# Towards a Systematic and Human-Informed Paradigm for High-Quality Machine Translation

**Aljoscha Burchardt, Kim Harris, Georg Rehm, Hans Uszkoreit**
DFKI GmbH
Alt-Moabit 91c, Berlin, Germany
{firstname.lastname}@dfki.de

**Abstract**

Since the advent of modern statistical machine translation (SMT), much progress in system performance has been achieved that went hand-in-hand with ever more sophisticated mathematical models and methods. Numerous small improvements have been reported whose lasting effects are hard to judge, especially when they are combined with other newly proposed modifications of the basic models. Often the measured enhancements are hardly visible with the naked eye and two performance advances of the same measured magnitude are difficult to compare in their qualitative effects. We sense a strong need for a paradigm in MT research and development (R&D), that pays more attention to the subject matter, i.e., translation, and that analytically concentrates on the many different challenges for quality translation. The approach we propose utilizes the knowledge and experience of professional translators throughout the entire R&D cycle. It focuses on empirically confirmed quality barriers with the help of standardised error metrics that are supported by a system of interoperable methods and tools and are shared by research and translation business.
**Keywords:** Machine Translation, Platforms, Human Evaluation

## 1. Introduction

Since the advent of modern statistical machine translation (SMT), much progress in system performance has been achieved that went hand-in-hand with ever more sophisticated mathematical models and methods. Numerous small improvements have been reported whose lasting effects are hard to judge, especially when they are combined with other newly proposed modifications of the basic models. Often the measured enhancements are hardly visible with the naked eye and two performance advances of the same measured magnitude are difficult to compare in their qualitative effects. On the other hand, most of the fundamental known barriers to MT quality have not yet overcome.

We sense a strong need for a paradigm in MT research and development, that pays more attention to the subject matter, i.e., translation, and that analytically concentrates on the many different challenges for quality translation. The approach we propose utilizes the knowledge and experience of professional translators throughout the entire R&D cycle. It focuses on empirically confirmed quality barriers with the help of a standardised parameterisable error metric. The metric is supported by a system of methods and tools and shared by research and translation business. These components, which have already been created and tested, are seen as core components of an envisaged cloud-based platform, which will be sketched out in the last part of the paper.

The remainder of this paper explains these ideas in more detail.

## 2. Human-Informed MT Development Cycle

The prevalent (S)MT development cycle consists of a number of experiments in which system parameters, feature sets, preprocessing steps, etc. are more or less systematically varied followed by a testing phase ("generate-and-test").

The power of SMT lies in its massive utilization of human translation expertise. In rule-based systems only those parts of human knowledge are used that could be encoded in the dictionaries and rule sets of the system usually a mix of intellectually compiled explicitly stated linguistic regularities and exceptions.

Statistical methods acquire implicit human knowledge about translation and linguistic well-formedness by learning huge numbers of patterns from texts, especially from translated texts in connection with their source texts. In this way they can model semantic and stylistic preferences and constraints that could not be encoded in any of the hand-crafted rule systems.

Within the testing phase human knowledge is again being used in a rather indirect and implicit way, i. e., by comparing the output of the MT engine with one or more human reference translations using simple mathematical measures such as BLEU.

If one has the goal of working towards High-Quality Machine Translation (HQMT), this approach is scientifically questionable at best, for a number of reasons including:

- It is widely known that simple automatic measures such as BLEU correlate only mildly with translation quality. If we rely on them, only optimising our systems towards BLEU scores exclusively, we run the risk of reporting spurious improvements, under- or overestimating system variants, oscillating on plateaus, etc.
- It has been shown that the highest BLEU improvements are often made on segments that are unintelligible anyway, i. e., completely unintelligible translations get a little less unintelligible, but, nevertheless, they remain unintelligible (but the BLEU score is improved). This approach does not contribute to the goal of working towards high-quality translation.
- BLEU relies critically on one or more reference translations used for the comparison. We have performed an internal study using a Chinese → English reference corpus comprising 11 documents (1000 sentences),

each translated by a different human expert to evaluate an online MT system. The results are startling: the choice of *one* reference document led to a variation in BLEU scores of up to 7.64 points, depending on which reference was chosen (average BLEU: 18.11). Using all 11 reference translations together led to a BLEU score of 53.42.

- While higher BLEU scores *indicate* improved translation quality, they cannot be taken as *scientific evidence*.
- Single scores do not provide many (scientific) insights. Tuning and optimisation steps are usually epiphenomenal. They are not suitable to generalise results, to apply them to new translation tasks, or to make predictions.
- BLEU scores do not detect errors, nor do they provide any information on the type or source of errors.

All researchers should be eager to analyse the results of their experiments as thoroughly as possible in order to compare them to the work of others and to be in the best possible position to generate hypotheses for improving the approach and to drive future experiments. There are two classes of quality indicators: (i) Translation errors and (ii) cases where a generated correct translation is better or worse that another possible correct translation. Whereas (ii) is rather important for human translation in the highest quality segment, for today's MT only errors matter.[1] So far we do not have any reliable ways for automatically detecting errors and their error types. Thus we are convinced that any serious attempt to improve translation quality must include feedback by human experts such as translators and linguists early in the development process. This feedback can be given, e. g., in the form of post-edited (i.e. corrected) translations or explicit annotation of errors using standardised markup. In the language industry, both approaches are established best practice for assessing (machine) translation quality.

If the MT research community wants to produce research results that are supposed to be meaningful to the language industry, we have to extend our approaches, systems and paradigms in such a way as to be able to assess and report translation quality in the required way (in addition to what is needed in the SMT R&D cycle, i.e., automatic scores like BLEU). Figure 1 shows how the current SMT development cycle can be extended to include human feedback. The blue box represents the typical existing SMT development cycle. In addition, we propose to include at certain intervals or checkpoints a language expert who inspects the translation results, annotates and classifies errors and provides feedback to the MT developer who then starts another development cycle based on the insights gained. At some point, the language expert will most probably compare newly generated translations to previous output to see if the intended improvements have materialised, to check if there have been any unintended side-effects, and to spot the most pressing quality barriers.

This proposed approach makes it necessary to reserve a budget for human language professionals in MT projects (and to make sure that the human analysis and annotation process is optimally supported by tools), but we are convinced that this investment will pay off. The data gathered from human analysis and annotation should be used to build linguistically informed methods for quality estimation and error detection to eventually support (semi-)automatic analytic workflows.

## 3. High-Quality Translation Paradigm

The use of MT is increasingly popular for 'gisting' (information-only translation) through free online systems such as Google Translate or Bing Translator. These services have created huge new markets for translation. Already back in 2012, Google alone automatically translated as much content in a single day as all professional translators combined in an entire year, and was used by more than 200 million people every month[2] – by now the usage figures are surely much higher.

Still, all popular, freely available online translation services follow the "one size fits all" approach, i.e., they are not customisable. This approach is inherently incompatible with Europe's pressing demand for being able to produce large volumes of high-quality outbound translations either fully automatically or through human translators supported by machines. In this high-quality scenario, MT has to behave much more like Translation Memories (TMs) that are widely used in translation industry, especially when dealing with repetitive material such as technical documentation. TMs support translators by suggesting perfect or almost perfect translations based on previous translations. The translator can then accept and edit the suggestion or translate from scratch.

Already in past publications such as, for example, (Rehm and Uszkoreit, 2013; Burchardt et al., 2014; Popović et al., 2014), we have made the point of breaking out of the dead-end the MT research landscape is currently trapped in by advocating a paradigm shift. Instead of only adjusting known SMT algorithms and features to produce marginally better results, we call for a different approach of carrying out MT research in Europe, an approach that addresses the goal of producing quality translations and that takes into account very thoroughly the needs and priorities of European MT and Language Service Provider (LSP) companies, thus initiating a close collaboration for creating new breakthroughs in research and business opportunities at the same time.

A trivial, yet far-reaching insight is that not all translations are equally useful for human translators. For simplicity, they are often divided into the three discrete classes of (i) error-free translations, (ii) translations that can efficiently be post-edited and (iii) translations that are so bad that they would not help a human translator. While class (iii) might

[1] BLEU score measurement also punishes correct translations if they differ from the reference translations, may they be better or worse.

[2] http://googleblog.blogspot.de/2012/04/breaking-down-language-barriersix-years.html

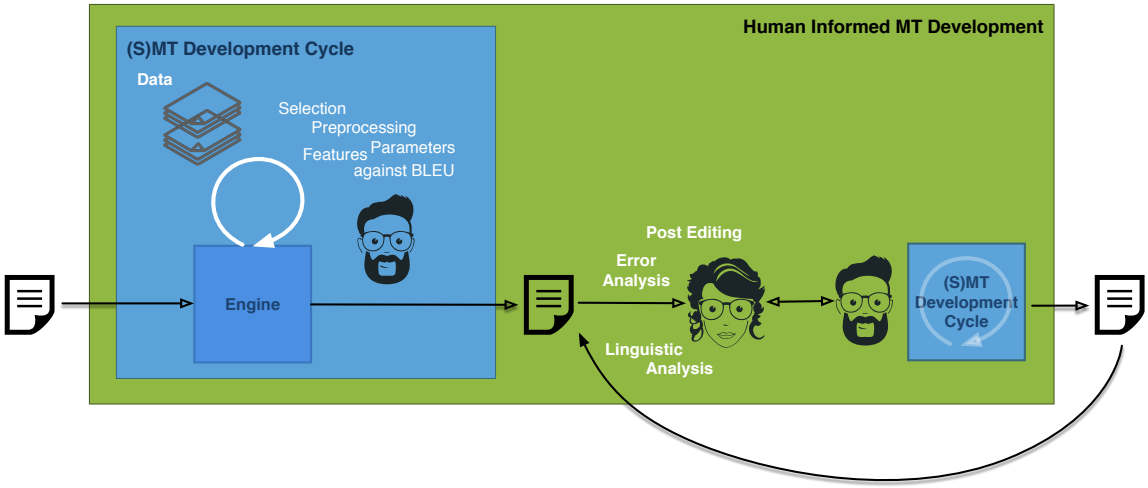


Figure 1: Human-informed MT development cycle

still provide some guidance to people in an information search scenario (gisting), they do not play a role in the quality translation scenario.

As a consequence, *improvement* in this paradigm, say from system variant $A$ to variant $A'$, requires that the proportion of the quality classes changes so that we have more translations of the "better" classes in the end. Figure 2 provides one such example where there has been an increase in the number of perfect, error-free translations. The precise criteria like error types, severity classes, scoring model, etc. need to be worked out between research and industry taking into account task-specific factors such as the language pair, document type, domain, target audience, etc.

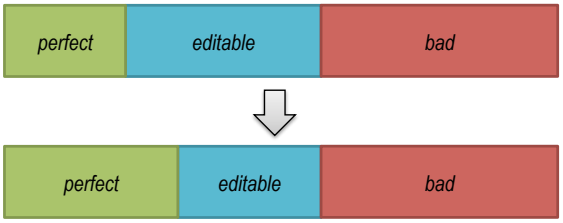


Figure 2: Example of improvement in hiqh-quality translation paradigm

This focus on HQMT requires new diagnostic methods. We provide some suggestions in the next section.

## 4. Standardised Error Metrics and Benchmarks

HQMT relies on improved translation models that must be based on novel, reliable and informative quality measures. Simplistic common measures such as BLEU or edit-distance based measures such as TER may even incorrectly punish perfectly adequate translations which differ from a given reference (or references), e. g., in completely legitimate word order and morphological realisation. Currently, the only way of assessing translation quality with an adequate level of reliability and granularity (word/phrase level) necessarily involves intellectual work such as post-editing or explicit error annotation.

In the new type of MT development, annotations can be added on the following three levels as needed:

**Phenomenological level** Annotation of issues in the translated output (target side) with translation errors such as, e. g., `Omission`, `Terminology`, or `Grammar`.

**Linguistic level** Annotation of the translation source or target side with information like part of speech, phrase boundaries or more specific phenomena under consideration such as long-distance dependencies or multi-word expressions.

**Explanatory level** Annotation of the source (also referencing the target) with (typically speculative) reasons for translation failure such as model class, $n$-gram size, data sparseness, etc.

The annotation on the phenomenological level usually involves language professionals like human translators while the other two levels require linguistic skills and expertise on the MT system level that researchers from linguistics, language technology, and related areas typically have.

**Standardised error markup with MQM** While the notions "error markup" and "issue markup" are often used interchangeably, there is an important difference that we only briefly sketch in this article. There is no transcendent, absolute notion of translation quality. Thus, an issue such as an inconsistency in terminology, for example, referring to an object as "PC" in one sentence and as "computer" in another, might be counted as an error, e. g, in a reference manual, but it can be perfectly acceptable, maybe even preferred, in a newspaper article. Translation quality is always relative to the intended communicative purpose and context that can best be captured in a formal specification. The Multidimensional Quality Metrics (MQM) (Lommel et al., 2014) is based on this principle of flexibility to translate different purposes into dimensions and selective subsets of

issues to be checked (and weighted). MQM was designed around a master vocabulary comprising 100+ issue types for describing task-specific metrics in a highly customisable way. It provides a unified approach for (diagnostic) evaluation of MT with approaches used for human translation quality checking in industry. It was designed as a non-strict superset of prominent metrics (LISA QA Model, SAE J2450, ATA certification, etc.). An early version was standardised in the W3C recommendation ITS 2.0.

**Post-Editing** Apart from direct error annotation, one can also make use of human feedback on the basis of automatically translated output that was corrected by human translators through post-editing. This output can be used to update, reinforce, and correct systems' translation hypotheses and together with explicit error markup will help to overcome real quality barriers and also fix relatively minor issues such as punctuation or agreement errors that seem to have been over-looked in the development of MT engines for gisting, yet render most output improper for outbound translation.

**Evaluation workflow** Projects such as QTLaunchPad and QT21 have developed valuable experience for what we hope becomes best practice in future evaluation scenarios. Given some MT translated corpus and initial hypotheses of what issues may be encountered, the following steps are included in an example evaluation workflow:

1. Definition of a concrete metric for the given purpose starting from an existing metric ("benchmark") or from scratch.

2. Filtering the translation corpus to be evaluated in a triage:

   a. Perfect translations.

   b. Almost good translations that need further analysis.

   c. Bad translations that do not qualify for further inspection.

   These steps can be performed manually, in a semi-automatic or even in an automatic way using sampling and filtering strategies or with the help of a quality estimation toolkit such as QuEst ((Shah et al., 2013)), depending on the size of the corpus, available human resources, and required precision and recall. What follows is:

3. Annotation/Post-editing of the segments of type b.

4. Inspection of the errors/edits to:

   - Confirm if the system output supports the hypotheses
   - Get a quantitative basis to decide on MT development priorities
   - Get a qualitative idea of remaining quality barriers

Figure 3 illustrates this proposed workflow. It is advisable to perform error annotation and post-editing in parallel so that analysis and correction are handled in a consistent manner as there are usually multiple ways of analysing and fixing a translation error.

**Test Suites** Test suites are a best practice instrument in areas such as grammar checking, to ensure that a parser is able to analyse certain sentences correctly or test the parser after changes to see if it still behaves in the expected way.
In the context of HQMT, we use the term "test suite" to refer to a selected set of input-output pairs that reflects interesting or difficult, error-prone cases. Test suites have not generally been used in MT research. Reasons for this might include the theoretical issue that there is no eternal notion of "good translation" and the more practical issue that there are usually many different good translations for a given input. Even if one could assume the existence some gold-standard translation, there would be no simple notion of deviation that could be used. In line with what we have argued for above, human analysis will be needed for evaluating MT performance on test suites.
Nevertheless, we think that testing system performance on empirically grounded error classes will lead to insights that can guide future research and improvements of systems. By using suitable test suites, MT developers will be able to see how their systems perform compared to scenarios that are likely to lead to failure and can take corrective action, e. g., by creating targeted training corpora focussing on certain error types. Test suites can also be the basis for new types of benchmarks and shared tasks that are based on empirically attested quality barriers; at the time of writing, we are working on a test suite for the language pair German–English, which will be published in 2016.

## 5. Integrated MT Development Platform

MT research has contributed to the development of a large set of tools required to build MT systems, training data, and evaluating corpora. These resources exist in various locations and often require substantial IT and system development skills to put them together and to make them work in an operating environment. As a result, most resources are not being reused to the extent they should be, some are not being reused at all after the end of the project in which they were created. Some of the tools that could have been useful outside the R&D community, especially to language service providers (LSPs), have appealed primarily to researchers and computer scientists rather than to language professionals and, thus, their use in and impact on the language industry has remained limited at best. Even the large volumes of valuable data accumulated over the years by the Workshop on Statistical Machine Translation (WMT) community (primarily in the projects EuroMatrix, EuroMatrix-Plus, MosesCore and CRACKER) have mostly been stored in hundreds of unconnected text files that are hard to search and to combine.
The field of leading-edge MT R&D has reached a level of complexity with its workflows, networks of people and communities, as well as resources and components involved that it is about time to discuss the pros and cons of an integrated development platform. An integrated

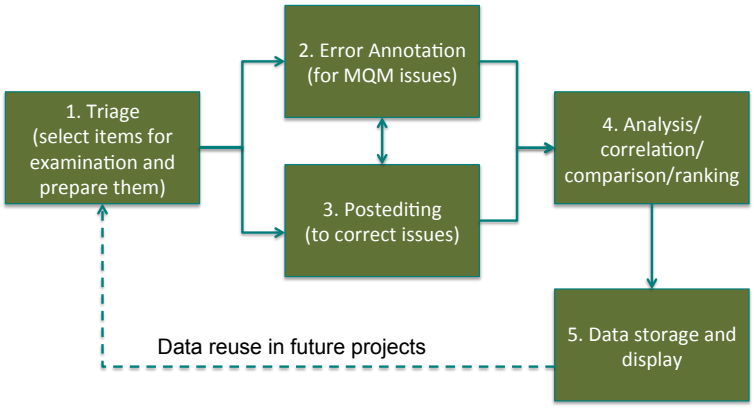


Figure 3: Post-Editing and Error Annotation Workflow Example

development environment will be even more necessary when we look at the *additional* ingredients needed for the human-informed and quality-driven MT paradigm we have sketched above, which will add to the already complexity no doubt. If designed the right way, the integrated development platform we have in mind will have positive effects on MT technology evolution by speeding up the search and evaluation cycles clustered around shared tasks and collectively approached challenges. A technology infrastructure for a major research effort in HQMT needs to serve several major purposes:

- It should help research groups to develop, test and demonstrate their new methods in realistic workflows with existing state-of-the-art components, realistic tasks, and benchmarks. It should also help them to compare the performance of their components with the best existing technology.
- It should support shared research and shared evaluations by providing the data, environments, workflows, and evaluation tools for collective system building and collective comparative assessment campaigns.
- It should preserve, document, administer and provide data, technologies and evaluations for new research groups, potential users, research planners, funding agencies and the media.
- It should provide access to state-of-the-art tools for functions and processes that may lie outside the core competency of any specific group.

To this end, the integrated platform needs to go beyond existing resources such as the open resource exchange infrastructure META-SHARE (Piperidis et al., 2014), the open source tools and core components for building SMT systems like Moses or Jane (Koehn et al., 2007; Vilar et al., 2010); tools for quality estimation and error analysis such as Qualitative (Avramidis et al., 2014), MTComparEval (Klejch et al., 2015) or Hjerson (Popović, 2011); web platforms for selecting or training MT systems like iTranslate4EU[3] or Let'sMT[4]; CAT tools and workbenches like MateCat and Casmacat (Federico et al., 2014; Alabau et al., 2014); Post-Editing and error-annotation tools like PET (Aziz et al., 2012) or MT-Equal (Girardi et al., 2014); the web-based service collection of PANACEA (Poch et al., 2012), or the accumulated data, tools, and scripts of the WMT shared-task repositories. The targeted infrastructure should incorporate as many existing useful resources as possible instead of rebuilding components, data, tools and service platforms. It must enable truly collaborative research and make resources more accessible to all.

Figure 4 provides an overview of the kind of platform we envisage. In the center there is a core research infrastructure that consists of a data repository that is connected through a back-end to a front-end that takes over the function of the overall cockpit. This core infrastructure should be linked to the different services, tools, data sources, workflows, and stakeholders included in the figure in coloured boxes. Note that the content of these boxes is not fully exclusive. For now, we have left it underspecified what services can and will actually be hosted by the backend and what should be accessed via APIs.

An important part of the cockpit is data management, i. e., a data model together with data collections, DB user interfaces and data maintenance tools for the management of MT-related data collections. For now, we propose a simple data model and suggest to base a first iteration of the cockpit on the the existing open source tool translate5.

### 5.1. Data Model

For the envisaged platform, we suggest to define a very general relational model for MT-relevant data including training data, reference data, benchmark data, evaluation results and test suites. This model can be employed for designing databases for existing and new data sets of various types. It also supports user interfaces for typical viewing and editing tasks.

The proposed data model is simple and versatile. It picks up an original idea of Harris (Harris, 1988) who proposed to store bi-texts in databases whose two dimensions are the

[3] http://itranslate4.eu/en/
[4] http://www.letsmt.eu

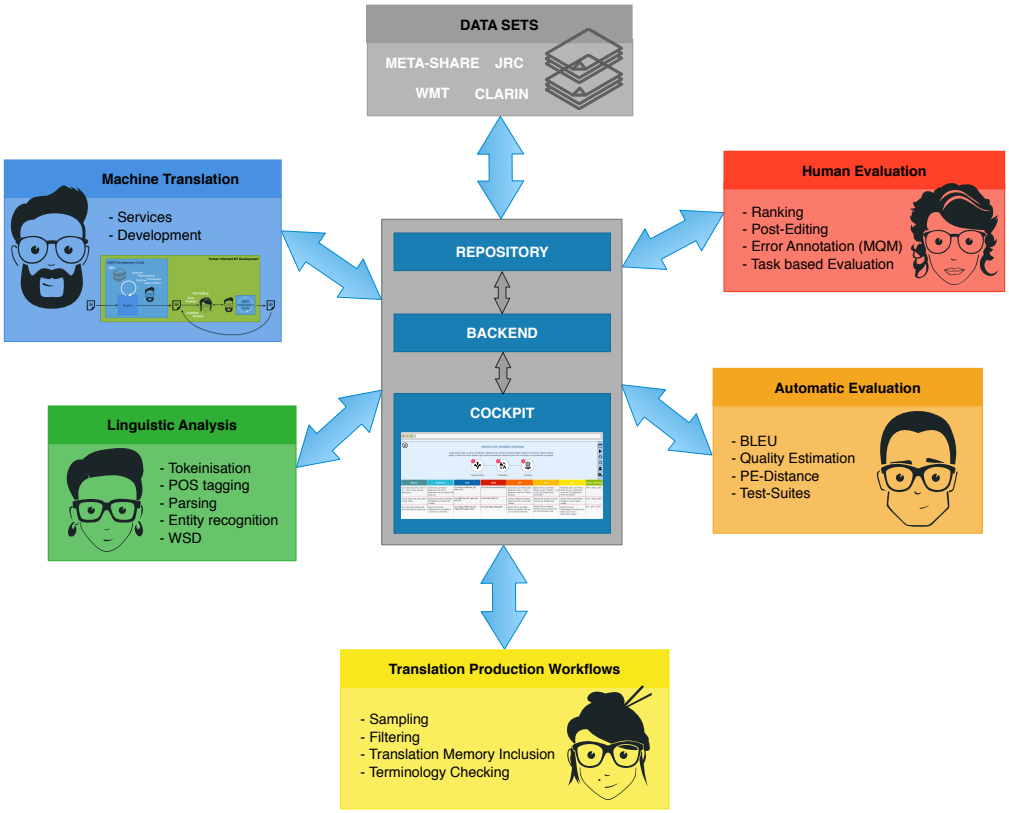


Figure 4: Integrated MT Development Platform

segments and the languages. Assume a tabular database in which every row contains all the translations and annotations of a segment. The columns are dedicated to different translations, versions and annotations of the segments. Closest to the original idea of bi-texts are columns for a translation of a text into another language. However, columns could also be used for storing edited versions of segments or texts. They can hold alternative translations by human experts or machines into the same target language. Columns can also hold comments such as assessment scores or annotations referring to the segments in another column, such as POS tagging, syntactic structures, or marked errors. They can accommodate stand-off annotation consisting of lists of mark-up tags with their respective scopes given in offset notation. However, they can also contain in-line markup applied to a copy of the raw text in another column. Finally, they can also be used to annotate relational information pertaining to two or more columns such as alignments or comparative quality rankings. A relational database management system would be needed to prevent overly complex databases. Assigned primary keys will facilitate linking and joining of the tables.

Authoring, translation, assessment and annotation can be conceptualised and realised as entering data into a new column. Pre-and post-editing could either be realised as editing data in an existing column or in a new, copied column, depending on the interest in documenting the editing step. In keeping with the user- and human-centred research paradigm, such a user interface would be suited for translation professionals who are used to working with similar (albeit less powerful) interfaces in many computer-assisted translation (CAT) tools. Figure 5 provides a mockup of an example workflow including preprocessing, translation, and human ranking results.

Modelled workflows in translation management can be easily supported by assigning and removing read/write privileges and by appropriate reporting functions. The same is true for workflows in collective research such as multi-site system testing and in shared evaluation tasks such as the application of alternative systems to the same texts or competitive quality assessments. Such workflows can include the evaluation tasks and the realisation and testing of second-order translation systems such as combos. In addition, full versioning of data sets will ensure that users will be able to trace the complete provenance of all data; in current workflows, multiple different versions of resources may be in circulation leading to situations in which it is not clear after the fact which version of a resource was used to generate another resource.

Just as several general architectures for text analytics use layers of annotation as the output interfaces between the modules, a general architecture for MT built on our data model would use new columns to display results.

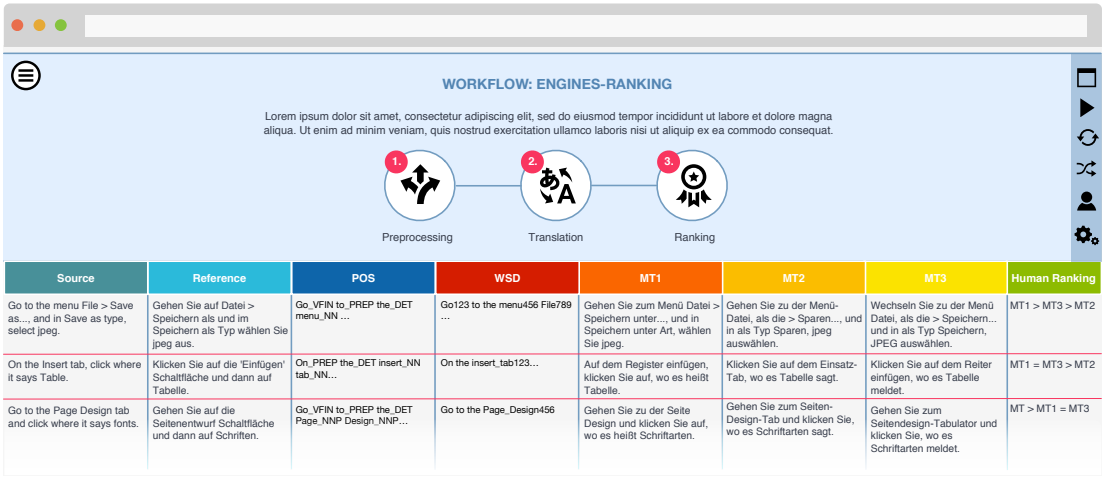


Figure 5: Mockup of the infrastructure's cockpit

### 5.2. Translate5

Translate5, implemented by MittagQI is a database-driven tool with a graphical user interface that implements the column-based data model sketched above (see Figure 6 for a screenshot).

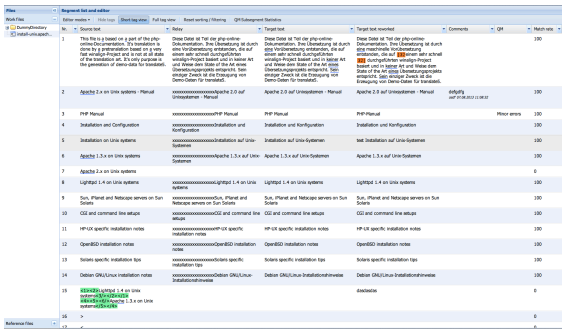

Figure 6: translate5 column view

It was originally implemented as a proofreading and post-editing environment for the translation industry. It uses a MySQL database, in which source texts, translations and annotations are stored. In the QTLaunchPad project we extended the tool to support MQM error annotation. In order to stress-test the system with large amounts of data, we imported all WMT data from 2008 to 2014 into translate5 without encountering any performance issues.

Translate5 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 first set of APIs has been implemented as well as queuing, management and load-balancing of external and internal processes including dependency management. Simple reporting functions are already in place, others will follow soon. Translate5 features user administration, as well as simple workflow specification facilities and client management functions. Translate5 currently supports post-editing, MQM error tagging, and simple ranking. Further improvements of the feature set will be provided with support of the project CRACKER.

## 6. Conclusion

In this paper, we have argued that research and development in Machine Translation has to make a more direct use of the knowledge of language experts such as translators and linguists. To this end, we suggest a human-informed development cycle that works on empirically confirmed quality barriers with the help of standardised error metrics and benchmarks.

As the technical foundation for a new kind of intensified collaboration between MT developers and language professionals, we outline a platform that assembles a system of methods and tools that are shared by research and the translation industry in MT R&D activities. One open source tool that could serve as the nucleus for this envisaged paradigm is translate5 that has been extended to support MQM error markup in the QTLaunchPad project and is currently further developed with support of the CRACKER project (Rehm, 2015).

In fact, some of the currently running European projects like QT21 and QTLeap are already implementing certain aspects of the emerging paradigm by including human annotation and evaluation into the MT development methodology, supported by CRACKER. Yet, we are convinced that implementing the vision put forward in this paper requires substantial support, both in terms of willingness on the side of the research community and in terms of support on the side of funding agencies and policy makers. The support of this quality-driven and analytical approach to MT development we see in industry is a step in the right direction.

## Acknowledgements

This article has received support from the EC's Horizon 2020 research and innovation programme under grant agreements no. 645452 (QT21) and no. 645357 (CRACKER). We thank the three anonymous reviewers for their valuable comments.

## 7. Bibliographical References


Alabau, V., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., Germann, U., González-Rubio, J., Hill, R. L., Koehn, P., Leiva, L. A., Mesa-Lao, B., Ortiz-Martínez, D., Saint-Amand, H., Sanchis-Trilles, G., and Tsoukala, C. (2014). CASMACAT: A computer-assisted translation workbench. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 25–28.

Avramidis, E., Poustka, L., and Schmeier, S. (2014). Qualitative: Open source python tool for quality estimation over multiple machine translation outputs. *The Prague Bulletin of Mathematical Linguistics*, 102:5–16, 10.

Aziz, W., Castilho, S., and Specia, L. (2012). PET: a Tool for Post-editing and Assessing Machine Translation. In *Eighth International Conference on Language Resources and Evaluation*, pages 3982–3987, Istanbul, Turkey.

Burchardt, A., Lommel, A., Rehm, G., Sasaki, F., van Genabith, J., and Uszkoreit, H. (2014). Language Technology Drives Quality Translation. *MultiLingual*, (143):33–39, April. Issue April/May.

Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M., Cattelan, A., Farina, A., Lupinetti, D., Martines, A., Massidda, A., Schwenk, H., Barrault, L., Blain, F., Koehn, P., Buck, C., and Germann, U. (2014). The matecat tool. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference System Demonstrations, August 23-29, 2014, Dublin, Ireland*, pages 129–132.

Girardi, C., Bentivogli, L., Farajian, M. A., and Federico, M. (2014). MT-EQuAl: a Toolkit for Human Assessment of Machine Translation Output. In *COLING 2014*, pages 120–123, Dublin, Ireland, August. Dublin City University and ACL.

Harris, B. (1988). Bi-text, a new concept in translation theory. *Language Monthly*, 54:8–10.

Klejch, O., Avramidis, E., Burchardt, A., and Popel, M. (2015). MT-ComparEval: Graphical evaluation interface for machine translation development. *The Prague Bulletin of Mathematical Linguistics*, 104:63–74.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, Richard a nd Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lommel, A. R., Burchardt, A., and Uszkoreit, H. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0(12):455–463, 12.

Piperidis, S., Papageorgiou, H., Spurk, C., Rehm, G., Choukri, K., Hamon, O., Calzolari, N., del Gratta, R., Magnini, B., and Girardi, C. (2014). META-SHARE: One year after. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, May.

Poch, M., Toral, A., Hamon, O., Quochi, V., and Bel, N. (2012). Towards a user-friendly platform for building language resources based on web services. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Popović, M., Avramidis, E., Burchardt, A., Hunsicker, S., Schmeier, S., Tscherwinka, C., Vilar, D., and Uszkoreit, H. (2014). Involving language professionals in the evaluation of machine translation. *Journal on Language Resources and Evaluation*, 48(4):541–559.

Popović, M. (2011). Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 96:59–68.

Georg Rehm et al., editors. (2013). *The META-NET Strategic Research Agenda for Multilingual Europe 2020*. Springer, Heidelberg, New York, Dordrecht, London. Buy this book at springer.com or amazon.de.

Rehm, G. (2015). CRACKER: Cracking the Language Barrier. In Ilknur Durgar El-Kahlout, et al., editors, *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT 2015)*, page 223, Antalya, Turkey, May.

Shah, K., Avramidis, E., Biçici, E., and Specia, L. (2013). QuEst – design, implementation and extensions of a framework for machine translation quality estimation. *The Prague Bulletin of Mathematical Linguistics*, 100:19–30, 9.

Vilar, D., Stein, D., Huck, M., and Ney, H. (2010). Jane: Open source hierarchical translation, extended with reordering and lexicon models. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT 2010)*, pages 262–270, Uppsala, Sweden, July.