BLEU and METEOR cannot be used for production and provide no indication concerning the suitability of a translated text for any particular purpose, organizations have no way to assess the suitability of MT for their business needs. In addition, because reference-based measures depend on the specific reference translation(s) used, they do not provide an absolute measure of quality, but rather a relative one, meaning that scores cannot even be compared with each other across projects.

As a result of these limitations, one of the key activities of QTLaunchPad was development of a system for building translation quality metrics that would apply to both HT and MT. This effort was carried out initially through a survey of major methods for assessing translation quality, with a focus on both human and automatic measures. Very early on it was decided that an *analytic* approach was needed, one in which specific errors could be identified and categorized. Such an approach was needed for research to understand the barriers that impact translation quality: without an analytic approach of some sort it is impossible to identify specific conditions that impact translation quality.

The results of this survey were the first version of the Multidimensional Quality Metrics (MQM), which served as the basis for the localization quality data categories in the Internationalization Tag Set (ITS) 2.0.[1] The unanticipated opportunity to incorporate this early version of MQM into an internationally recognized standard required that development of MQM be accelerated and that certain planned features were not included in the first version, but resulted in incorporation of this project result into an internationally recognized standard. This version featured 26 categories (ITS 2.0 added one additional category) in a flat list organized with the principle that the *first* applicable issue be used if multiple issues apply. This principle worked because the specification has an implicit hierarchy in which more specific issues are listed before less specific issues. MQM version 1 operated, however, at a relatively high level of granularity, with only the most common issue types included with relatively little differentiation between subtypes.

• terminology	• grammar	• misspelling	• whitespace
• mistranslation	• legal	• typographical	• internationalization
• omission	• register	• formatting	• length
• untranslated	• locale-specific-content	• inconsistent-entities	• *non-conformance (ITS 2.0 addition)*
• addition	• locale-violation	• numbers	
• duplication	• style	• markup	• uncategorized
• inconsistency	• characters	• pattern-problem	• other

Table 1. Issue types from MQM version 1

[1] See http://www.w3.org/TR/its20/#lqissue and http://www.w3.org/TR/its20/#lqissue-typevalues.

MQM v. 1 (as represented in the *localization quality issue type* category in ITS 2.0) has now been incorporated in a number of systems that implement ITS 2.0, both commercial and open source.

MQM version 2 extended version 1 considerably. Among the notable changes:

- MQM version 2 incorporates an explicit hierarchy with varying levels of granularity possible. For example, an issue can be categorized as *Grammar* or as a subtype (such as *Word order*) depending on the nature of the issue and the particular issues being considered.
- It adds the notion of task-specific issue type selections. Rather than representing a master list of categories to be used in all circumstances, it provides a *common vocabulary* of issue types from which custom selections can be made
- It incorporates an explicit definition of translation quality that draws on international standards (ISO/TS-11669 and ASTM F2575) to help define appropriate metrics.
- It provides a default scoring system.
- It provides an extension mechanism that allows MQM to be adapted for unanticipated needs (MQM 1 required that all issues be categorized as one of the issue types listed in Table 1).

This system proved to be quite flexible and was used as the basis for the Translation Quality Error Corpus (available as "Round 1" at http://qt21.eu/deliverables/annotations/).[2] This corpus was created in close collaboration with GALA and working with LSPs who used the translate5 environment to annotate segments from the leading statistical, rule-based, and hybrid systems from WMT 2012 (research data) and from customer data (mostly SMT, but also some RbMT) according to a project-specific subset of MQM (Figure 1) in English↔German and English↔Spanish. The specific subset was selected to provide analytic insight into the specific issues seen in MT and proved to be highly successful in providing a clear picture of the differences between the various types of MT systems and the four language pairs (ultimately and unexpectedly, the kind of text annotated was found to be an insignificant variable), as described in D1.2.2[3] and D1.3.1[4].

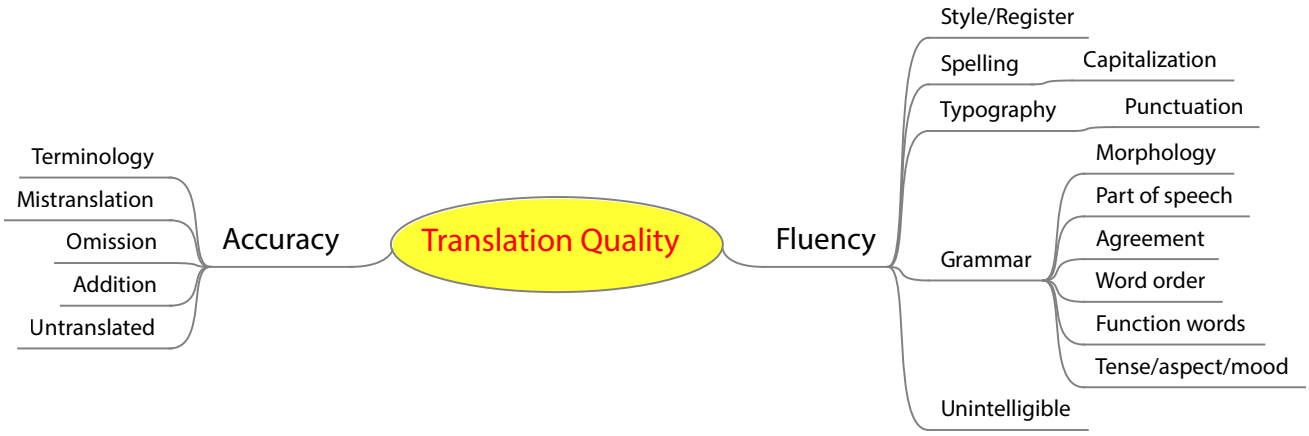

Figure 1. Initial MQM issue type selection used in annotation tasks.

Based on feedback from the annotators and from analysis of the resulting corpora (D1.2.2), the project team discovered that additional changes needed to be made to MQM, both to the core

[2] http://metashare.dfki.de/repository/browse/qtlaunchpad-mqm-annotated-corpora/11f7bcb62f7811e4944d003048d082a428d0e56f4fc849d8a76f8faaf138654a/
[3] http://www.qt21.eu/launchpad/system/files/deliverables/QTLP_deliverable_1_2_2_0.pdf
[4] http://www.qt21.eu/launchpad/system/files/deliverables/QTLP-Deliverable-1_3_1-v2.0.pdf

structure and to the subset used in the annotation tasks. The changes were done largely to exclude issue types that were found to occur infrequently in the context of the annotation tasks and to simplify selection of issue type in cases that had proved to be difficult for the annotators (such as distinguishing between *Mistranslation* and *Terminology*). The updated selection (Figure 2) proved to be more reliable and useful for analytic purposes.

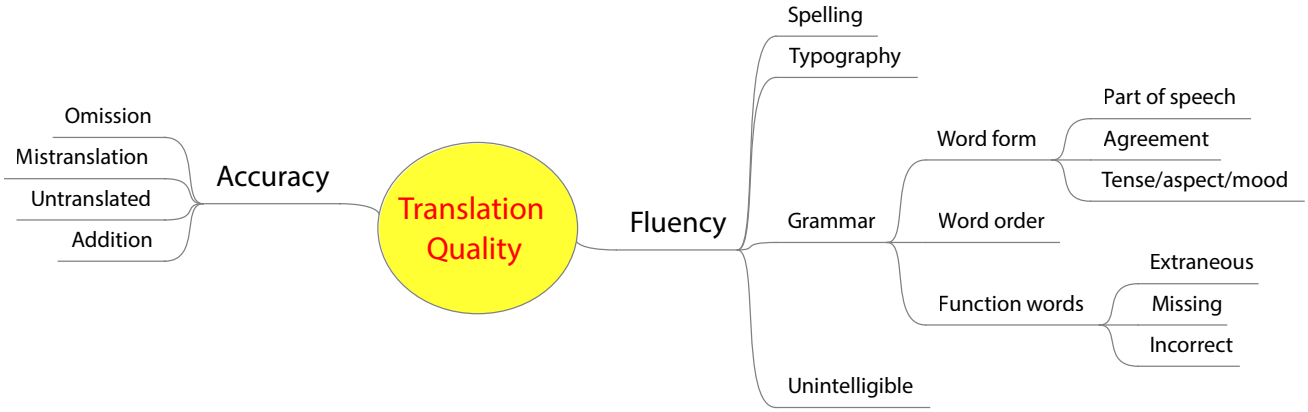

Figure 2. Updated issue type selection used in later rounds of annotation.

Based on the annotation of "real" (WMT and customer) data and of translations produced from sentences included in the TSNLP grammar-testing suite, the project was able to produce two translation test suites[5] (for English→German and German→English) that exemplify common problems and difficult scenarios for MT (D1.4.1). The intention of these test suites is to provide a set of segments that system developers can translate in order to determine if their systems can correctly address the phenomena contained in the examples. Based on analysis of the test suite and of the corpora, the project was able to identify a number of common grammatical and semantic phenomena that are particularly likely to cause difficulty for MT, such as use of "*-ing*" forms and non-genitive *of* in English, omission of subject pronouns in Spanish, and long-distance separation of verbs from subjects in German. While many of these phenomena are well-known from anecdotal evidence, the analysis enabled through use of MQM provided clearer insight into the distribution of these problems and their relative importance.

The resources developed in QTLaunchPad address the needs of both developers and users of MT services and help to align their concerns in a positive fashion.

3.1.1 Relationship between MQM and TAUS Dynamic Quality Framework (DQF)

One point of concern for many individuals in the translation industry and research has been the relationship between MQM and the Dynamic Quality Framework (DQF), a framework developed for quality evaluation by the Translation Automation User Society (TAUS). Both MQM and DQF claim to offer a dynamic, task-specific approach to evaluating translation quality, but are structured very differently, in line with their different purposes, and are largely complementary. While MQM focuses on a generalizable framework for describing and constructing translation quality metrics in a shared vocabulary, DQF provides ready-made translation quality metrics for a variety of common translation scenarios. Many of the DQF metrics use methods quite different from MQM that could not provide the research insights available from MQM (note, however, the TAUS does not claim that DQF is a research tool, but rather that it is suitable for translation production environments).

---

[5] http://www.qt21.eu/deliverables/test-suite/, also at http://metashare.dfki.de/repository/browse/qtlaunchpad-mt-test-suite/857ea5822f7d11e492fa003048d082a4581e7cf1f30c4b6f854c08913c0dd64f/.

MQM and DQF do overlap in one particular area however: DQF has an error typology that is quite similar in some respects to the MQM error typology. As a result there is a need to ensure compatibility between MQM and DQF. Pursuant to this end, representatives from DFKI met with TAUS on 8.September 2014 to address a process to harmonize MQM and relevant portions of DQF. The meeting resulted in the following action items:

- DQF will adopt the MQM definition of quality and incorporate it into a forthcoming release of DQF.
- TAUS and DFKI will work on a common glossary of quality-related terms and promote it.
- Issues around domain/text type (MQM) and industry/content type (DQF) will need to be revisited to see how to reconcile them.
- In the context of forthcoming projects that will further develop MQM, TAUS and the future project will release a joint statement outlining the mapping and relationship of the two projects. This statement and an accompanying report will focus on a more in-depth comparison of the methods/models for each framework and how they are similar and different, emphasizing the importance to both frameworks of translation specifications (based on the ISO/TS-11669:2012 ASTM International F2575:2014 specifications).
- The DQF error typology will be put in a mind map format and differences between MQM and DQF resolved. (In most cases resolution will be simple, but in some other cases DQF may have "extra" features where it includes process-oriented features out of scope for MQM. Such features will be clearly identified in the joint statement referred to above.)
- MQM will add a fourth severity level ("neutral") for cases when it is useful to identify an issue but the severity is unknown or where it is not useful to indicate it.
- The two frameworks will agree on a default set of issue severity multipliers. Although MQM defined the default multipliers as minor = 1, major = 5, critical = 10, based on the LISA QA Model, subsequent consultation with evaluation experts has led to the recommendation that the default weights become 1, 10, and 100 respectively. As this change is a major one, it will require considerable discussion.

The harmonization effort will also address issues around the creation of customized DQF metrics and how to ensure that they are compatible with MQM principles.

This harmonization is anticipated to take place in the first half of 2015 within the QT21 and CRACKER projects currently in grant agreement preparations.

3.1.2 Implementing MQM in tools

As the core of MQM (and the only *required* aspect of MQM) is the list of issue types, tools are free to implement MQM as they see fit, provided they adhere to the MQM issue types. Other aspects, like the default scoring module and XML formats, are optional. Parties interested in adopting MQM should consult the MQM definition (http://qt21.eu/mqm-definition/) for guidance on using MQM. At present there is no default MQM API or other pre-built libraries for implementing it, although such may be developed in the future. However, a more robust API for translate5 that includes support for its MQM tagging functionality is planned for development and promotion as part of the work envisioned for the QT21 and CRACKER projects. (Note that one commercial product, XTM, has already implemented MQM based on the MQM definition, showing that it is possible to implement it based on the information contained in it.)

3.2 Quality estimation & test of methodology and infrastructure in a shared task (WP 2 & 5)

Machine Translation (MT) systems have been increasingly adopted in recent years for different purposes, including gisting and aiding humans to produce professional quality translations (i.e., post-editing-based workflows). Since the quality of automatic translations tends to vary significantly across text segments and MT systems, methods to predict translation quality become more and more relevant. This problem is referred to as Quality Estimation (QE). Different from standard MT evaluation metrics, QE metrics do not have access to reference (human) translations; they are aimed at MT systems in use. Applications of QE include:

- Deciding which segments need revision by a human translator.
- Deciding whether a reader gets a reliable gist of the text.
- Estimating how much effort will be needed to post-edit a segment.
- Selecting among alternative translations produced by different MT systems.

QE was has played a crucial role within QTLaunchPad. In order to achieve the project goals of pushing research towards high quality machine translation, metrics to measure quality in the absence of humans or reference translations, are essential (see above). Such metrics (from simplistic numeric to elaborate MQM-based) have been used for the following main applications within the project:

- Automatically grouping translations into different quality bands for different uses, e.g., gisting vs. dissemination, or different levels of post-editing needed to achieve publication level quality (including retranslation).
- Supporting systematic analysis of translation quality barriers by automatically selecting translations of low-medium quality for analysis by humans.
- (Partially) automating this process of analysis by pin-pointing potential errors in translations, an application directly connected to the work on multidimensional quality metrics (MQM) for (manual) error analysis.
- Improving MT systems by selecting high quality translations that can be directly used, e.g., to supplement training corpora in statistical systems.

The research done towards improving the state of the art in QE within QTLaunchPad involved four major steps: software development, benchmarking of the software on different datasets, preparation of datasets and organisation of shared tasks, experiments with using QE to improve MT.

Although work on QE started about a decade ago (Blatz et al., 2004), most of it focused on estimating automatic metrics such as BLEU and WER. However, these metrics are difficult to interpret, particularly at the sentence level. In addition, before the QTLaunchPad project, there was no quality estimation software (open source or even commercial), and only very few small scale datasets were available, which had been used by individual research groups, i.e., no comparison among various different approaches had been performed on common datasets. The first shared task on the field was proposed by USFD mid-2012 (Callison-Burch et al., 2012), where a simple baseline software tool was made available to participants. No previous work had successfully addressed the use of QE to directly improve MT. In what follows, we cover the main outcomes of the project with respect to QE in the four aforementioned directions.

3.2.1 Software development

Further developing QuEst—an open source software to improve over this simple baseline—was the first major QE step within QTLaunchPad. QE is generally addressed as a supervised machine-learning task using algorithms to induce models from examples of translations described through a number of features and annotated for quality. One of most challenging aspects of the task is the

identification, design and implementation of feature extractors to capture relevant aspects of quality. This was therefore the focus of the software development task within QTLaunchPad. A modular framework for feature extraction was developed in Java that allows different subsets of features to be extracted, and to facilitates the addition of new features. For machine learning, existing algorithms, and in particular an existing toolkit with various algorithms (scikit-learn: http://scikit-learn.org/) were integrated into the framework.

The two versions of QuEst released during the project covered: (i) language- and MT system-independent features from source and translation texts; and (ii) a wide range of advanced, linguistically motivated and MT system-dependent features from source and translation texts, external resources (e.g., source and target language corpora, language models, topic models) and tools (e.g., parsers, part-of-speech taggers), and the MT system that generated the translations, or other MT systems (pseudo-references). Lists of existing features (up to approximately 150 depending on the language pair) are available from the project website (http://www.quest.dcs.shef.ac.uk/). Some of the more advanced and most useful features were reported in recent publications (Almaghout and Specia, 2013; Rubino et al., 2013, Shah et al., 2013a).

This variety of features played a key role in QE, but it also introduced a few challenges. Datasets for QE are usually small because of the cost of human annotation. Therefore, large feature sets raise sparsity issues. In addition, some of these features are more costly to extract as they depend on external resources or require time-consuming computations. Finally, different datasets (e.g., language pair, MT system, or specific quality annotation such as post-editing time vs. translation adequacy) can benefit from different features. Another important component of QuEst was therefore the investigation of feature-selection techniques to help not only select the best features for a given dataset, but also understand which features are in general more effective. We proposed a technique based on Gaussian Processes that proved very effective. The relevance of features is learned as a hyperparameter during model building. As result of a number of benchmarks on different datasets, we were able to study the effectiveness of features across language pairs, text domains, MT systems and quality labels (Specia et al., 2013; Shah et al., 2013b).

Regarding annotation costs, work on selecting unlabelled samples that contribute maximally to improvements of the prediction model, and thus minimise the number of samples to be annotated, was also developed within QTLaunchPad (Beck et al., 2013a). We found that by cleverly selecting data for labelling using active learning techniques it was possible to achieve the same (and sometimes better) QE performance with only about 30% of the data. The fact that better performance could be achieved with less labelled data can be due to the highly subjective (and bias-prone) nature of the process of having humans labelling translations for quality. In other words, some of the translations may be labelled differently by different annotators. As a result, the common procedure of simply averaging different annotators' scores for a given translation or using different annotators to judge different translations is suboptimal. We proposed a way of modelling annotator biases by using multi-task learning Gaussian Process, where each annotator is treated as a "task" and the algorithm learns the inter-annotator correlations and their contribution to the overall quality prediction problem. This approach led to significantly improved results for existing datasets with multiple annotators (Cohn and Specia, 2013).

The current version of QuEst (Shah et al., 2013a) forms the full pipeline necessary for quality estimation: from data pre-processing and feature extraction to feature selection, building and testing of models. The software is available for download from http://www.quest.dcs.shef.ac.uk/. It was used as the official baseline system in the WMT13-14 shared tasks on quality estimation, organised by USFD, as described below.

The latest and final version of QuEst also includes improvements with respect to efficiency and user-friendliness, with a web interface for remote access to facilitate the use of the framework by non-

expert users. This interface allows users to submit files with source and one or more translation sentences (or to get translations from Bing Translator) and get features and predictions for language pairs with models pre-built offline, as well as to perform a ranking of alternative translations of each source segment based on quality predictions when more than one translation is provided (Shah et al., 2014a; Shah et al., 2014b).

3.2.2 Benchmarking

Extensive benchmarkings on the first and second versions of QuEst were performed within QTLaunchPad (Deliverables D2.1.1, D2.1.3; Shat et al., 2013b; Shah et al., 2014b). The benchmarking experiments cover the intrinsic and extrinsic evaluation of QuEst over a number of datasets (up to 13), mostly focusing on the application of quality predictions for dissemination purposes by estimating post-editing effort. A small dataset annotated with quality labels at different levels of granularity was also used in an attempt to predict multidimensional quality metric (MQM) scores, but it turned out to be too small to yield any conclusive results. The intrinsic experiments compared:

- Variants of learning algorithms, more specifically, Support Vector Machines (SVM), which are known to perform well for this task, against a more modern algorithm: Gaussian Process. Other algorithms such as Partial Least Squares and Logistic Regression were also used, but not extensively benchmarked. Gaussian Process was found to outperform SVM, particularly when used jointly with feature selection.
- Different subsets of features, more specifically, a standard "baseline" set with 17 simple features, the set of all features available for the dataset, including new features such as those based on Information Retrieval (IR) techniques, and subsets of features selected using feature selection algorithms. Overall, subsets of features resulting from feature selection algorithms led to the best results across all datasets. Where available for the dataset, a comparison was also made between black-box (MT system-independent) and glass-box (derived from the MT system, such as model component scores from the decoder) features. Glass-box features on their own were found to perform worse than black-box features alone, but improvements were found when both back-box and glass-box features were combined, followed by feature selection.
- Variants of feature selection algorithms, more specifically, a well-known method: Randomised Lasso, and the method adopted within QTLaunchPad: Gaussian Process. The latter was found to outperform the former for all datasets.

As extrinsic task in the benchmarking, quality predictions were used to rank alternative translations from multiple MT systems best-worst. Different methods were compared, including building individual quality estimation regression models for each MT system, building pooled regression models with all MT systems' data together, and building pairwise models (Avramidis, 2013). The results varied for different datasets, but overall it was found that building individual models for each MT system leads to the best results. This extrinsic evaluation of quality predictions was also addressed as a subtask in the 2013 WMT Quality Estimation shared task, described in the following section.

3.2.3 Shared tasks

Two shared tasks on QE were organised as part of QTLaunchPad, both held with WMT, the Workshop on Machine Translation: WMT13 (Bojar et al., 2013) and WMT14 (Bojar et al., 2014). The goals of these shared tasks were to further develop the field of QE by providing a platform for comparing different approaches: common datasets and resources (MT system internal data for some subtasks, corpora, etc.), evaluation metrics and baseline systems. The tasks attracted 14 teams (with

55 system entries) and 10 teams (with 57 system entries), respectively in 2013 and 2014. The baseline feature set from QuEst was used as the official baseline both years. In addition, USFD participated with the latest versions of QuEst in both years, and some of the techniques described above, including learning and feature selection with (multi-task) Gaussian Processes, and active learning (Beck et al., 2013b; Beck at al., 2014). Submissions on advanced systems were also made by DCU, based on Referential Translation Machines with a number of novel features (Bicici, 2013; Bicici and Way, 2014a; Bicici and Way, 2014b), and DFKI, based mainly on advanced parsing features and ranking algorithms (Avramidis and Popovic, 2013; Avramidis, 2014).

In 2013, four subtasks were proposed. Three covered English-Spanish sentence-level prediction of post-editing effort (HTER) and post-editing time, and English-Spanish and German-English ranking of up to five MT systems for every source sentence translation. DCU contributed experiments with using F1 (Bicici, 2013) as automatic evaluation metric to train QE systems and obtained top results. A final task covered, for the first time, word-level quality prediction, more specifically, a binary version to decide, for each word, whether it should be kept as is (i.e., the word was correct) or edited, and a decision between three possible operations: keep as is, substitute or delete. The latter task was done on a very small scale English-Spanish dataset automatically annotated based on heuristics applied on post-editions of machine translations (and therefore, prone to errors). The submissions from USFD and DCU topped the post-editing effort estimation subtask. The submissions from DFKI and DCU came top in the ranking task, with DFKI achieving correlations which were even higher than those obtained by reference-based MT evaluation metrics for German-English.

In 2014, data was collected in closer connection with other tasks within QTLaunchPad. More specifically, it resulted from a joint effort to collect and annotate data for systematic (manual and semi-automatic) analysis quality barriers and for quality estimation. The task included three sentence-level substasks: estimating post-editing effort (HTER), estimating post-editing time, estimating quality bands (1-3); and a word-level task to predict core quality issues from MQM. The datasets covered four language pairs (English↔Spanish, English↔German), and for some subtasks they included translations from 2-3 MT systems (statistical, rule-based and hybrid) and from a professional human translator. These subtasks were set such that participants would not know which systems (or humans) had produced a given translation. This represented a more challenging setting for participants. In the subtask of predicting 1-3 quality bands with 2-3 MT systems plus a human translation, submissions from DCU were ranked top for all language pairs. DCU (tied with USFD and a few others) also came top in the subtask of predicting post-editing time, as well as in the subtask of predicting post-editing effort (according to one of the two official evaluation metrics).

Finally, DCU was the only project partner that also contributed to the word-level quality prediction task in 2014. It outperformed the baseline for all language pairs. Overall, the word-level subtask proved even more challenging than we had anticipated. Despite major efforts to collect a substantial number of segments annotated with MQM, the variety of language pairs, translation systems, text domains, and MQM issue types in the datasets made the annotations very sparse, with very few examples for most issue types for a specific language pair. The shared task however set the ground for further work on MQM prediction, with the establishment of guidelines for MQM annotation (in this case using the open source tool translate5 (www.translate5.net) for annotation, core issue types, protocols for training annotators, etc. As more data is collected, we expect better systems for word-level prediction to be proposed.

Overall, the results from both WMT shared evaluation tasks showed that the systems proposed within QTLaunchPad, all based on QuEst, present the state of the art performance in most variants of QE tasks.

3.2.4 QE to improve MT

The goal of this final contribution towards high quality translation within QTLaunchPad was to use the QuEst framework to improve MT systems' quality. Different sets of experiments were performed, as described below.

The first set of experiments (Deliverable D2.2.1) used QuEst models to improve the performance of statistical machine translation systems without changing the internal functioning of such systems. The experiments included the following approaches: (i) n-best list re-ranking, where translation candidates (segments) produced by a machine translation system are re-ranked based on predicted quality scores such as to get the best translation ranked top; (ii) n-best list re-combination, where sub-segments from then n-best list are mixed using a lattice-based approach, and the complete generated segments are scored using quality predictions and then re-ranked as in (i); (iii) system selection, where translations produced by multiple machine translation systems and a human translator are sorted according to predicted quality to select the best translated segment, including the challenging case where the source of the translation (i.e., which system/human produced it) is unknown (same datasets as for the 1-3 quality bands prediction in WMT14), and (iv) diagnosis of statistical machine translation systems by looking at internal features of the decoder and their correlation with translation quality, as well as using them to predict groups of errors in the translations.

Experiments were performed on a number of datasets and feature sets produced using QuEst. We observed improvements with the n-best re-ranking approach in terms of BLEU scores and other automatic metrics on the set of top translations (1-best) selected from the n-best list, particularly with n=100. For word-level re-combination, we observed performance increases with larger n, with the best results obtained when n=1000. With system selection, we observed promising results with a blind setting where the sources of the translations (i.e., MT systems) are unknown. Finally, as for the diagnosis of statistical translation systems, we found that statistical machine translation system-related features coming from the decoder can provide very useful information in understanding overall translation quality, as well as in predicting specific error types.

The second set of experiments (Deliverable D2.2.2) used QuEst models to improve the performance of statistical machine translation systems by supplementing their training corpora or by using the ITERPE model described in (Bicici and Specia, 2014). The experiments used different quality-informed active learning strategies to select, among alternative machine translations, those which are: (i) predicted to have high quality, and thus can be added to the machine translation system training set; (ii) predicted to have low quality, and thus need to be corrected/translated by humans, with the human corrections added to the machine translation system training set. The experiments were performed in an iterative manner with translations produced by a statistical machine translation system that was updated at every iteration, and in a static manner with translations from a rule-based machine translation system used to update the statistical system. Improvement is measured by the increase in the performance of the overall machine translation systems on held-out datasets, in terms of an automatic evaluation metric (BLEU) comparing the scores of the original machine translation system against the score of the improved machine translation system after additional material is used. Both strategies perform better than random selection, with both the statistical and the rule-based system translations. More important, adding machine-translated segments predicted to have high quality leads to improvements that are comparable to adding reference translations. The iterative experiments with the statistical translation system show that the system learns to produce better translations over time, as more machine-translated segments are added to its training corpus. The ITERPE model consists in automatically grouping translations into different quality bands for instance for retranslation or for post-editing (Bicici and Specia, 2014). These results can be helpful in automatic identification of quality barriers in MT to achieve high quality machine translation.

Finally, USFD also experimented with using quality predictions as a metric for the tuning of statistical MT systems. A number of experiments directly using QE scores on themselves, or combined with reference-based metrics like BLEU, led to no significant improvements on translation quality. We hypothesize that this is due to the high levels of similarity amongst hypotheses in n-best lists of statistical MT systems, for which a more fine-grained quality prediction approach would be needed. However, the use of QE indirectly proved positive. In (Song et al., 2014), we use a few QE features (not predictions) as a way to select tuning data that potentially has higher quality from a pool of noisy data (Common Crawl) in order to train the parameters of a statistical MT system. Significant improvements were obtained even in comparison with using pre-defined, professionally created tuning sets. More interestingly, these were obtained based on the selection of the tuning set using purely QE features and other heuristics comparing the source and target segments (such as alignment scores), without considering the similarity of these segments to the test set, nor decoding-based features.

QE by itself is not as informative as reference-based metrics for the tuning of SMT systems, as the task of distinguishing very similar translation candidates in an n-best list for a given source sentence in terms of their quality is arguably more challenging than that of distinguishing quality in candidate translations for different source sentences or by different MT systems. This has been shown in previous work on QE for re-ranking n-best lists and is also reported in D2.2.1. Therefore, using QE as the *sole* metric led to worse overall translation quality.

The combination of QE and reference-based metrics is a more promising avenue, as we believe the two types of metrics are complementary. However, this task involves two significant challenges: (1) QE has to be done at a granularity level that allows it to distinguish between very similar translation candidates in the n-best list—i.e., using word-level prediction—but appropriate word-level models are still not very effective; (2) one needs to tune the weights of the metrics, as a simple unweighted linear combination has not proved effective. This tuning is a complex problem on itself as it would require as gold-standard translations assessed by a metric that is more reliable than the ones we are combining (possibly human annotations). However, human annotation of hundreds of thousands of candidate translations in n-best lists is simply not feasible. Experiments with various manually set weights led to no improvements, and therefore were not reported.

3.3 MT-specific infrastructure & extension of a platform for resource sharing for large-scale collaborative MT research (WP 3 & 4)

3.3.1 Introduction

The availability and access to appropriate language resources that support the full range of machine translation paradigms and multilingual information processing in general is central to MT technology for high-quality translation and associated multilingual applications development. Most research groups spend a significant amount of time on tasks like searching for and preparing corpora, defining of development sets, and searching for necessary tools. On the other hand, time-consuming tasks like human evaluation are normally not performed due to the high complexity required to organize them.

A major goal of the QTLaunchPad project has been to ease the development and improvement of translation systems, by facilitating and allowing for a much quicker development-testing-evaluation cycle than was previously possible. QTLaunchPad's objective has thus been to provide both resources relevant to MT and an extended infrastructure in order to improve research in machine translation by providing not only access to large amounts of data, but also automation of common tasks in machine translation research, such as pre-processing, text alignment, annotation etc. Furthermore, special attention has been given to evaluation workflows and the implementation of tools for quality estimation and human evaluation of machine translation output.

Therefore, parallel actions were taken in two directions: first, to define and provide the necessary infrastructure for supporting different workflows for the development of quality machine translation systems; and second, to detect, acquire, document, and share appropriate language resources.

3.3.2 Definition of workflows and tool interfaces

The project identified and documented the common workflows[6] particular to the development and evaluation of MT systems[7]. Workflow definitions were based on the results of a survey on the Requirements of Relevant Stakeholder Groups[8]. This survey addressed the needs of a diverse group of stakeholders with a specific interest in translation and translation technology: freelance translators, academic researchers, lecturers, managers, directors/owners of an institution/company/organisation, as well as representatives from government organisations and the industrial sector. The workflows described address the following tasks:

- **Acquisition:** the most important components in statistical machine translation (SMT) systems are parallel corpora, the size and the quality of which affect the SMT system output.
- **Preparation:** corpus preparation involves corpus filtering, focusing on homogenizing the corpus in terms of the lengths of the sentences it contains, tokenization, and casing, mainly to reduce the number of out-of-vocabulary items found during translation. Tokenization, casing, and morphological analysis may not be necessary for all corpora and may depend on the task.
- **Translation:** including training, translation and target language modeling, decoding/testing, minimum error training and development.
- **Evaluation:** translators, linguistics, localisation engineers, and QA specialists are the typical roles involved. Evaluation is widely recognised as an integral part of MT development. Approaches to evaluation can be categorised into two main paradigms: human evaluation and automatic evaluation. Human evaluation is often regarded as the gold standard for MT evaluation, but it can be resource-expensive in terms of time, tools, and expertise. Alternatively, automatic evaluation metrics (AEMs) often represent a faster and cheaper method for error analysis, system comparison and system optimisation. Special focus was given so as to allow for a more efficient and accessible integration of human evaluation processes into MT development.

Many of the workflows identified require the use of external tools, like for example tokenizers, par-of-speech taggers, parsers, etc.[9] Interfaces for such tools were defined, at an appropriate level of abstraction, to facilitate integration and linking to a machine translation infrastructure[10]. For each tool category the interface specification concentrated on common aspects (e.g. standardized data formats for input/output), but allowed flexibility in selecting among alternative implementations and consequently enabling realistic expectations to be made at different sections of the QTLP information flow pipeline.

3.3.3 Implementation

The specification of workflows and component interfaces highlighted the desired features a machine translation infrastructure should have in order to meet development and evaluation requirements.

---

[6] Workflow is the flow of information beginning from input to output including any transformations and the tools involved in between.
[7] See QTLaunchPad Deliverable D3.1.1.
[8] See QTLaunchPad Deliverable D6.5.1.
[9] For the availability and documentation of such annotation tools see QTLaunchPad Deliverable D4.4.1).
[10] See QTLaunchPad deliverable D3.2.1.

Initial prototypes[11] with limited functionality, focusing on key parts and related components of an MT infrastructure were made available in order to collect feedback from the users and incorporate it into the development process and as a result improved infrastructure components were delivered[12]. Taking into consideration QTLaunchPad's specific objectives and focus, special attention has been given to the evaluation workflows and to the data acquisition, documentation and annotation components of the infrastructure including the following:

- Scorecard and Metric Builder, a tool for building MQM (Multidimensional Quality Metrics);
- QuEst, a system for quality estimation;
- translate5, an MQM based quality evaluation interface, and
- QT21 repository, a META-SHARE compliant repository for resource documentation, sharing and processing/annotation.

3.3.3.1 The MQM Metric Builder

The **MQM Metric Builder** tool provides a way to build custom MQM metrics based on the core issue types. Users are guided to define MQM dimensions and select corresponding MQM issue-type inventories. An MQM-based online quality scorecard system is integrated with the Metric Builder and allows users to generate error count-based scores without doing detailed quality markup[13]. Source code for the Scorecard and Metric Builder can be downloaded from http://www.qt21.eu/downloads/scorecard-source-2014-06-30.zip. The scorecard can be accessed online at http://scorecard2.gevterm.net. (NOTE: As of December 2014 we are currently not promoting this version as we are working on a new version that will be tied into the Scorecard tool.)

3.3.3.2 QuEst

The QuEst (for **Qu**ality **Est**imation) system from the University of Sheffield provides automatic quality estimation for submitted content. Users can either get a ranked comparison of two or more translations submitted to QuEst or can get a subsample of results that are estimated to have the highest quality levels[14].

Pre-built MT system-independent quality prediction models for the language pair and text domain at hand are used to produce quality estimates for the translations. For the subsampling functionality, the dashboard allows the user to specify the proportion of the dataset to be annotated. In future versions, it will also allow the user to specify whether the sampling should be done based on translations from one particular MT system, or—as it is implemented currently—whether it will be done taking all systems into account (i.e., the best ranked translation among all options for each source sentence will be used as the basis for the sampling). Source code for QuEst is available at http://www.qt21.eu/downloads/quest-source-2014-06-24.zip.

3.3.3.3 Translate5

Translate5, implemented by Marc Mittag (MittagQI - Quality Informatics, subcontracted by DFKI) is a powerful tool, implementing a user interface to databases and incorporating many of the features described in D3.4.1. It provides an easy-to-use interface for annotating data with MQM issues, as

---

[11] See QTLaunchPad deliverable D3.4.1
[12] See QTLaunchPad deliverables D3.4.2, D3.4.3
[13] Output from this component can be applied *only* to entire datasets or sampled subsets, not to individual segments, but a revised version is currently under development that will track which segments are visible when issues are added. This version will occupy a space between a simple score card and the more advanced functionality available through translate5.
[14] Note that quality estimation does not yet necessarily conform to any specific MQM quality assessment level, but is instead intended to be an indicative assessment for further work.

well as for post-editing, ranking, and other common translation research tasks. The interface works on MySQL databases, in which source texts, translations and annotations are stored. It can be used for manual translation, pre- and post-editing and for quality assessment. For automatic processing steps, as well as for data import and export, a programmable interface is foreseen. One of the key features of translate5 is that it allows multicolumn comparisons and selections. Thus, it enables comparisons and analyses that are currently difficult to accomplish because data are stored in heterogeneous files with no relational structure. Translate5 can make use of the result from QuEst or other tools if downloaded in CSV format and uploaded as a new project. Current developments will result in an API-driven connector that will allow for more integrated access.

The translate5 system does not currently have an installer and requires extensive knowledge to install this software. An installer is currently under development and planned for the second half of 2014. Although many of translate5's features (multi-target project support, MQM annotation, user management, among others) have been funded by QTLaunchPad, it should be noted that the source code for translate5 contains many features funded by other parties. This mixture of features and code is an example of how QTLaunchPad was able to leverage existing commercial development to deliver a more capable infrastructure component than would have been possible with an independent development project. The latest build of translate5 is available at http://dfki.translate5.net. It is possible to log into the system with the user name **dfki-test** and password **playground**.

3.3.3.4 The QT21 repository

3.3.3.4.1 Identification, acquisition and documentation of resources

The acquisition of high-quality corpora and the availability of processing tools are essential for high-quality MT as indicated by the acquisition and preparation workflows (D3.2.1). In order to optimize the process of choosing the appropriate corpus for a specific experiment according to its features, it is important to maintain a repository of available corpora specifying metadata for each corpus, including language, type, genre, sources used, etc. To provide all support mechanisms needed for documentation, sharing, search and retrieval of all MT-related resources, QTLaunchPad builds upon and extends the META-SHARE infrastructure (www.meta-share.eu, www.meta-share.org). A QTLaunchPad–dedicated META-SHARE node/repository, http://qt21.metashare.ilsp.gr/, has been set up and populated with MT-related language resources and/or their metadata-based descriptions. The documentation of resources follows the META-SHARE metadata model. The source code of the META-SHARE software is distributed under the BSD license and can be downloaded from https://github.com/metashare/META-SHARE.

All resources with permissive terms of use have been replicated on the QTLaunchPad repository for reasons of service efficiency. The language resources with which the repository was populated are, as mentioned before, primarily but not exclusively parallel corpora. Moreover other types of datasets have been selected: monolingual and multilingual corpora, raw or annotated, parallel or comparable, as well as lexical and conceptual, in principle structured, datasets[15,16]. In parallel, a selection of tools and processing services was made, namely tokenizers and sentence splitters, part-of-speech taggers

[15] Lexical Conceptual Datasets are Computational Lexica, FrameNets, Lexica, Machine Readable Dictionaries, Terminological Resources, Ontologies, Thesauri, Wordnets and Word Lists. The list provided is according to the sub-types of the lexicalConceptualResourceType component of the META-SHARE metadata schema.

[16] A detailed documentation is available in D4.1.1 Inventory and documentation of existing monolingual and multilingual data.

and lemmatizers, syntactic parsers, named entity recognizers and parallel text aligners, as they prove necessary during the preparation of a corpus for translation[17].

The resources selected to populate the QT21 repository come from the following sources:

i. resources already available in the META-SHARE network, as well as language resources currently widely used in MT research (e.g. resources from the Euromatrix(+) and OPUS),
ii. datasets coming from the QTLaunchPad consortium partners,
iii. domain and genre-specific textual data being discovered automatically from the web in order to fill in missing datasets - this type of (monolingual, bi- and multli-lingual) data has been identified through focused web-crawling tools.

The QTLaunchPad language resources cover the languages of interest for the project (English, German, Portuguese and Greek) as well as the corresponding language pairs. The datasets were also examined with regards to the defined four Research and Innovation Application Scenarios (RIAS)[18]. The monolingual resources gathered from the META-SHARE network and the QTLaunchPad consortium partners offer a good starting point for the "horizontal" RIAS ("Public" and "Media"), and to a sufficient extent for the vertical "Medical" RIAS.

As indicated by the limited size of the resources available at the project's start for the "Medical" and "Automotive" domains, collections (for all project languages) were created using the ILSP focused crawler (ILSP-FC), a tool that was extended and enhanced in the context of QTLaunchPad. ILSP-FC[19] integrates modules for normalization (encoding detection and conversion to UTF8), language identification, boilerplate detection, text classification, link extraction, URL ranking, (near) duplicate detection and removal, and exporting of acquired documents in an XML variant of the cesDOC Corpus Encoding Standard (http://www.xces.org/). ILSP-FC has been enhanced to include new functionalities for, among others, data acquisition (e.g. text extraction from PDF documents), language identification (i.e. integration of alternative language identifiers), web page genre classification and metadata extraction (e.g. detection of Creative Commons (CC) licenses). For genre classification, the following five genres were defined: Discussion (e.g. forum web sites), Reference (e.g. web pages containing scientific articles, frequently answered questions, wiki pages, etc), News/Journalism, Commercial and Other. For the classification task, a heuristics-based genre classifier was used examining the metadata of a web page (e.g. URL and title) and searching for specific terms/patterns (e.g. forum, wiki, advices, how to, news, noticias, /store/, /obidos/, productid=,) implying that a web page is relevant to one of these genres. An enhanced method for near de-duplication was applied during data acquisition: a de-duplicator module identifies (near) duplicate documents and discards the shortest in term of words.

The ILSP-FC[20] can, depending on user-defined configuration, employ processing workflows for the creation of either domain-specific or general monolingual or bilingual collections. For the latter, a module that would decide which acquired pages could be considered as pairs of parallel documents has been implemented. This was achieved in the following way: after in-domain pages are downloaded, the Crawler Pair Detector module uses three methods to identify pairs of pages that could be considered parallel. The first method compares the URLs of the acquired web pages and considers them candidate translations if the differences only concern special patterns like /l1/ and /l2/, etc; the second method is based on co-occurrences, in two documents, of images with the same filename, while the third takes into account structural similarity. The main difference between using the ILSP-FC for building monolingual and bilingual collections concerns the selection and

[17] See QTLaunchPad deliverable D4.4.1.
[18] See QTLaunchPad deliverables D6.2.x
[19] ILSP-FC is available as an open-source Java project from http://nlp.ilsp.gr/redmine/projects/ilsp-fc
[20] See QTLauchPad deliverables D4.2.1 and D4.3.1

prioritization of extracted links. In case the crawler is used for acquiring monolingual data all links are added to the list of the links to be visited and the score link is mainly influenced by the "domainess" of its surrounding text. When ILSP-FC is used for building multilingual collections, only the links that point inside the targeted web sites are selected and the link score is powered by the probability that the link under consideration originates from a web page in L1 and "points" to a web page that is probably in L2. This is the case when, for example, the targeted languages are EN and DE, and the anchor text of a link in an English web page contains strings like "de", "Deutsch", etc. Thus, the crawler is forced to visit candidate translations before following other links. The collections of monolingual corpora created with the ILSP-FC have been documented, stored and rendered downloadable, on the condition that their associated licences permit so, through the QT21 repository.

Members of the QTLP and QT21 communities are able to register, access and download the resources residing in QT21. Resource providing members are able to modify their resources and their descriptions, as well as add new ones, thus enriching and keeping the repository and inventory updated. Providers also have the ability to provide their actual data selecting the appropriate data format and respecting the repository size limitations (max. 35MB). Repository population is and will be an ongoing process, aiming to offer a reference point including all datasets and tools available, that are relevant and fit for MT research and development, respecting all legal (and possibly other) restrictions and preferences.

3.3.3.4.2 Integration of language processing tools

Extending the META-SHARE functionalities and providing for the implementation of data preprocessing and annotation workflows described above, the QT21 repository offers access not only to datasets, but also the ability to use the QT21 repository tools and services for preprocessing and annotating them. The implemented language processing layer[21] provides a growing number of services: tokenisation, sentence splitting, pos tagging, lemmatisation, chunking, dependency parsing, and named entity recognition, both for monolingual and multilingual textual datasets as well as text alignment for multilingual textual datasets. Language processing tools are documented with the appropriate metadata and are provided as web services through the language-processing layer. These services are currently offered for the all the project languages and datasets coming in various formats (txt, xces/xml, tmx, Moses). The services can be locally or remotely deployed and chained into appropriately defined, internally interoperable workflows. Currently, OpenNLP services are deployed for English, German and Portuguese, Panacea-DCU services for English, DFKI-Heart-of-Gold services for German, LX-Center/University of Lisbon for Portuguese, and ILSP NLP services for Greek. Each QT21 natural language processing workflow chains together components or services of the same suite/family of tools. To accommodate cases where the services deployed belong to different suites, the appropriate converters have been developed. Enabling the user to define and deploy custom workflows, cross-suite or not, is on our agenda for the immediate future.

In a typical scenario, when the user selects to process a dataset, a list of all available annotation services for each relevant annotation level (e.g. tokenization & sentence splitting, POS tagging, lemmatization, alignment) are provided for the given language, and resource type. As soon as the user selects a tool, the server invokes a service that dispatches the corpus to the specific web service for processing. The system informs the user about the requested job via the messaging service of the platform. When the processing has been completed, the new (annotated) dataset is automatically stored and indexed in the repository, and the user is appropriately informed. The repository keeps track of the processing requests so that if the user, for any reason, requests to process a dataset with a

[21] See QTLaunchPad deliverable D3.3.1.

specific tool, and this dataset has already been processed by the specific tool, then the system will just forward the user to the automatically generated metadata record of the processed dataset that has been created and stored in the repository.

3.3.4 Testing

The correctness of the implementation of the different components of the MT infrastructure was guaranteed by running multiple tests[22]. The translate5 platform underwent multiple rounds of testing in the second year of the QTLaunchPad project through annotation tasks, use in workshops hosted by QTLaunchPad (the Machine Translation Evaluation, MTE, LREC workshop on May 26, 2014 in Reykjavik), and use by LSPs cooperating with the QTLaunchPad project. DFKI staff used the software to administer the on-going annotation tasks and to prepare for the workshop organised by the project. Furthermore, the user interface and instructions for use of translate5 were evaluated by an expert in usability and reliability testing and specific suggestions for improvement have been obtained and used to improve the system.

The scorecard system was not deployed as extensively within the project, but nevertheless received extensive testing through a pilot deployment (two small sprints for Firefox 28) and subsequent production deployment with the Mozilla Project. It was subsequently rolled out into a test of 20 languages (14 European). Finally, as of June 2014, Mozilla is planning to adapt the scorecard to address the scope of the entire Firefox project (90 languages). The scorecard was also assessed by a U.S./Germany language service provider (Multiling), which provided extensive feedback and suggestions.

Feedback from users, professional translators among others, and the industry has been gathered through multiple testing rounds resulting in making user demands clear and allowing for these systems to be improved in various ways (interface, technical abilities, data format, ergonomics etc.).

QuEst has been used by a number of people from different institutions beyond the QTLaunchPad project partners, showing how effective it is in predicting segment-level quality. Among others, it has been used in the MATECAT project predict the quality of sentences while dynamically adapting the models to specific users, and in the SUMAT project to estimate the post-editing effort for subtitles, filtering out segments that are too bad to be post-edited. In addition, a subset of it with 17 features has been used as the official "baseline" system in the WMT12-13-14 shared tasks on quality estimation (estimation of post-editing time, edit distance, and post-editing effort, and ranking of machine translation systems), achieving very competitive performance. QuEst has also been tested by companies like Yandex and Multilizer, which participated in the WMT14 shared task.

On the QT21 META-SHARE side, automatic unit tests have been performed for all the (sub) processes involved in the META-SHARE extension. Functionalities for concurrent users and size of data to be processed have also been tested. Tests were intended to make sure that each function in the resource processing chain returns the correct value for a specific input and the repository performs in accordance with the design choices made. In addition the production of the correct metadata for the new processed resource has been tested. The unit tests constitute the complete set of automatic procedures implemented for testing the whole QT21 repository data management and processing lifecycle, from data documentation and uploading to processing and storing back the respective results.

[22] See QTLaunchPad deliverables D3.5.1 and D3.5.2

3.3.5 Legal framework

Throughout the project, key legal issues were investigated with regard to Language Resources re-use, particularly in MT and Machine Processing (MP) settings. The LR processing requires the use and re-use of information of various kinds, in a variety of ways, by different types of organisations and, as a result, it involves a wide range of legal regimes. The issues relate primarily to Intellectual Property Rights (IPR), but it may also involve Personal Data Protection, Public Sector Information (PSI) and other related Regulations.

In order to assess the conditions under which LR re-use in MT and MP may lawfully take place, we had to understand the acts it involves, the degree to which they are regulated by different types of laws, and the permissions that someone needs to obtain in order to perform such acts. The legal study undertaken in QTLaunchPad elaborates on and explains the core concepts from the main legal regimes that influence the LR-based MT & MP, explains how the flows of rights are to be treated in a paradigmatic MT & MP scenario, classifies the key emerging issues, and presents the core conclusions with regard to the way LR-based MT & MP are to be treated.

A set of core use cases that may be used as classes of scenarios of use has been created to illustrate different aspects of the use of LR for MP/MT purposes. Special attention has been given, mainly, but not exclusively, to textual datasets derived from the web through automatic crawling techniques and for which no clear licensing conditions are available, so as to determine under what terms of use they can be made available. This mainly involves the delineation of copyright exceptions, in particular for text and data mining operations, monolingual or multilingual. These use cases scenarios have been explored with regard to the re-use of data so as to present the best long term solutions, as widely accepted as possible, fostering use and reuse of existing content and textual data, structured or unstructured, and their derivatives for the purpose of training language technology tools The problem statements of each use case and their answer is organized in a FAQ form and constitutes an integral part of the study[23].

In tandem, QTLaunchPad has followed developments in legal framework reform in the EU and the associated public consultations on copyright including Text and Data mining issues. In this context, QTLaunchPad submitted its own contribution as respondent to the EU consultation (http://ec.europa.eu/internal_market/copyright/initiatives/index_en.htm), promoting the peculiarities of data use and reuse in the text mining, machine translation and language technology in general.

A rather simplified operational model has been adopted for the QT21 repository infrastructure. It permits only openly licensed, with no no-derivatives (ND) restriction datasets to be processed by openly licensed services and workflows. In future versions, QT21 will be equipped with a business logic that will allow processing of otherwise licensed datasets and services supporting the appropriate business models.

3.4 Preparing a big QT action & Outreach (WP 6)

3.4.1 Preparing QT Action

The QT21 action will provide a dedicated and systematic attack on quality barriers using cutting edge competitive research and innovation linked with strong community and stakeholder engagement. The novelty of this approach to MT research and development will consist of our growing network of motivated translation and localisation professionals coupled with an established relationship with the European translation industry.

---

[23] See QTLaunchPad deliverable D4.5.1

A key strategic factor for achieving high-quality translations in the QT21 action is to concentrate on specific domains and types of content by systematically collating, pre-processing, and exploiting domain- and content-specific corpora effectively. The missing key technologies needed for these highly specific processes and data will be developed. This is a contrast to the "one size fits all" approach employed by all publicly available online translation services, which cannot deliver high-quality translation for many specialised areas.

The QT21 action will therefore concentrate on the selected domains and content types as described in various Research Innovation and Application Scenarios (RIAS) (see D6.7.1). Owing to the inclusion of actual users as partners in the services to be developed through this action, individual projects will be able to involve them more closely and consistently in order to adapt technologies and methods to meet their needs. At the same time this allows for the sharing of valuable user feedback and expertise and allows access to vast repositories of existing data sets.

The QT 21 Action will happen in 4 phases:

- Phase 1: Defining the RIASes
- Phase 2: Negotiating specific activities with partners for the RIAS to produce a roadmap and master plan
- Phase 3: Planning QT21 activities, projects, etc.
- Phase 4: Satellite actions/projects, updating roadmap, and masterplan

The QT21 Planning Panel is a board consisting of stakeholders who would qualify as potential participants of QT21. The planning process, which is driven and coordinated by the QTLaunchPad Extended Steering Board, consists of several phases. The first phase is centred around the definition of the RIASes , that should later guide R&D in QT21. The first meetings of the planning panel took place as follows:

- September 13–14, 2012 at DFKI in Berlin
- October 25, 2012 at Hotel Crowne Plaza in Wiesbaden (co-located with tekom/TCworld fairs)
- March 13, 2013 at FAO Headquarters in Rome (co-located with Multilingual Web Workshop)

At the meetings, large space was given to the presentation and discussion of use cases and the respective research strategies needed.

The main goal of the third planning phase was to turn the use cases (RIASes) for QT21 into a master plan, including a leading-edge research plan and roadmap for the field of MT. The roadmap for MT developed within QTLaunchPad is a continuation and concretization of the roadmap for the whole field of language technologies as defined in the META-NET Strategic Research Agenda for LT in Europe presented publicly in January 2013.[24] This Roadmap and master plan is described in more detail in D6.4.2

In addition, work in this phase included negotiating with potential partners for specific aspects of the QT 21 Action. From these negotiations we were able to elicit both feedback on our work and also identify interested partners to work in these areas. Consequently a number of MoUs between national funding agencies were drawn up and letters of support from relevant industry stakeholders were collected (See D6.8.2 for more details).

[24] http://www.meta-net.eu/sra-en

Within the context of QTLaunchPad, extensive planning was made for the QT21 action with "satellite" projects. These would include implementation-driven projects seeking to apply QTLaunchPad/QT21 results in various industry areas as well as locally funded infrastructure development projects. These satellites will play a key role in anticipated large MT action by (a) providing a way for the results to be implemented and tested in real-world scenarios to ensure that they work not only in research situations and (b) extending the impact of EU funding to additional areas through use of funding from member states.

3.4.2 Overview of QTLaunchPad Stakeholder Engagement and Outreach Activities

Stakeholder engagement and outreach activities undertaken throughout the project were essential to raise awareness of the work conducted by the consortium among both the academic community and the industry, and to gather feedback on the concerted European action in HQMT. Multiple formats and channels were used to maximise the impact of these engagement and outreach activities, to effectively reach diverse stakeholder groups and exploit all the opportunities for interaction and feedback gathering.

The frequency and intensity of all these diverse outreach activities progressively grew during the progress of the project. The conferences and events where the project was represented are all well-established in the field, and in addition there were a number of QTLaunchPad workshops and events collocated with major conferences. A full listing of outreach efforts is provided in **Section 4, item A2** (below).

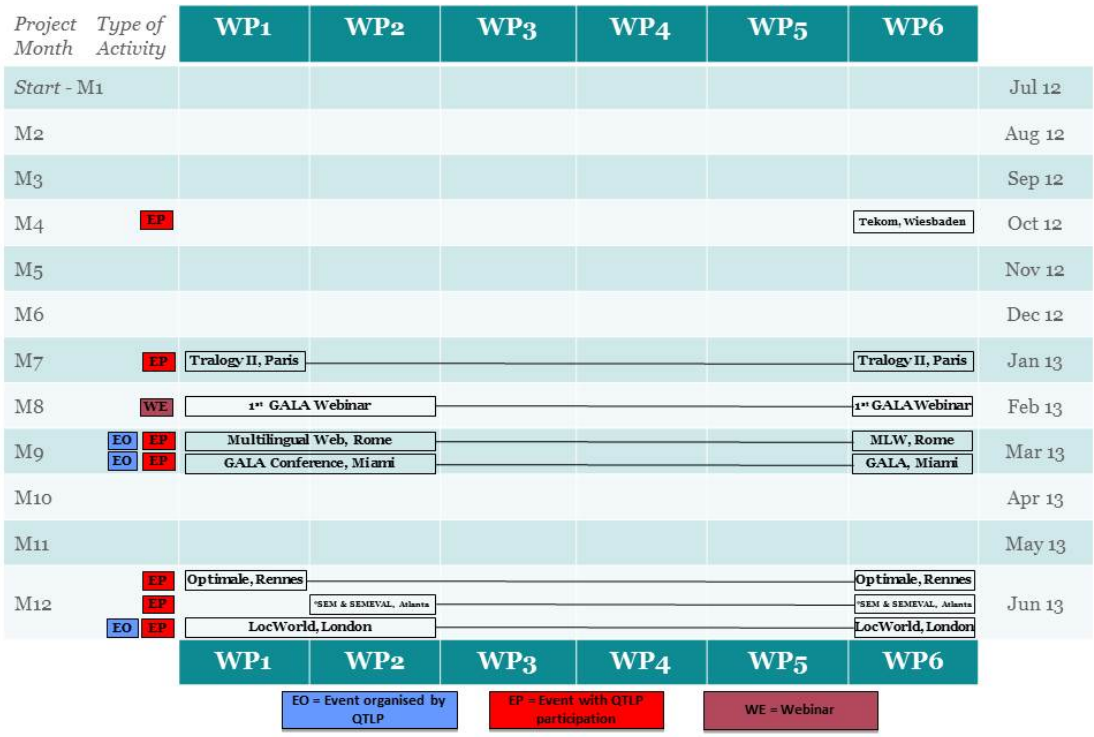

Figure 1: Overview of stakeholder engagement and outreach activities in year 1.

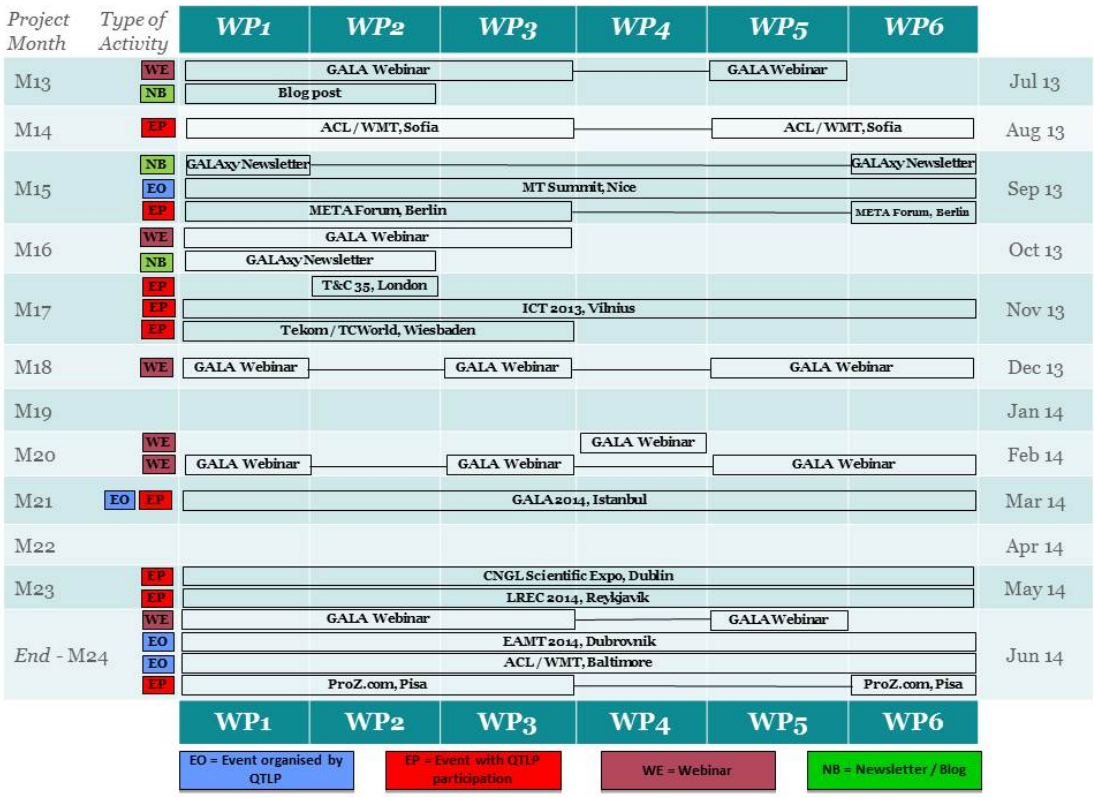

Figure 2: Overview of stakeholder engagement and outreach activities in year 2.

4 Use and dissemination of foreground

4.1 Section A (public)

4.1.1 Section A1: List of Scientific (Peer-Reviewed) Publications

NO.	Author(s)	Title	Year	Title of the periodical or the series	Pages	Open access?
1	Toral, Antonio	Hybrid selection of language model training data using linguistic information and perplexity	2013	*Proceedings of the 2nd Workshop on Hybrid Approaches to Translation*	8–12	yes
2	Specia, Lucia, Kashif Shah, José Guilherme Carmago de Souza & Trevor Cohn	QuEst - a translation quality estimation framework	2013	*Proceedings of the 51st Annual Meeting of the Association for Computational Lingustics (ACL13): Demonstrations*	79–84	yes
3	Song, Xingyi, Lucia Specia & Trevor Cohn	Data selection for discriminative training in statistical machine translation	2014	*Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT14)*	45–52	yes
4	Shah, Kashif, Trevor Cohn & Lucia Specia	An investigation on the effectiveness of features for translation quality estimation	2013	*Proceedings of MT Summit XIV*	167–74	yes
5	Shah, Kashif, Marco Turchi & Lucia Specia	An efficient and user-friendly tool for machine translation quality estimation	2014	*Proceedings of the 9th International Conference on Language Resources and Evaluation*	3560–64	yes
6	Shah, Kashif, Eleftherios Avramidisi, Ergun Biçici & Lucia Specia	QuEst - Design, implementation and exstensions of a framework for machine translation quality estimation	2013	*Prague Bulletin of Mathematical Linguistics* 100	19–30	yes
7	Shah, Kashif & Lucia Specia	Quality estimation for translation selection	2014	*Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT14)*	109–16	yes
8	Rubino, Raphael, José Guilherme Camargo de Souza, Jennifer Foster & Lucia Specia	Topic Models for translation quality estimation for gisting purposes	2013	*Proceedings of MT Summit XIV*	295–302	yes
9	Rubino, Raphael, Antonio Toral, Santiago Cortés Vaíllo, Jun Xie, Xiaofeng Wu, Stephen Doherty & Qun Liu	The CNGL-DCU-Prompsit translation systems for WMT13	2013	*Proceedings of the 8th Workshop on Statistical Machine Translation (WMT13)*	213–18	yes

NO.	Author(s)	Title	Year	Title of the periodical or the series	Pages	Open access?
11	Piperidis, Stelios, Harris Papageorgiou, Christian Spurk, Georg Rehm, Khalid Choukri, Olivier Hamon, NIcoletta Calzolari, Riccardo Del Gratta, Bernando Magnini & Christian Girardi	META-SHARE: one year after	2014	*Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC14)*	1532–38	yes
12	Papavassiliou, Vasilis, Prokopis Prokopidis & Gregor Thurmair	A modular open-source focused crawler for mining monolingual and bilingual corpora from the web	2013	*Proceedings of the 6th Workhop on Building and Using Comparable Corpora*	43–51	yes
13	Lommel, Arle, Maja Popović & Aljoscha Burchardt	Assessing inter-annotator agreement for translation error annotation	2014	*Proceedings of the MTE Workshop on Automatic and Manual Metrics for Operational Translation*	31–37	yes
14	Lommel, Arle, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis & Hans Uszkoreit	Using a new analytic measure for the annotation and analysis of MT errors on real data	2014	*Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT14)*	165–72	yes
15	Labropoulou, Penny, Christopher Cieri & Maria Gavrilidou	Developing a framework for describing relations among language resources	2014	*Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC14)*	3843–50	yes
16	Gaspari, Federico, Antonio Toral, Arle Lommel, Stephen Doherty, Josef van Genabith & Andy Way	Relating translation quality barriers to source-text properties	2014	*Proceedings of the MTE Workshop on Automatic and Manual Metrics for Operational Translation*	61–70	yes
17	Doherty, Stephen, Federico Gaspari, Declan Groves, Josef van Genabith, Lucia Specia, Aljoscha Burchardt, Arle Lommel & Hans Uszkoreit	Mapping the industry I: findings on translation technologies and quality assessment	2013	*European Commission Report*	—	yes
18	Doherty, Stephen & Joss Moorkens	Investigating the experience of translation technology labs: pedagogical implications	2013	*Journal of Specialised Translations* 19	122–36	yes
19	Cohn, Trevor & Lucia Specia	Modelling annotator bias with multi-task Gaussian processes: an application to machine translation quality estimation	2013	*Proceedings of the 51st Annual Meeting of the Associaton for Computational Lingustics (ACL13)*	32–42	yes
20	Bojar, Ondrej, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Hervé Saint-Armand, Radu Soricut, Lucia Specia & Aleš Tamchyna	Findings of the 2014 workshop on statistical machine translation	2014	*Proceedings of the 9th Workshop on Statistical Machine Translation (WMT14)*	12–58	yes
21	Biçici, Ergun, Qun Liu & Andy Way	Parallel FDA5 for fast deployment of accurate statistical machine translation systems	2014	*Proceedings of the 9th Workshop on Statistical Machine Translation (WMT14)*	59–65	yes
22	Biçici, Ergun, Declan Groves & Josef van Genabith	Predicting sentence translation quality using extrinsic and language independent features	2013	*Machine Translation Journal* 27	171–92	yes

NO.	Author(s)	Title	Year	Title of the periodical or the series	Pages	Open access?
23	Biçici, Ergun & Josef van Genabith	CNGL: Grading student answers by acts of translation	2013	*Proceedings of the 2nd Joint Workshop on Lexical and Computational Semantics*	585–91	yes
24	Biçici, Ergun & Josef van Genabith	CNGL-CORE: Referential translation machines for measuring semantic similarity	2013	*Proceedings of the 2nd Joint Workshop on Lexical and Computational Semantics*	234–40	yes
25	Biçici, Ergun & Andy Way	RTM-DCU: Referential translation machines for semantic similarity	2014	*Proceedings of Semantic Evaluation Exercises -- International Workhop on Semantic Evaluation (SemEval)14*		yes
26	Biçici, Ergun & Andy Way	Referential translation machines for predicting translation qualitz	2014	*Proceedings of the 9th Workshop on Statistical Machine Translation (WMT14)*	313–21	yes
27	Biçici, Ergun	Referential translation machines for quality estimation	2013	*Proceedings of the 8th Workshop on Statistical Machine Translation*	343–51	yes
28	Biçici, Ergun	Feature decay algorithms for fast deployment of accurate statistical machine translation systems	2013	*Proceedings of the 8th Workshop on Statistical Machine Translation (WMT13)*	78–84	yes
29	Beck, Daniel, Lucia Specia and Trevor Cohn	Reducing annotation effort for quality estimation via active learning	2013	*Proceedings of the 51st Annual Meeting of the Associaton for Computational Lingustics (ACL13)*	543–48	yes
30	Beck, Daniel, Kashif Shah, Trevor Cohn and Lucia Specia	SHEF-Lite: When less is more for translation quality estimation	2013	*Proceedings of the 8th Workshop on Statistical Machine Translation (WMT13)*	337–42	yes
31	Beck, Daniel, Kashif Shah & Lucia Specia	SHEF-Lite 2.0: Sparse multi-task Gaussian processes for translation quality estimation	2014	*Proceedings of the 9th Workshop on Statistical Machine Translation (WMT14)*	307–12	yes
32	Avramidis, Eleftherios	Efforts on machine learning over human-mediated translation edit rate	2014	*Proceedings of the 9th Workshop on Statistical Machine Translation (WMT14)*	302–6	yes
33	Almaghout, Hala & Lucia Specia	A CCG-based quality estimation metric for statistical machine translation	2013	*Proceedings of MT Summit XIV*	223–30	yes

4.1.2 Part A2. List of Dissemination Activities

Note that QTLaunchPad had no dedicated outreach work package. Dissemination and outreach was distributed throughout the various work packages, although some tasks (e.g., Task 6.5, Stakeholder Networks) focused more on these aspects. Much of the outreach and dissemination was conducted by subcontractors, notably GALA and, to a lesser extent, FIT. The table below lists the major dissemination activities.

NO.	Type of activity	Main Leader	Title	Date/Period	Venue/Location	Type of audience	Size of audience
1	Other (blog)	DCU	Translation Quality Models and Tools – Is There Room for Improvement?	2013-07-18	GALA Blog	Industry	1000+
2	Other (newsletter)	DCU	Mapping the Trends and Needs of the Translation and Localization Industry	2013-06-28	GALAxy Newsletter	Industry	1000+
3	Other (newsletter)	DCU	The Value of Effective Training in Translation and Localization	2013-09-01	GALAxy Newsletter	Industry	1000+
4	Other (newsletter)	DCU	Which Approach to Human Translation Quality Evaluation and Why?	2013-10-10	GALAxy Newsletter	Industry	1000+
5	Presentation	DFKI	Translation Quality Metrics for Human and Automatic Translation	2013-01-17	Tralogy Conference (Paris, France)	Research	
6	Presentation	DCU	Focussing on the End Users of Machine Translated Documentation: A Usability Study	2013-06-07	OPTIMALE Conference (Rennes, France)	Research	
7	Presentation	DFKI	Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment	2013-06-14	Localization Word (London, UK)	Industry	80
8	Presentation	DCU	Presentation of two papers	2013-06-14 2013-06-15	SemEval (International Workshop on Semantic Evaluation)	Research	300
9	Presentation	SHFU	Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation	2013-08-05	ACL 2013 (Sophia, Bulgaria)	Research	500
10	Presentation	SHFU	Reducing Annotation Effort for Quality Estimation via Active Learning	2013-08-06	ACL 2013 (Sophia, Bulgaria)	Research	500
11	Presentation	DFKI	From Quality Translation to Innovation and Commercial Services (session with four presentations)	2013-09-19	META-FORUM 2013 (Berlin, Germany)	Research, Industry	250
12	Presentation	DFKI	Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment	2013-09-26	Media4All (Dubrovnik, Croatia)	Research, Industry	80
13	Presentation	DFKI	Multidimensional Quality Metrics	2013-10-09	TAUS Translation Quality Summit (San Jose, California, USA)	Industry	40
14	Presentation	DCU	Cracking the Language Barrier	2013-11-06	ICT 2013 (Vilnius, Lithuania)	Research, Industry, Policy Makers	5000+
15	Presentation	DFKI	Measuring Translation Quality in Today's Automated Lifecycle	2013-11-07	Tekom (Wiesbaden, Germany)	Industry	100
16	Presentation	DFKI	A Unified Model for Document and Translation Quality Assurance	2013-11-09	Tekom (Wiesbaden, Germany)	Industry	80
17	Presentation	DFKI	Multidimensional Quality Metrics: A Flexible System for Assessing Translation Quality	2013-11-28	ASLIB Translating and the Computer (London, UK)	Research, Industry	75
18	Presentation	DFKI	On the Road to Quality: Translation for the 21st Century (QT21)	2014-03-26	GALA Conference 2014 (Istanbul, Turkey)	Industry	50

NO.	Type of activity	Main Leader	Title	Date/Period	Venue/Location	Type of audience	Size of audience
19	Presentation	DFKI	Multidimensional Quality Metrics (MQM): A New Framework for Translation Quality Assessment	2014-04-23	JIAMCATT 2014 (Strasbourg, France)	Policy Makers	100
20	Presentation	DFKI	Quality Models, Linked Data and Standardisation Efforts for a Multilingual and Linked Web	2014-05-07	Seventh MultilingualWeb Workshop (Madrid, Spain)	Research, Industry	100
21	Presentation	DCU	CNGL Scientific Expo	2014-05-13 2014-05-14	CNGL Scientific Expo (Dublin, Ireland)	Research	50
22	Presentation	DCU	Relating Translation Quality Barriers to Source-Text Properties	2014-05-26	MTE Workshop at LREC 2014 (Reykjavik, Iceland)	Research, Industry	36
23	Presentation	DFKI	Assessing Inter-Annotator Agreement for Translation Error Annotation	2014-05-26	MTE Workshop at LREC 2014 (Reykjavik, Iceland)	Research, Industry	36
24	Presentation	DFKI	Using a New Analytic Measure for the Annotation and Analysis of MT Errors on Real Data	2014-06-17	EAMT 2013 (Dubrovnik, Croatia)	Research, Industry	
25	Presentation	DFKI	Relations between Different Types of Post-Editing Operations, Cognitive Effort and Temporal Effort	2014-06-18	EAMT 2013 (Dubrovnik, Croatia)	Research, Industry	
26	Webinar	DFKI	Multidimensional Quality Metrics	2013-02-21	GALA Webinar	Industry	80
27	Webinar	DFKI/ USHF	Multidimensional Quality Metrics Version 2 and Software Infrastructure	2013-07-06	GALA Webinar	Industry	40
28	Webinar	DCU	Understanding and Implementing Effective Translation Quality Evaluation Techniques	2013-10-03	GALA Webinar	Industry	46
29	Webinar	DCU	Effective Post-Editing in Human and Machine Translation Workflows: Critical Knowledge and Techniques	2013-12-02 2013-12-07	GALA Webinar	Industry	138/16
30	Webinar	ILSP	Discovering, Sharing and Processing Language Data: One Step Beyond	2014-02-17	GALA Webinar	Industry	48
31	Webinar	DFKI	Using MQM for translation quality evaluation and annotation	2014-02-27	Webinar	Industry	25
32	Webinar	mittag-qi[25]	Web-based Proofreading and Post-Editing with translate5 and MQM	2014-06-05	GALA Webinar	Industry	39
33	Workshop	DFKI	Workshop on “Quality Metrics”	2013-03-14	Sixth MultilingualWeb Workshop (Rome, Italy)	Industry, Research, Policy Makers	20

[25] As subcontractor to DFKI

NO.	Type of activity	Main Leader	Title	Date/Period	Venue/Location	Type of audience	Size of audience
34	Workshop	DCU/DFKI	Workshop on “Quality Metrics”	2013-03-19 2013-03-20	GALA Conference 2013 (Miami, USA)	Industry	50
34	Workshop	DFKI	QTLaunchPad “Roundtorial” on “An Open Model for Assessing and Estimating Translation Quality”	2013-06-12	Localization Word (London, UK)	Industry	50
36	Workshop	DCU	Full-day Workshop: “User-Centric MT and Evaluation”	2013-08-03	Machine Translation Summit XIV (Nice, France)	Research, Industry	55
37	Workshop	DCU	Half-day pre-conference workshop on “Translation Quality Assessment”	2014-03-23	GALA Conference 2014 (Istanbul, Turkey)	Industry	50
38	Workshop	USHF	MTE 2014: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation	2014-05-26	LREC 2014 (Reykjavik, Iceland)	Research, Industry	36
39	Workshop	ILSP	Half-day workshop on “Legal Issues in Language Resources and Infrastructures”	2014-05-27	LREC 2014 (Reykjavik, Iceland)	Research	50
40	Workshop	DCU	QTLaunchPad Pre-conference Workshop on “Quality Translation: Where Are We Now, and Where Are we Going”	2014-06-15	EAMT 2014 (Dubrovnik, Croatia)	Research, Industry	30
41	Workshop	USHF	WMT 9 Shared Task on Quality Estimation	2014-06-26 2014-06-27	52nd ACL (Baltimore, USA)	Research	
42	Presentation	DCU	It Is dangerous to Lean out of the Ivory Tower of Babel... Or Is It? Machine Translation, Quality and Post-Editing	2014-06-29	ProZ.com 2014 International Conference (Pisa, Italy)	Industry	115