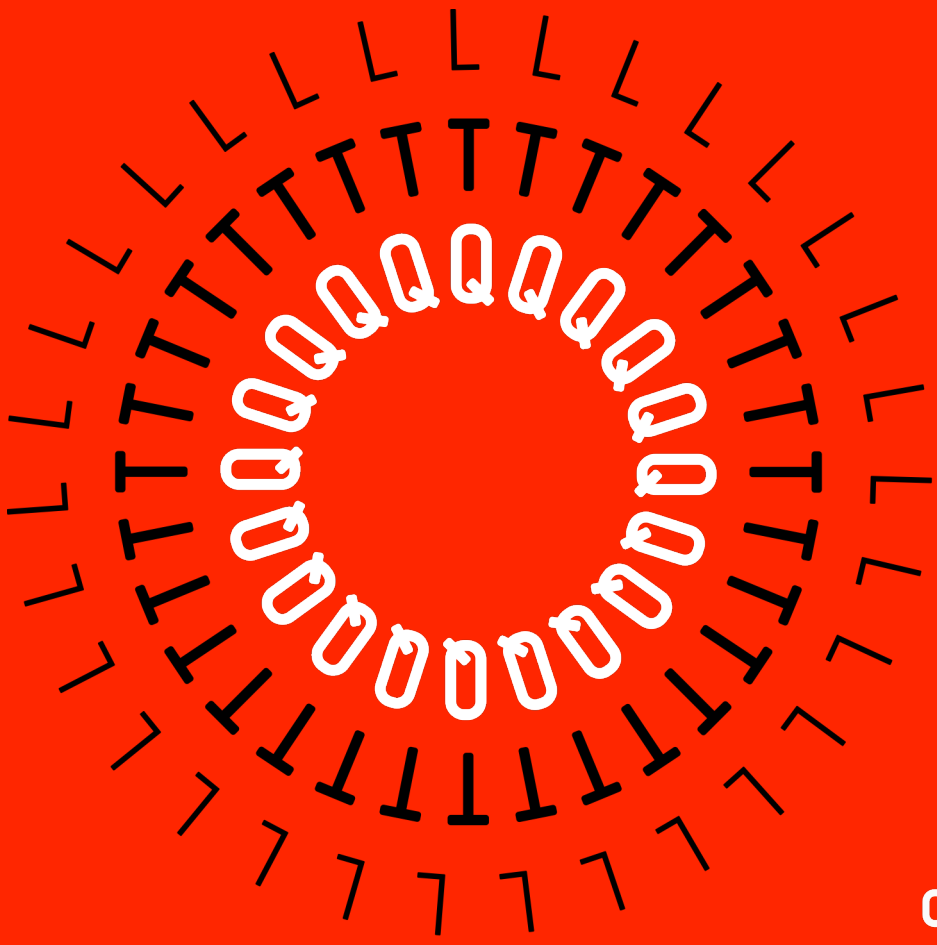


qtleap
quality translation by deep language engineering approaches

REPORT ON EVALUATION METRICS AND BASELINES FOR THE PROJECT

DELIVERABLE D2.2

VERSION 1.9 | 2015 June 15

QTLeap

Machine translation is a computational procedure that seeks to provide the translation of utterances from one language into another language.

Research and development around this grand challenge is bringing this technology to a level of maturity that already supports useful practical solutions. It permits to get at least the gist of the utterances being translated, and even to get pretty good results for some language pairs in some focused discourse domains, helping to reduce costs and to improve productivity in international businesses.

There is nevertheless still a way to go for this technology to attain a level of maturity that permits the delivery of quality translation across the board.

The goal of the QTLeap project is to research on and deliver an articulated methodology for machine translation that explores deep language engineering approaches in view of breaking the way to translations of higher quality.

The deeper the processing of utterances the less language-specific differences remain between the representation of the meaning of a given utterance and the meaning representation of its translation. Further chances of success can thus be explored by machine translation systems that are based on deeper semantic engineering approaches.

Deep language processing has its stepping-stone in linguistically principled methods and generalizations. It has been evolving towards supporting realistic applications, namely by embedding more data based solutions, and by exploring new types of datasets recently developed, such as parallel DeepBanks.

This progress is further supported by recent advances in terms of lexical processing. These advances have been made possible by enhanced techniques for referential and conceptual ambiguity resolution, and supported also by new types of datasets recently developed as linked open data.

The project QTLeap explores novel ways for attaining machine translation of higher quality that are opened by a new generation of increasingly sophisticated semantic datasets and by recent advances in deep language processing.

www.qtleap.eu

Funded by

QTLeap is funded by the 7th Framework Programme of the European Commission.

Supported by

And supported by the participating institutions:

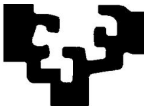
Faculty of Sciences, University of Lisbon

German Research Centre for Artificial Intelligence

Charles University in Prague

Bulgarian Academy of Sciences

Humboldt University of Berlin



University of Basque Country

University of Groningen

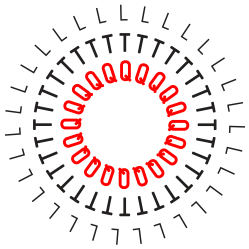
Higher Functions, Lda

Revision History

version	date	author	organisation	description
0.1	2014 APR 1	Aljoscha Burchardt	DFKI	First draft
1	2014 APR 9	Aljoscha Burchardt	DFKI	Prefinal including feedback from partners
1.1	2014 APR 9	Eleftherios Avramidis	DFKI	Addition of references and correction of typos and formatting
1.2	2014 APR 10	Eleftherios Avramidis	DFKI	Data, system descriptions and comments for Basque Bulgarian, Portuguiese, Spanish, Czech
1.3	2014 APR 11	Aljoscha Burchardt	DFKI	Included reviewing feedback from CUNI
1.4	2014 APR 30	Eleftherios Avramidis	DFKI	Addition for German and Dutch data, BLEU scores. Final version
1.5	2014 MAY 1	Eleftherios Avramidis	DFKI	METEOR scores, minor formatting issues
1.6	2015 June 8	Eleftherios Avramidis, Aljoscha Burchardt, Arle Lommel using input from all partners	DFKI	Revision after annual review
1.7	2015 June 10	Eleftherios Avramidis, Aljoscha Burchardt	DFKI	Included reviewing feedback from CUNI (Martin Popel)
1.8	2015 June 12	Eleftherios Avramidis	DFKI	References
1.8	2015 June 15	Aljoscha Burchardt	DFKI	Final comments by partners

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.



REPORT ON EVALUATION METRICS AND BASELINES FOR THE PROJECT

DOCUMENT QTLEAP-2014-D2.2
EC FP7 PROJECT #610516

DELIVERABLE D2.2

completion

FINAL

status

SUBMITTED

dissemination level

PUBLIC

responsible

JAN HAJIČ (WP2 COORDINATOR)

reviewer

MARTIN POPEL

contributing partners

DFKI, FCUL, CUNI, IICT-BAS, UPV/EHU, UG, HF

authors

ALJOSCHA BURCHARDT, ELEFTHERIOS AVRAMIDIS

Contents

1 Executive summary

The goal of the QTLeap project is to develop Machine Translation (MT) technology that goes beyond the state of the art in terms of "depth" of the methods and knowledge used (cf. Deliverable D2.1 for an overview of the state of the art in MT). As already described in the DoW, phrase-based statistical MT (SMT) systems will serve as baseline that will help to measure the effects of "getting deeper".

Our notion of "state of the art" is based on comparisons of openly documented (if possible through academic publications) and available systems, since systems that we do not have access to (e.g., Google Translate) may contain resources, components, and settings that make it hard to reproduce results. At the QTLeap Kick-Off meeting it had already been decided to use Moses as baseline engine. Three remaining issues that have been discussed in subsequent meetings were: i) what training (development/test) data to be used, ii) how to evaluate MT quality, and iii) how to ensure comparability of the baselines. These questions have been discussed at two virtual meetings ("WP 3 update meeting", December 20, 2013 and "MT Pilot 0 meeting", January 29, 2014) as well as at the face-to-face "Strategic project meeting" March 14/15, 2014. All meetings have been documented in the intranet. This document consolidates the results of the discussion.

This Deliverable documents part of Milestone B "Baselines and pipelines". It is closely related to D3.2 "Report on the workbench for developers" (due M9) and D3.3 "Report on the extrinsic evaluation metrics" (due M9). To avoid overlap, we will touch the respective issues only marginally here.

In the remainder of this document, we will describe the baseline systems and evaluation metrics. Finally, we will give an outlook on next steps.

2 Machine Translation Baselines ("Pilot 0")

The QTLeap usage scenario is provided by the partner HF (see, e.g, D3.1). In this IT helpdesk scenario, a user query in language X is translated into English, an answer is found in the English database, then translated into language X, and sent to the user. The more important translation direction in terms of quality is the outbound direction (English→X), which is in line with the European demand for better support of more languages. The query direction (X→English) will thus not receive the same attention in QTLeap, it can be seen as case of cross-lingual information retrieval rather than high quality MT. For the SMT baselines, however, both directions will be set up.

2.1 Systems

The responsible partners set up their own baselines individually according to the state-of-the-art settings for their language(s), e.g. by adopting best-performing settings from WMT13 (Bojar et al., 2013):

- FCUL: Portuguese↔English
- DFKI: German↔English

- CUNI: Czech↔English
- IICT-BAS: Bulgarian↔English
- UPV/EHU: Basque and Spanish↔English
- UG: Dutch↔English

After some discussion, it has been decided to not assume comparability across the languages and instead leave some freedom as to which corpora are being used (see below).

Still, all relevant system settings (for training and decoding), scripts, and links to the data[1] have been stored in the workbench/repository to ensure transparency and replicability (see D3.2).

The Moses [Koehn et. al, 2007] systems were set up in a phrase-based setup with the configurations documented in Table 1.

2.1.1 Generic notes for all language pairs

Defining the state-of-the-art settings upon the development of the phrase-based systems has been majorly based on the comparative development stages as published within the Shared Task of the Workshop in Machine Translation, as this is the basic event where such settings have been explored and published openly over the years. For those languages that have participated in the Translation Task of the Workshop on Statistical Machine Translation (WMT, see e.g., http://statmt.org/wmt15/), the existing configuration of the WMT state-of-the-art systems was followed, whereas for the rest of the languages, each partner adapted those configurations to their own setup, given the data available. For the latter, definition of "state of the art" is more difficult since there are few open publications on the building of systems on those languages, but we assume that following the successful settings from the other language pairs should suffice.

All systems have been trained based on a phrase-based model by Giza++ or mGiza with "grow-diag-final-and" symmetrization and "msd-bidirectional-fe" reordering (Koehn et. al, 2003). For the language pairs where big quantities of domain-specific monolingual data were available (Czech, Spanish and German) along with the generic domain data, separate language models (domain-specific and generic) were interpolated against our domain-specific development set (Koehn et. al, 2012). For LM training and interpolation, the SRILM toolkit (Stolcke, 2002) was used. The method of "Truecasing" (Koehn et. al, 2008) has been adopted for several language pairs where it proved useful.

Below, you find more details on the different project languages.

[1] Due to licensing issues, one corpus resource for Basque provided by the ABPU member Eleka was not meant to be shared. We did not want to reject this kind offer and used it.

	bg	eu	cs-en	en-cs	es	de-en	en-de	nl	pt
Pre-processing									
Sentence length limit	80	75	99	99	80	80	80	80	80
Tokenization	Moses tokenizer	Stanford CoreNLP, Eustagger	Treex	Treex	Moses tokenizer	Moses tokenizer	Moses tokenizer	Moses tokenizer	Moses tokenizer
Lemmatiza–tion	no		Treex	Treex	no	no	no	no	no
Casing	lowercase +recasing	lowercase +recasing	Treex	Treex	Moses trucaser	Moses truecaser	Moses truecaser	Moses truecaser	Moses truecaser
Compound splitting	no	no	no	no	no	Moses splitter	no	no	no
Training									
Factored model	No	no	yes	yes	no	no	no	no	no
LM order	5	5	6	8 [morph tags] 6 [full form]	5	5	5	5	5
LM training	SRILM	KenLM	SRILM	SRILM	SRILM	SRILM	SRILM	IRSTLM	SRILM
LM KN smoothing	yes	yes	yes	yes	yes	yes	yes	yes	yes
LM interpolation	no	no	SRILM+ MERT	SRILM+ MERT	SRILM	SRILM	SRILM	no	SRILM
Word-alignment	Giza++ on full forms	mGiza on lemmas	Giza++ on lemmas	Giza++ on lemmas	mGiza on full forms	Giza++ on full forms	Giza++ on full forms	Giza++ on full forms	Giza++ on full forms
Factor setup				target: Full form +morph tag					
Factor fallback	no	no	Lemma	no	no	no	no	no	no
Tuning	MERT	MERT	MERT	MERT	MERT	MERT	MERT	MERT	MERT

Table 1: State-of-the-art settings for all language pairs.

2.1.2 Basque

Not much research on MT for the English-Basque language pair has been published. Although there are two systems available for use (developed by Google and Lucy Software), they are proprietary software and are not accessible for research in general. Most research on Basque is focused on Spanish-to-Basque translation direction, where both rule-based [Mayor et al, 2011] and statistical approaches have been developed. Regarding the SMT approach, Basque segmentation [Díaz de Ilarraza et al., 2009-a) and Spanish pre-reorder rules [Díaz de Ilarraza et al, 2009-a] improved BLEU results on reduced corpora, but those improvements were not as significant when the size of the training corpus is bigger [Labaka, 2009].

Our Basque system uses language-specific preprocessing tools for tokenization, lemmatization and lowercasing. In particular, Stanford CoreNLP is used for the English side and Eustagger for the Basque one. The length threshold for filtering the training sentences has been adjusted to a maximum of 75 words per sentence in order to meet the language-specific length properties. Contrary to other language pairs, in order to achieve better performance, the word alignment is based on lemmas, which is then projected to lowercased full word forms for the rest of the training process. After translation a recasing process perform based on the tools available in Moses toolkit.

1.1.1 Bulgarian

A state-of-the-art baseline for the translation between English and Bulgarian is difficult to establish as there are only a very small number of publications dealing with this translation pair and it has to our knowledge not been included in any MT workshop so far. Therefore, our translation system baseline was built to closely follow the recommendations of WMT13, as with other less studied translation pairs. The translation system includes a standard phrase-based translation model and a 5-gram language model created with the SRILM toolkit (see Table 1).

Tools provided as part of the Moses toolkit were used to prepare the data for training, including limiting sentence length to 80 words, lowercasing, and tokenization. A Moses recaser model was trained and applied to the translation output for each translation direction.

1.1.2 Czech

Our Czech baseline system follows the CU-Bojar system [Bojar et al., 2013]**.** For Czech to English we trained a factored phrase-based model based on full truecased forms with the fallback to translate Czech lemma into English form if the Czech form is not known. Two LMs for English forms were used whereas their weights were optimized with MERT: a 6grams LM from the monolingual part of news and political corpora (see 2.2) and a 6grams LM from the English side of a bilingual Czech- English corpus. For English to Czech we trained a factored phrase-based model based on truecased forms translated directly to the pair <truecased form, morphological tag>. There were three LMs for Czech:

- 8grams of morphological tags from the monolingual part of news and political corpora,
- 6grams of forms from the monolingual part of news and political corpora and
- 6grams from the Czech side of a bilingual Czech-English corpus.

The pre-processing of this SMT system has been harmonized with the pre-existing version of Tecto-MT: Tokenization and lemmatization is handled by Treex followed by further tokenization at any letter-digit-punctuation boundary (e.g. hyphenated words or words that mix numbers and letters, e.g. 2good -> '2 good'). Additionally, casing is handled by a Czech-specific 'supervised truecasing' method. The output of the lemmatizer is used, as names have lemmas capitalized, the the casing of the lemma is cast to the token (lowercasing non-names at sentence beginnings, lowercasing also ALL CAPS if correctly lemmatized). Finally, the translation is done using case-sensitive tokens and finally the first letter in every sentence is only capitalized.

1.1.3 Dutch

There is no active work on SMT for Dutch, to the best of our knowledge, so our system incorporates successful settings from other language pairs and development efforts focused more on the data used.

For the creation of the baselines for Dutch-English and English-Dutch, the Moses baseline system was trained on several corpora, containing both general vocabulary as domain data to see which one would yield the best BLEU score. Each model was tested on 1% of the corpus and on the HF-development data. Models trained on corpora that gave good BLEU scores were then combined for the creation of a new system to see if this would increase the scores. The Dutch Parallel Corpus in combination with KDE4 gave the best result on the development data for both translation directions and was therefore used for the final Pilot 0 version.

The words are aligned with GIZA++ and tuning with was done with MERT on a sample of 1% of the training data. The applied heuristics for the Dutch baselines where set to "grow-diag-final-and" alignment and "msd-bidirectional-fe" reordering. For the creation of the language models IRSTLM was used to train a 5-gram language model with Kneser-Ney smoothing on the monolingual part of the training corpora.

1.1.4 German

For both German-English and English-German we follow several optimal system settings as indicated in WMT13 [Bojar et al, 2013]. As the best system UEDIN-SYNTAX [Nadejde et. al, 2013] included several components which were not openly available, we proceeded with adopting several settings from the next best system UEDIN [Durrani et. al, 2013], since the difference of the ranking position is minimal (0.586 vs 0.608 for German-English and 0.587 vs 0.614 for English-German which was not statistically significant as a difference). Commercial systems like Google Translate have the best performance in both directions; although we cannot use them as a state-of-the-art baseline at the current state since they are not openly available, we may consider investigating their potential contribution in a future stage.

In our system we follow the practice of augmenting the training data with domain-specific data, and building relevant extensive language models, interpolated as described above. When having German as a source language, compounds were split using the frequency-based method [Koehn and Knight, 2003].

1.1.5 Portuguese

Besides the Koehn et al [2009] we are not aware of other papers reporting on Portuguese-English Machine Translation. Relatively to the system described in that paper, the only difference seems to be the corpus (they used the JRC-Acquis) but since QTLeap is being evaluated on a QA corpus, we trained our system on Europarl. This seems more adequate as it contains more first and second person text. Concerning the rest of the settings, the Portuguese system follows the parameters adopted for other Romance languages.

1.1.6 Spanish

Spanish state-of-the-art is shown as part of the WMT13. Our system followed the majority of the successful settings for both directions. We proceeded with adopting several settings from UEDIN [Durrani et. Al, 2013], the best system that did not use external resources. Our system was trained using Open-Source tools, and made use all the corpora available in WMT13 workshop, as well as smaller domain-specific corpora (KDE and OpenOffice). Tools available in Moses toolkit were used for tokenization and truecasing, while mGiza was used for word alignment. For language modeling, we use SRILM to train a different LM for each of the corpora available and to combine them by means of LM interpolation.

2.2 Data

Two best practices have guided the selection and creation of data in QTLeap:

- Practice by language service providers (LSPs): use only *original* translation direction for testing (and training)
- Practice by Workshop on Statistical Machine Translation (WMT): allow for variable training data and use the same test data for all language pairs (sometimes the same test set is used in both directions of a given pair by switching the notion of source and target)

2.2.1 Development and Test Data

To adhere to the LSP practice, fresh data has been produced and will be continuously extended within the project, namely translations of the HF customer data from English into the project languages (for the not so important direction from the project language into English, we decided to deviate from the LSP best practice).

Two out of four data sets containing 1000 interactions (question-answer pairs) will be used for the baselines. One set for development and one for testing.

2.2.2 Training Data

By training data, we refer to both monolingual data for training language models and parallel data for training phrase-tables. The availability of training data for the project languages is different. While general vocabulary (GV) data (e.g. Europarl) is widely available, the situation regarding domain ("customer") data is difficult. Although it would be advantageous to do in-domain training, for pragmatic reasons the project decided to train the baselines on a mix of GV and domain data. Following WMT best practice, we allow for variation in the training data used for each project language.

The available corpora are documented for project-internal use in the intranet[2] and are stored in the data repository[3]. The basic sources for data, common for most language pairs are:[4]

[2] http://qtleap.eu/intranet//index.php/WP2baselines
[3] http://194.117.45.196:7777/index.php/apps/files?dir=/Shared/Project/Pilot0
[4] Again, the consortium decided to leave some freedom to the partners in the design of the training corpus.

- **Europarl** [Koehn 2005]
- **News Corpus** and **News Commentary** from WMT12 and WMT13 [Callison-Burch et al., 2012; Bojar et al., 2013]
- **UN corpus** [Eisele et. al, 2010]
- **OPUS corpus** [Tiedemann, 2009], which contains the SeTimes, the EMEA, the JRC Acquis and the Openoffice corpora.

In detail, the following corpora have been used for the baselines:

- **Basque**:
    - Bilingual corpora
        - TMs by Elhuyar Foundation (1.1M sentences), containing academic books, software manuals and software localizations
        - Web crawl with PaCo2 tool by Elhuyar Foundation (400K sentences)
    - Monolingual corpora
        - Monolingual part of parallel corpus (1,5M sentences)
        - Elhuyar TMs (7.4M sentences): Basque part of Spanish-Basque TMs, mainly administrative texts.
- **Bulgarian**:
    - Parallel Corpora
        - SeTimes (cleaned version[5], 154K sentence pairs)
        - Europarl (380K sentence pairs)
        - LibreOffice Document Foundation (71K sentence pairs)
        - Bulgarian-English BTB lexicon (10K word translations)
    - Monolingual Corpora
        - Bulgarian National Reference Corpus[6] (1.4M sentences)
        - English Europarl (2M sentences)
- **Czech**:
    - Parallel Corpora
        - CzEng 1.0 (Czech-English Parallel Corpus[7], 15M sentences)
    - Monolingual Corpora
        - News and political corpora: Monolingual part of News Commentary, Europarl and News as provided by WMT12[8]
        - Czech side of CzEng 1.0 (15M sentences)
- **Dutch**:
    - Parallel Corpora:

[5] http://bultreebank.org/EMP/
[6] http://www.webclark.org/
[7] http://ufal.mff.cuni.cz/czeng/czeng10/
[8] http://www.statmt.org/wmt12/translation-task.html

        - Dutch Parallel Corpus[9] (180K sentence pairs)
        - KDE4 documentation (192K sentence pairs)
    - Monolingual corpus:
        - Monolingual part of Dutch Parallel Corpus
        - Monolingual part of KDE4 documentation
- **German**:
    - Parallel data
        - Generic Data:
            - Europarl (1.7M sentence pairs)
            - News Commentary (154K sentence pairs)
            - UN (156K sentence pairs)
            - CommonCrawl (2.4M sentence pairs)
        - Data of technical domain:
            - The Document Foundation: Libreoffice Help (47K sentence pairs), Libreoffice User Interface (35K parallel entries), The Document Foundation Terminology (690 translated terms), The Document Foundation Website (226 sentence pairs)
            - Chromium browser (6,3K parallel entries)
            - Ubuntu Documentation (6,3K sentence pairs), Ubuntu Saucy (183K parallel entries)
            - Drupal web-content management (5K parallel entries)
    - Monolingual corpus
        - Europarl (2.2M sentences)
        - News Corpus (212K sentences)
        - News Crawl (38.4M sentences)
        - UN (167K sentences)
        - Target side of domain-specific data (285K segments)
- **Portuguese**:
    - Parallel Corpora:
        - Europarl (2M sentence pairs)
    - Monolingual corpus:
        - Europarl (2M sentence pairs)
- **Spanish**:
    - Bilingual corpora
        - Europarl (2M sentence pairs)

[9] described at Paulussen (2006), available at http://tst-centrale.org/nl/producten/corpora/dutch-parallel-corpus-niet-commercieel/6-65

        - UN corpus (11M sentence pairs)
        - News Commentary (174K sentence pairs)
        - Common Crawl corpus (1.8M sentence pairs)
    - Monolingual corpora
        - Europarl (2M sentences)
        - UN corpus (11M sentences)
        - Common crawl (1.8M sentences)
        - News Corpus (13M sentences)
        - Target side of domain-specific data (256K segments)

3 Evaluation metrics, methods, and tools

At a time when the translation industry has been thinking about more adequate and flexible metrics and appropriate tools for QA, research in MT also arrived at a stage where their commonly used quality metrics, automatically computed scores of superficial proximity to one or more human reference translations, do not suffice anymore. When MT reaches a certain quality level as targeted in QTLeap, these scores become less reliable and too much biased toward certain techniques. At this a stage of technological sophistication, truly analytical metrics are needed for guiding research. We have to know which properties of language and which quality demands cause certain drops and gains in quality. Only a multidimensional assessment can deliver the findings required for a methodical investigation of quality barriers.

As pointed out in its DoW, QTLeap is seeking close cooperation with the EC-funded QTLaunchPad project (lead by DFKI), which has as one of its objectives the development of a multidimensional translation quality metric (MQM, see below). QTLaunchPad has invested considerable efforts in the development of a metric for translation quality suited for human and machine translation involving professional translators, LSP companies and MT researchers into the process from the start on.

3.1 State of the Art

QTLaunchPad performed an extensive survey and examined the following commonly used quality assessment (QA) metrics, methods, and tools:

- **Metrics**
    - LISA QA Model. The LISA QA Model is the most frequently implemented common metric.
    - SAE J2450. Standard for assessing quality of automotive industry translations.
    - ISO CD 14080. ISO CD 14080 was withdrawn as a work item from IOS TC 37 in June, 2012, but contained a useful list of categories presented as a standard for assessing all translation.

  - ATA Certification grader criteria. The categories examined are those used for assessing ATA certification exams.
  - SDL TMS Classic. SDL TMS Classic is a simple metric implemented in SDL systems.
  - XLIFF:doc. The XLIFF:doc format created by the Interoperability Now! group contains a set of quality markup categories.
- **Tools**
  - Acrocheck
  - ApSIC XBench
  - Okapi CheckMate
  - Yamagata QA Distiller
  - LPET (Language Proficiency Evaluation Tool) from the Center for Advanced Study of Language (CASL)

In addition the assessment task addressed the following measures for MT quality:

- **Manual**
  - Adequacy and Fluency
  - Rankings
  - Error Analysis
- **Automatic**
  - BLEU [Papineni et. al, 2002]
  - NIST
  - METEOR [Lavie et. al 2007]
  - WER-/PER/(h)TER/TERp [Levenshtein, 1966; Snover et. al 2009]
- **Other** metrics
  - Post-Editing time [Beck et. al 2013]
- **Quality Estimation**
- **MT Error Classification and Diagnostic MT Evaluation**
- **Tools for MT Error Analysis**
  - Addicter [Berka et. al, 2012]
  - AMEANA
  - BLAST
  - DELiC4MT
  - Hjerson [Popović et al, 2011]
  - TerrorCat [Fishel et al, 2011]
  - Woodpecker

The project's detailed report on the comparison of these items can be made available on request. It found that there is little consistency in the methods considered, with a particularly noticeable divide between the human and machine translation QA methods mentioned previously. Key findings include the following observations:

- **Human translation**
    - There is tremendous variability in the number of issue types recognized by the various metrics and tools, ranging from a low of 6 (SAE J2450) to 65 (the full documentation of the LISA QA Model).
    - The only issue shared across all human-oriented metrics examined is *terminology.*

Different metrics and tools focus on different items. For example, some metrics do not examine white space as a specific issue, while Okapi CheckMate identifies eight separate issues related to whitespace.

    - Automatic tools, not surprisingly, tend to focus on easily identifiable formal characteristics of the text, while metrics aimed at human reviewers focus on issues such as meaning, clarity, and other items that are more difficult to identify formally.
    - Differences in granularity, scope, and overlap present a problem for comparing quality scores.
    - Some metrics (particularly the LISA QA Model) combine/confuse different types of quality. For example, the LISA QA Model includes project and process quality measures, as well as measures for assessing the product quality.
    - No real "tuning" of several existing metrics for various needs is possible.

- **Machine translation**
    - Machine translation assessment methods all use some form of "distance" from an ideal human translation (or multiple translations) as the measure of quality. As a result they are sensitive to the translations chosen, and scores will change if the reference(s) in use change.
    - Although considered "automatic" metrics, these in fact rely on the existence of human-produced reference translations and so are in fact quite manual and labour intensive.
    - Many of the MT metrics are very useful for research purposes, but they are costly and difficult to implement for production environments.
    - Large improvements in scores such as BLEU may be possible without improving the perceptual quality if the improvement in scores happens in the lowest-quality segments. Thus improvements may not correspond to human judgment.
    - The maintenance of separate methods for MT presents a significant barrier to adoption of MT since potential users have difficulty understanding how the quality of MT relates to that of typical human translation.

Based on this analysis, the QTLaunchPad project concluded that there is a significant need for a common system (metaframework) for describing and comparing quality systems, capable of emulating existing systems. At the same time, due to differences in the metrics, it needed to be an extensible framework. The Multidimensional Quality Metrics (MQM) proposed by QTLaunchPad addresses these needs.

3.2 MQM

MQM[10] provides a master catalog of issue types suitable for various tasks. MQM lets you declare your quality metric in a shared vocabulary. Figure 3 in the Appendix shows the full MQM hierarchy with its top level branches including accuracy, fluency and verity (the latter accounts for pragmatic/localisation errors where the translation does not match the state of the world even if it is fine as a translation). The idea is to build custom metrics for a given task. For example, Figure 1 shows a metric that has been used for MT diagnosis in experiments that have been conducted within QTLaunchPad involving several LSPs while Figure 2 shows a metric that models the industry standard SAJ 2450.

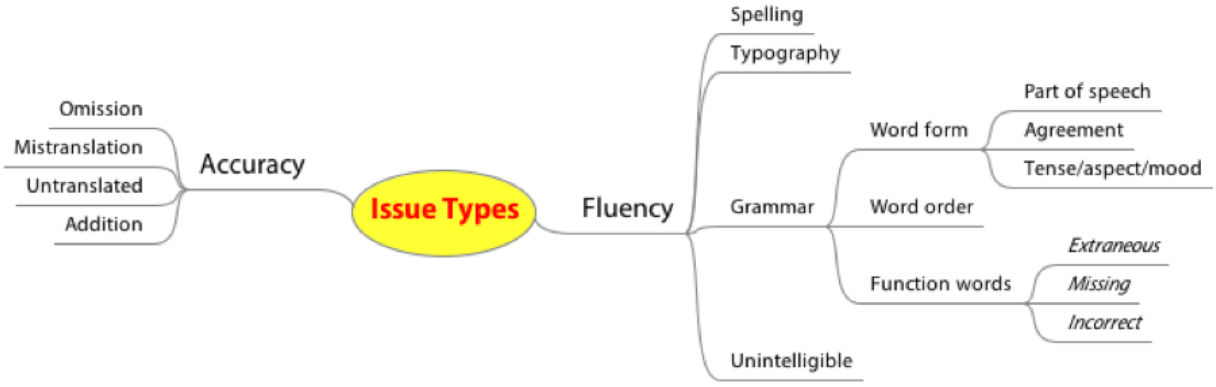


Figure 1: Metric for MT diagnostics

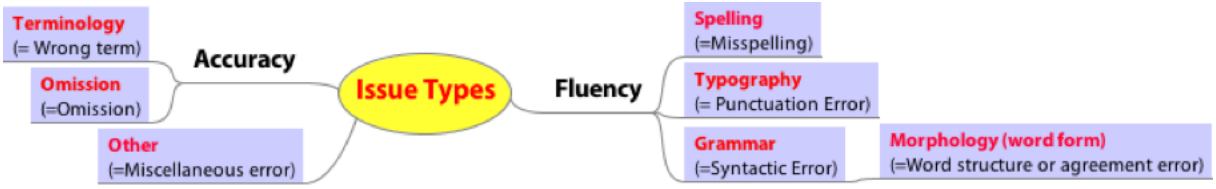


Figure 2: Metric modeling SAJ 2450

3.2.1 Tools for MQM

MQM is free and open and can thus be implemented in any tool. The QTLaunchPad project has supported the implementation in the open-source editor translate5[11] (see Figure 4 in the Appendix) and provides two different score cards that can be used for error assessment (see Figure 5 and Figure 6 in the Appendix).

[10] http://www.qt21.eu/launchpad/content/multidimensional-quality-metrics
[11] http://www.translate5.net/

4 Translation Quality Evaluation in QTLeap

Overall, in QTLeap, the following types of quality evaluation should be used:

1. Reference translations plus automatic measures (BLEU, METEOR, etc.) for setting up baselines, building corpora, quick comparisons, etc.
2. Human raters for pairwise comparison of improvement from Pilot to Pilot (simple scalar rating better-worse-same)
3. Post-edits for determining the productivity gain in the production scenario (see D3.3)
4. Explicit error markup for analytical insights into quality barriers and measuring improvement using an MQM metric

As 2) to 4) require human intervention, these types of evaluation can be performed only if resources permit and only on subsamples of the full test corpora. Suggestions have been made to team up with interested translation departments or companies from the QTLeap board of potential users and ask if they can contribute some man power to evaluation in their own interest.

4.1 Evaluation of baselines

The baselines have been tuned on usage data as this is a common practice to ensure that the SMT parameters fit the particular usage scenario. The baselines have been evaluated against the reference corpus provided by HF (1000 interactions) using automatic measures (BLEU, METEOR).

	Basque	Bulg.	Czech	Dutch	German	Portug.	Spanish
from [en]	20.24	23.52	26.26	31.50	32.42	19.91	45.86
to [en]	28.84	35.61	29.89	37.00	46.74	24.79	51.38

Table 1: Automatic BLEU scores for baselines

	Basque	Bulg.	Czech	Dutch	German	Portug.	Spanish
from [en]	n/a[12]	n/a	24.03	29.66	47.96	32.45	65.48
to [en]	26.68	25.36	30.25	30.45	36.13	25.61	41.10

Table 2: Automatic METEOR scores for baselines

An extrinsic evaluation of Pilot 0 implementing the baselines can be found in Deliverable D3.6. Human user evaluation rounds for optimising the SMT systems would have gone beyond the scope and capacities of this project.

It is recommended that the partners in the future perform an analytic, in-depth human evaluation on a random sample of at least 100 segments (see Section 3).

[12] Bulgarian and Basque are not supported by METEOR

5 Outlook

Together with development of the extrinsic evaluation metric in Task 3.5, an MQM metric was designed for QTLeap[13] starting from the metric in Figure 1: Metric for MT diagnostics. Some baselines have already served in its development.

When the deeper systems are being developed, we will re-inspect the user data, baselines, and metric again to make sure that the targeted phenomena are covered.

6 References

Daniel Beck, Lucia Specia, and Trevor Cohn. 2013. *Reducing Annotation Effort for Quality Estimation via Active Learning*. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 543–548, Sofia, Bulgaria, August. Association for Computational Linguistics

Jan Berka, Ondrej Bojar, Mark Fishel, Maja Popović, and Daniel Zeman. 2012. *Automatic MT Error Analysis: Hjerson Helping Addicter*. In 8th International Conference on Language Resources and Evaluation, pages 2158–2163

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. *Findings of the 2013 Workshop on Statistical Machine Translation*. In 8th Workshop on Statistical Machine Translation, Sofia, Bulgaria. Association for Computational Linguistics.

Ondřej Bojar, Rudolf Rosa, and Aleš Tamchyna. 2013. *Chimera - Three Heads for English-to-Czech Translation*. In Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 92–98, Sofia, Bulgaria, August.

Arantza Díaz de Ilarraza, Gorka Labaka, and Kepa Sarasola. 2009. R*elevance of Different Segmentation Options on Spanish-Basque SMT*. In Proceedings of the 13th Annual Conference of the European Association for Machine Translation, pages 74–80, Barcelona, Spain.

Arantza Díaz de Ilarraza, Gorka Labaka, and Kepa Sarasola. 2009. *Reordering on Spanish-Basque SMT.* In MT Summit XII: proceedings of the twelfth Machine Translation Summit, pages 207–213, Ontario, Canada.

Nadir Durrani, Barry Haddow, Philipp Koehn, and Kenneth Heafield. 2014. *Edinburgh's Phrase-based Machine Translation Systems for WMT-14*. In Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 97–104, Baltimore, Maryland, USA, June.

Nadir Durrani, Barry Haddow, Kenneth Heafield and Philipp Koehn. 2013. *Edinburgh's machine translation systems for European language pairs*. In Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 114-121, Sofia, Bulgaria, August.

[13] See, e.g., Deliverable D2.4 for the metric used in QTLeap.

Andreas Eisele and Yu Chen. 2010. *MultiUN: A Multilingual Corpus from United Nation Documents*. In Daniel Tapias, Mike Rosner, Stelios Piperidis, Jan Odjik, Joseph Mariani, Bente Maegaard, Khalid Choukri, and Nicoletta Calzolari (Conference Chair), editors, Proceedings of the Seventh conference on International Language Resources and Evaluation. International Conference on Language Resources and Evaluation (LREC-2010), May 19-21, La Valletta, Malta, pages 2868–2872. European Language Resources Association (ELRA)

Mark Fishel. 2013. *Ranking Translations using Error Analysis and Quality Estimation*. In Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 405–407, Sofia, Bulgaria, August. Association for Computational Linguistics.

Philipp Koehn. 2005. *Europarl: A parallel corpus for statistical machine translation*. In Proceedings of the tenth Machine Translation Summit, volume 5, pages 79–86, Phuket, Thailand.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. *462 machine translation systems for Europe*. In MT Summit XII: proceedings of the twelfth Machine Translation Summit, pages 65–72, Ontario, Canada.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation*. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pages 177–180, Prague, Czech Republic, June.

Philipp Koehn and Kevin Knight. 2003. *Empirical Methods for Compound Splitting*. In Proceedings of the 10th Annual Conference of the European Association for Machine Translation, volume 1, page 8, Budapest, Hungary.

Gorka Labaka. 2009. *EUSMT: Incorporating Linguistic information to Statistical Machine Translation for a morphologically rich language. Its use in preliminary SMT-RBMT-EBMT hybridization*. Ph.D. thesis, University of the Basque Country.

Alon Lavie and Abhaya Agarwal. 2007. *METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments*. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 228–231, Prague, Czech Republic, June. Association for Computational Linguistics.

Vladimir Levenshtein. 1966. *Binary Codes Capable of Correcting Deletions and Insertions and Reversals*. Soviet Physics Doklady, 10(8):707–710.

Aingeru Mayor, Iñaki Alegria, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, and Kepa Sarasola. 2011. *Matxin, an open-source rule-based machine translation system for Basque*. Machine Translation, 25(1):53–82.

Maria Nadejde, Philip Williams, and Philipp Koehn. 2013. *Edinburgh's Syntax-Based Machine Translation Systems*. In Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 170–176, Sofia, Bulgari

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation*. In Proceedings of the

40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

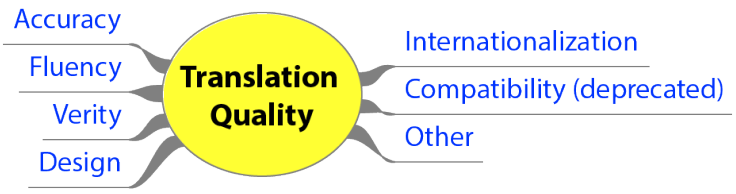
Hans Paulussen, Lieve Macken, Julia Trushkina, Piet Desmet, and Willy Vandeweghe. *Dutch Parallel Corpus: a multifunctional and multilingual corpus.* Cahiers de l'Institut de Linguistique de Louvain 32, no. 1-4 (2006): 269-285.

Maja Popović. 2011. *Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output*. The Prague Bulletin of Mathematical Linguistics, 96(-1):59–68. From Duplicate 3 (Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output - Popović, Maja) .

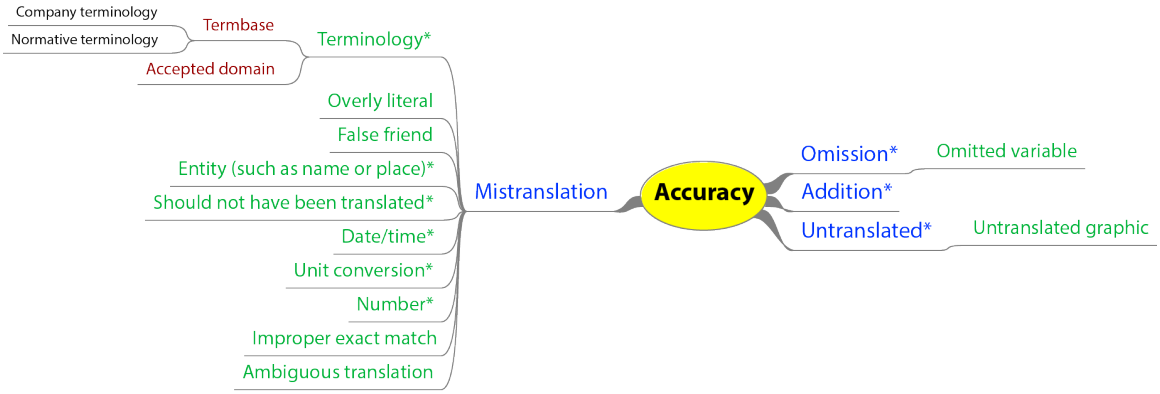
Matthew Snover, Nitin Madnani, Bonnie J Dorr, and Richard Schwartz. 2009. *Fluency, Adequacy, or HTER?: Exploring Different Human Judgments with a Tunable MT Metric*. In Proceedings of the Fourth Workshop on Statistical Machine Translation, pages 259–268, Stroudsburg, PA, USA, March. Association for Computational Linguistics.

Jorg Tiedemann. 2009. *News from OPUS—A Collection of Multilingual Parallel Corpora with Tools and Interfaces.* In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, Advances in Natural Language Processing, volume V, chapter V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

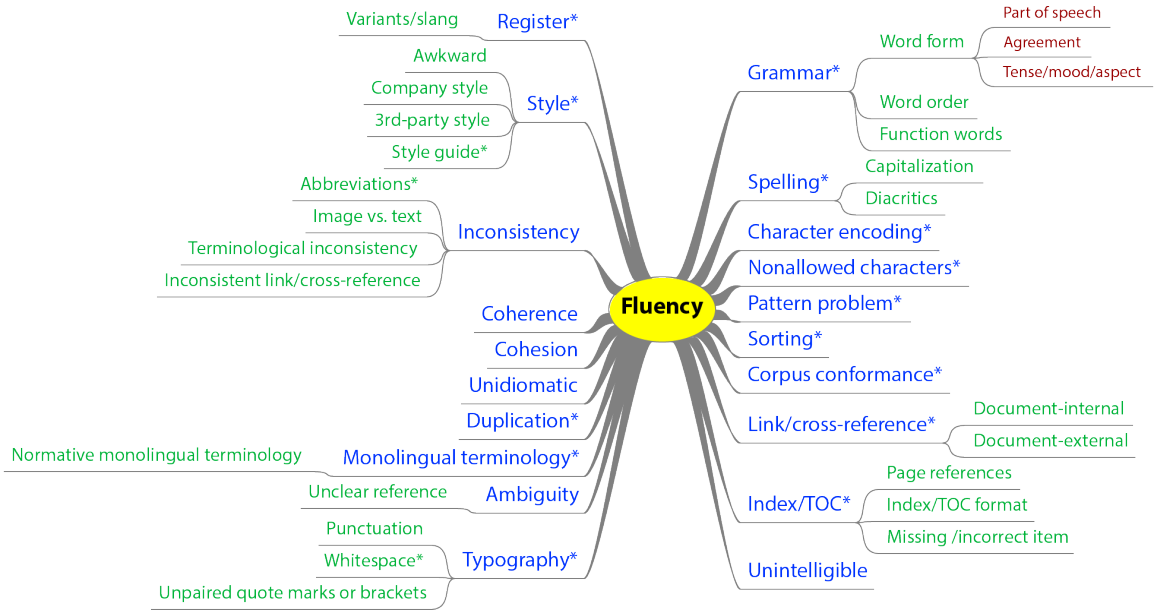
7 Appendix



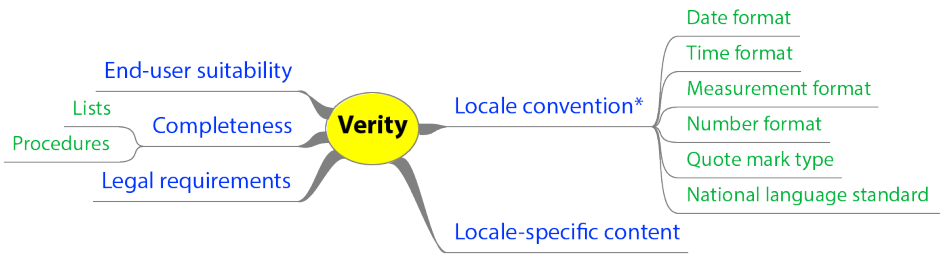
a. High-level structure



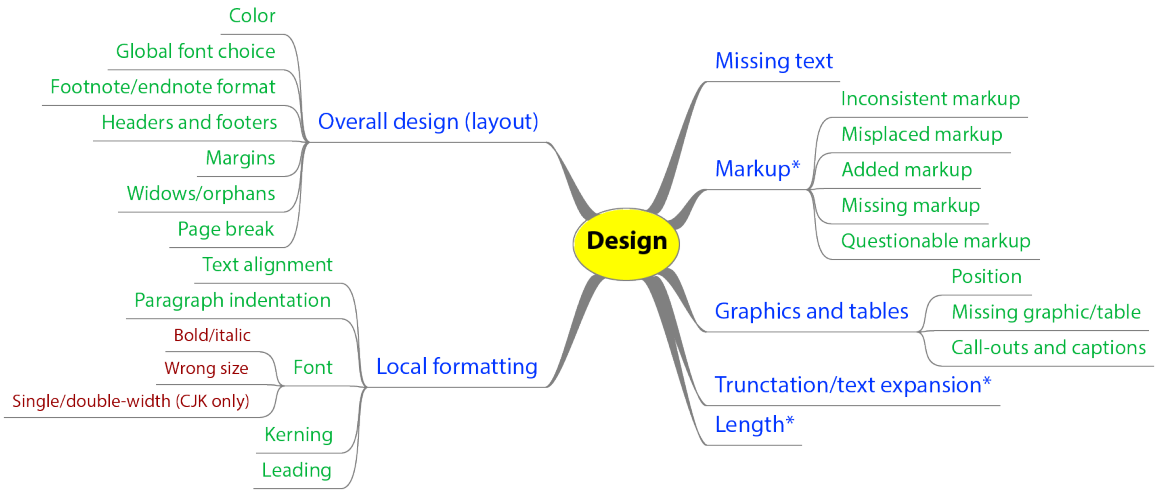
b. Accuracy dimension



c. Fluency dimension



d. Verity dimension



e. Design dimension

Figure 3: Full MQM. The version of MQM depicted above contains no issues in the Internationalization dimension and does not show deprecates issues from Compatibility. (Note: This version of MQM has since been replaced by a newer version with substantial differences. The depicted version corresponds to the version at http://www.qt21.eu/mqm-definition/issues-list-2014-03-19.html. The current version of MQM is always accessible at http://qt21.eu/mqm-definition.)

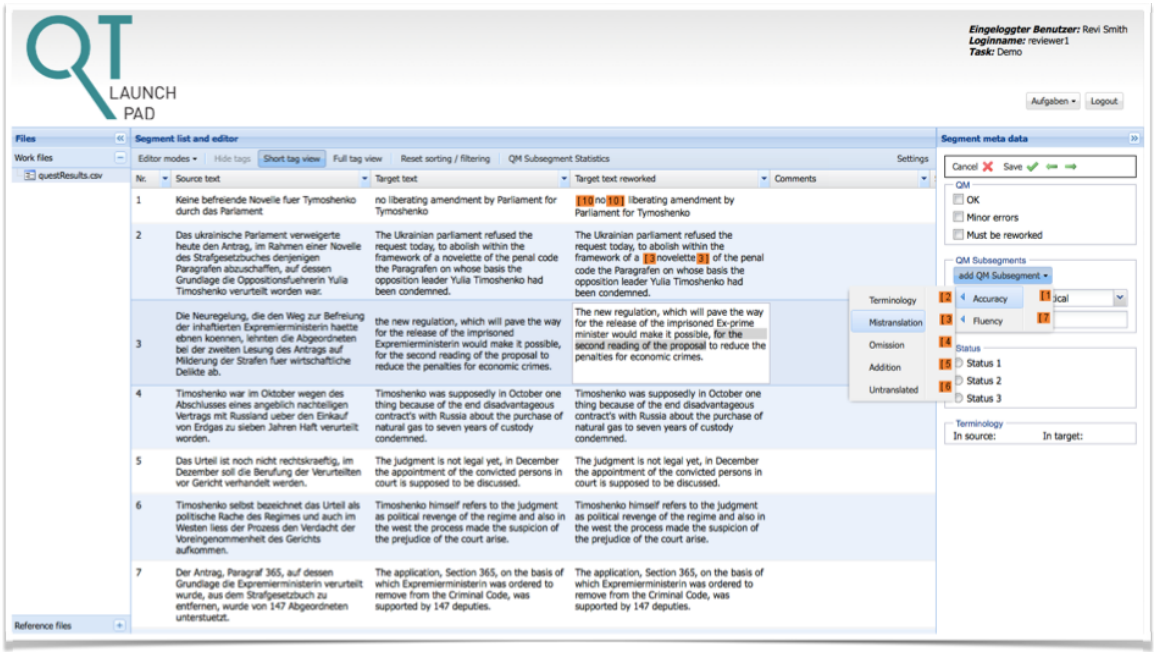


Figure 4: MQM error markup in translate5

Information about the text	
Word count (source):	1000
Word count (target):	10000

Severity multipliers	
Minor (default = 1)	
Major (default = 5)	
Critical (default = 10)	

←These values may be overridden by the user

I am performing a source quality check

Issue Type	Weight (default = 1)	Target Text Issues: Issue counts Minor	Major	Critical	Penalty Raw	Penalty Adj.	Target Subscore	Source Text Issues: Issue counts Minor	Major	Critical	Penalty Raw	Penalty Adj.	Source Subscore	Adjusted (Delta) Scores
Accuracy														
Terminology		2			2	2	100.0%							100.0%
Mistranslation		3			3	3	100.0%							100.0%
Omission		1			1	1	100.0%	Not applicable						100.0%
Untranslated			4		20	20	99.8%							99.8%
Addition					0	0	100.0%							100.0%
Accuracy subtotal						26	99.7%							**99.7%**
Fluency														
Content														
Register		1			1	1	100.0%							100.0%
Style					0	0	100.0%							100.0%
Inconsistency			2		10	10	99.9%							99.9%
Mechanical														
Spelling		1			1	1	100.0%							100.0%
Typography			2		10	10	99.9%							99.9%
Grammar			2		10	10	99.9%							99.9%
Locale convention		3			3	3	100.0%							100.0%
Unintelligible			5		25	25	99.8%							99.8%
Fluency subtotal						35	99.7%							**99.7%**
Verity														
Completeness			9		45	45	99.6%							99.6%
Legal requirements					0	0	100.0%							100.0%
Locale applicability					0	0	100.0%							100.0%
Fluency subtotal						45	99.6%							**99.6%**
TOTAL						106	98.9%							**98.9%**

Figure 5: Tabular scorecard

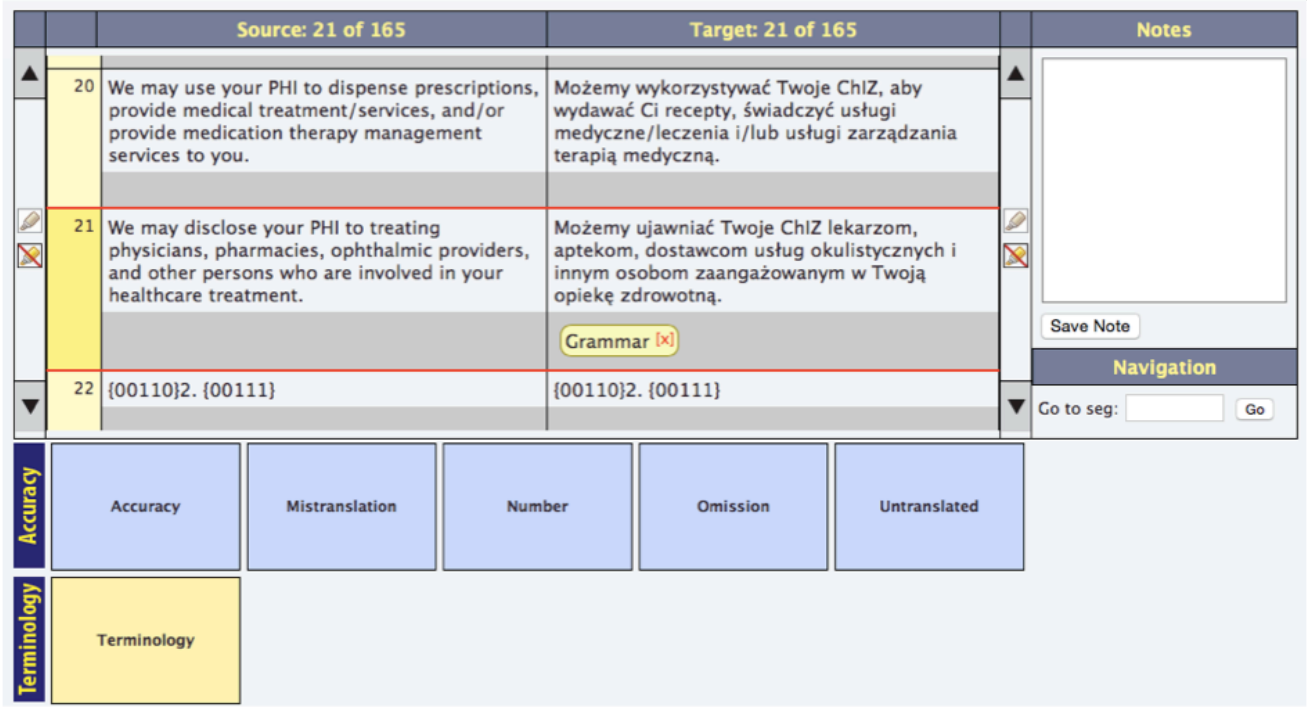


Figure 6: Ergonomic scorecard