



Deliverable D 1.3.1

Barriers for High-Quality Machine Translation

Author(s): Aljoscha Burchardt (DFKI)
Federico Gaspari (DCU)
Kim Harris (text&form)
Arle Lommel (DFKI)
Maja Popović (DFKI)
Antonio Toral (DCU)
Hans Uszkoreit (DFKI)

Dissemination Level: Public

Date: 2014.12.12
(original version: 2014.07.02)



This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 296347.

D1.3.1 Barriers for High-Quality Machine Translation

Grant agreement no.	296347
Project acronym	QTLaunchPad
Project full title	Preparation and Launch of a Large-scale Action for Quality Translation Technology
Funding scheme	Coordination and Support Action
Coordinator	Prof. Hans Uszkoreit (DFKI)
Start date, duration	1 July 2012, 24 months
Distribution	Public
Contractual date of delivery	December 2013
Actual date of delivery	original: February 2014; version 2: July 2014; version 2.5: December 2014
Deliverable number	1.3.1
Deliverable title	Barriers for HQMT
Type	Other (Report & Data)
Status and version	2.5, revised
Number of pages	79
Contributing partners	DCU, USFD
WP leader	DFKI
Task leader	DFKI
Authors	Aljoscha Burchardt, Federico Gaspari, Kim Harris, Arle Lommel, Maja Popović, Antonio Toral, Hans Uszkoreit
EC project officer	Aleksandra Wesolowska
The partners in QTLaunchPad are:	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
	Dublin City University (DCU), Ireland
	Institute for Language and Speech Processing, R.C. "Athena" (ILSP/ATHENA RC), Greece
	The University of Sheffield (USFD), United Kingdom

For copies of reports, updates on project activities and other QTLaunchPad-related information, contact:

DFKI GmbH

QTLaunchPad

Dr. Aljoscha Burchardt

Alt-Moabit 91c

10559 Berlin, Germany

aljoscha.burchardt@dfki.de

Phone: +49 (30) 23895-1838

Fax: +49 (30) 23895-1810

Copies of reports and other material can also be accessed via <http://www.qt21.eu/launchpad>

© 2014, The Individual Authors

This work is licensed under a Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).

D1.3.1 Barriers for High-Quality Machine Translation

Contents

1. Updated Results	4
1.1. Data	4
1.2. Inter-annotator agreement	6
1.3. Data structure	6
1.4. Filtering data	7
1.5. Analysis of issue types	8
2. Executive Summary	9
3. Experimental setup	11
3.1. Selection of “near miss” candidates, calibration sets, and additional data	11
3.2. Selection of annotators	12
3.3. Training of annotators	13
3.4. Annotation of the calibration set (Phase I)	13
3.5. Selection and annotation of customer data and additional WMT data (Phase II)	14
3.6. Analysis	14
4. Results of the Annotation Task	15
4.1. Error profiles	15
4.2. Inter-Annotator Agreement	26
5. Comparison Data from Automatic Error Analysis	31
5.1. Edit types	31
5.2. Reordering distances	32
5.3. Edit clustering	33
6. Relating Translation Quality Barriers to Source-Text Properties	35
6.1. Data and processing	36
6.2. Diagnostic Evaluation Results	37
6.3. Comparison of DELiC4MT Results with Other Findings	44
6.4. Tabular Data for Diagnostic Evaluation	46
7. Lessons Learned in the Annotation Task	51
8. References	52
9. Appendix: Description of DELiC4MT Functionality	53
9.1. General Presentation of DELiC4MT	53
9.2. Novelty of the Analysis with DELiC4MT	54
10. Appendix: Annotation Guidelines	56
10.1. Original MQM Issue Annotation Guidelines	56
10.2. Revised Annotation Guidelines	64
11. Revised training materials	71
12. Change log	79
12.1. July 2014	79
12.2. December 2014	79

D1.3.1 Barriers for High-Quality Machine Translation

1. Updated Results

This version of QTLaunchPad Deliverable D1.3.1 contains the original versions of D1.3.1 (sections 2–10). It adds information on a second round of annotation undertaken to address issues encountered in the work described in the original version (see also D1.2.1 for description of some of the issues addressed). Notably, the second annotation round utilized the revised set of MQM issue types described in **Section 7** and annotation of both previously annotated and “fresh” data. This update does *not* present a reanalysis of all the work included in the original version of this deliverable report, but rather addresses some key points found in the first annotation round and in discussion with reviewers.

1.1. Data

The second annotation round resulted in quadruple annotation of 300 segments of translated data per language pair: 200 segments of WMT data, including some previously annotated data and some that had not previously been annotated; and 100 segments of previously annotated “customer” data. Unlike in the first round, all annotators annotated the same data. Although some of the same companies participated in this round as in the first annotation round, we requested that the individual annotators not be the same as the annotators from the previous round (to remove the variable of using experienced annotators in the second round).

The annotated data is available at <http://www.qt21.eu/deliverables/annotations/index.html>. That website includes the data from the first round of annotations (described in D1.2.1 and the original version of this deliverable) as well as the second round described in this version. The website provides a variety of ways to filter and search the annotated data in order to make it more accessible and useful.

1.1.1. Annotators

Translated data was annotated by the following language service providers (LSPs):

- EULE (DE>EN)
- iDisc (ES>EN)
- Linguaserve (ES>EN and EN>ES)
- Rheinschrift (EN>DE)
- text&form (DE>EN and EN>DE)
- Treeloc (EN>ES)

Each LSP provided two independent annotators per each language pair they covered. The annotators used the translate5 system to annotate data. The results were converted from the CSV export provided by translate5 into a “pretty print” HTML format using a tool developed by text&form. Additional data manipulation was carried out at DFKI to provide additional functionality (such as filtering).

D1.3.1 Barriers for High-Quality Machine Translation

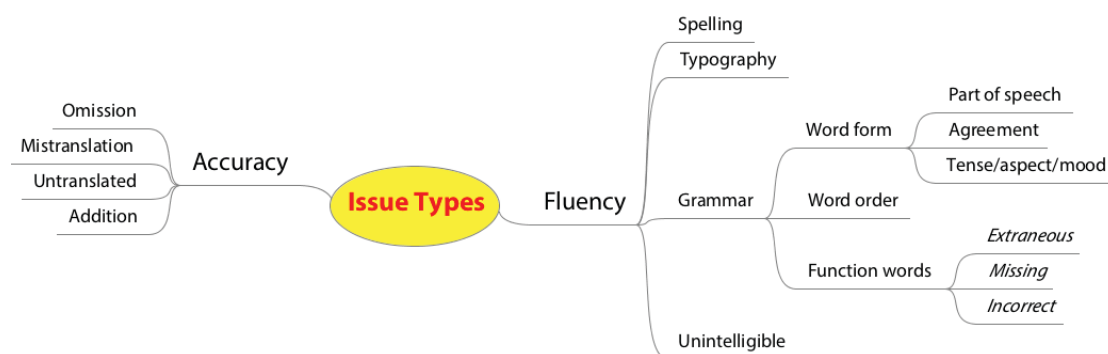


Figure 1. Revised MQM Issue Type hierarchy

Language pair	F-Scores (Sentence-Level Agreement)			Cohen's Kappa (Span-Level Agreement)		
	Round 1	Round 2	Change	Round 1	Round 2	Change
DE>EN	34.7	47.1	+12.4	0.29	0.36	+0.07
EN>DE	36.5	47.8	+11.3	0.25	0.39	+0.14
ES>EN	28.1	51.2	+23.1	0.32	0.34	+0.02
EN>ES	32.5	48.9	+16.4	0.34	0.42	+0.08
Average	33.0	48.8	+15.8	0.30	0.38	+0.08

Table 1. Changes in inter-annotator agreement between round 1 and 2. (F-scores are presented with a maximum possible score of 100 and Cohen's kappa with a maximum of 1.0.)

1.1.2. Annotation costs

Annotation costs can only be estimated approximately in advance. For annotation tasks of a complexity similar to those described in this document, the estimated direct cost for triple annotation (to establish inter-annotator agreement) is approximately 20€–30€/1000 (target) words. The actual cost depends on the language pair (which can impact the cost of qualified annotators), number of errors in the text, text domain, training of annotators (experienced annotators are generally faster than untrained annotators), etc. For higher-quality text the costs would be lower and for texts with significantly more errors, the costs would be higher, and human translations would generally be less expensive to annotate than machine translation; the cost range cited here reflects work with relatively inexperienced annotators, and might become substantially lower over time. When adding in management, analysis, and other overheads, fully loaded costs would be 100–200% higher. For more details on cost see the QTLaunchPad supplemental report *Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality*¹.

1.1.3. Issue types

Based on the experience documented in **Section 7. Lessons Learned in the Annotation Task** (page 51), the second round of annotations used the updated set of MQM issue types shown in **Figure 1** (above). The change in issue type selection means that some data from the first annotation round cannot be compared directly with data from this round (e.g., there is no equivalent to the *Style/register* issue type in the first round).

¹ Available at <http://qt21.eu/downloads/MQM-usage-guidelines.pdf>.

D1.3.1 Barriers for High-Quality Machine Translation

Annotations				
[derstandant_at2012/12/01/141907-2_de_SMT]				
"Ich finde das Gebäude lustig, es sieht futuristisch aus, und endlich gibt es wieder etwas Interessantes zu sehen", sagt Lisette Verhaig, eine Passantin am Straßenrand.				
1 (SMT)	"I find it funny, it looks futuristic, and finally there is something interesting to see again," said Lisette Verhaig, a technician at the roadside.			2-
AA	"I find <Mistranslation>-it<Mistranslation> funny, it looks futuristic, and finally there is something interesting to see again," <Tense/aspect/mood>-said <Tense/aspect/mood>-Lisette Verhaig, a <Mistranslation>-technician <Mistranslation>-incorrect function word-at<Incorrect function word> the roadside.	4	<ul style="list-style-type: none">• Mistranslation [it]• Tense/aspect/mood [said]• Mistranslation [technician]• Incorrect function word [at]	
BB	"I find <Mistranslation>-it<Mistranslation> funny, it looks futuristic, and finally there is something interesting to see again," said Lisette Verhaig, a <Mistranslation>-technician <Mistranslation>-at the roadside.	2	<ul style="list-style-type: none">• Mistranslation [it]• Mistranslation [technician]	
MM	"I find <Omission>-it<Omission> funny, it looks futuristic, and finally there is something interesting to see again," said Lisette Verhaig, a <Mistranslation>-technician<Mistranslation> at the roadside.	2	<ul style="list-style-type: none">• Omission [it]• Mistranslation [technician]	
NN	"I find <Mistranslation>-it<Mistranslation> funny, it looks futuristic, and finally there is something interesting to see again," <Mistranslation>-said<Mistranslation>- Lisette Verhaig, a <Mistranslation>-technician<Mistranslation> at the roadside.	3	<ul style="list-style-type: none">• Mistranslation [it]• Mistranslation [said]• Mistranslation [technician]	
gold	"I find <Mistranslation>-it<Mistranslation> funny, it looks futuristic, and finally there is something interesting to see again," said Lisette Verhaig, a <Mistranslation>-technician<Mistranslation> at the roadside.	2	<ul style="list-style-type: none">• Mistranslation [it]• Mistranslation [technician]	
[derstandant_at2012/12/01/141907-4_de_SMT]				
Aber ich frage mich, wofür wir heute noch eine Bücherei brauchen.				
2 (SMT)	But I wonder what we still need a library .			1-
AA	But I wonder <Mistranslation>-what <Mistranslation>-we still need a library<Omission> <Omission>.	2	<ul style="list-style-type: none">• Mistranslation [what]• Omission []	
BB	But I wonder what we still need a library<Missing function word>-<Missing function word>	1	<ul style="list-style-type: none">• Missing function word [,]	
MM	But I wonder <Mistranslation>-what<Mistranslation> we still need a library.	1	<ul style="list-style-type: none">• Mistranslation [what]	
NN	But I wonder what we still need a library<Omission> <Omission><Omission> <Omission>.	2	<ul style="list-style-type: none">• Omission [,]• Omission [,]	
gold	But I wonder <Mistranslation>-what<Mistranslation> we still need a library.	1	<ul style="list-style-type: none">• Mistranslation [what]	

Figure 2. Annotated data in online viewer.

1.2. Inter-annotator agreement

One of the concerns about the original inter-annotator agreement (IAA) figures from the first round of annotation was that they were considered "fair" but were not ideal. Annotators in this round were provided with improved training materials (see the revised annotator guidelines included in **Section 11** and the video explanation at <http://www.qt21.eu/downloads/annotationWebinar-2014-06-11.mov>). These materials included a "decision tree" to help annotators select issues more consistently. Feedback from annotators indicated that this tool was particularly useful.

The revised training materials were effective, and this round demonstrated a significant improvement in measures of IAA, as shown in **Table 1** (previous page), which presents both F-scores (showing sentence-level agreement about the issue types) and Cohen's kappa (showing span-level agreement). The improvements shown were found with minimal training and novice reviewers and we anticipate that with further training that agreement would increase (especially as an examination of the data shows that not all of the annotators followed the written instructions as closely as would be desirable).

1.3. Data structure

The annotated data is presented as a set of HTML files at <http://www.qt21.eu/deliverables/annotations>. The data is presented in a tabular format, as shown in **Figure 2**. The data includes the following information:

- **Segment ID value.** A unique ID value based on either the WMT identifier (for WMT data) or an internal code (for customer data).
- **Source segment.** The source segment that was translated.
- **Row number.** A row number that can be used to identify rows in the data set.
- **Translation type.** An indication of what type of system (SMT = statistical MT, RbMT = Rule-based MT, hybrid = Hybrid MT, or Human = human translation) that produced the translation.
- **Target segment.** The raw output from the indicated translation system.

D1.3.1 Barriers for High-Quality Machine Translation

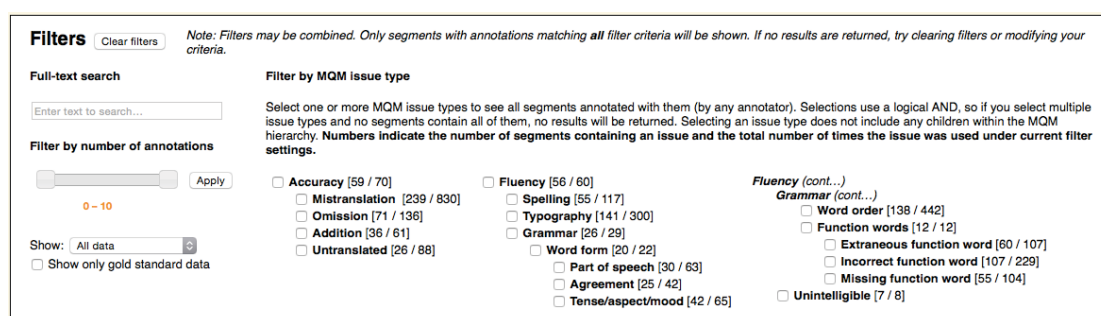


Figure 3. Filter options.

- **Range of annotations.** A set of numbers indicating the highest and lowest number of annotations provided by the annotators (e.g., 1–4 would indicate that at least one annotator only noted one issue and one noted four issues)
- For each annotator a row containing the **annotator's ID**, the **annotated segment**, the **number of issues annotated**, and a **listing of annotations**.
- For some segments, an **expert-produced annotation** (identified as “gold” in **Figure 2**, although this label will be changed).

The data is also available in an XML format to allow for uses not supported by an online HTML file.

1.4. Filtering data

The online data sets support advanced filtering to assist users in selecting relevant data for various purposes, as shown in **Figure 3**. The data supports the following filter options:

- **Full-text search.** Content can be searched based on a full-text search of data about each segment.
- **Number of annotations.** Data can be filtered by the number of annotations provided for each segment. For example, selecting “2–4” will show those segments annotated with 2 to 4 annotations. Filtering is done taking into account *all* annotations for a given segment, so with a setting of “2–4” if three annotators noted three issues and one noted five issues, the segment would *not* be shown.
- **Data source.** The system can show just WMT data, just customer data, or both.
- **MQM issue type.** Selecting an MQM issue type will show those segments where at least one annotator chose that issue type. If more than one issue type is chosen, only segments where *all* issue types selected appear will be shown. Numbers in the filter interface indicate how many segments contain each issue type (and the total number of times that issue type is annotated in those segments) under the current filter settings.

Note that filters can be used in combination with one another (i.e., the operate with a logical AND) to provide detailed searches.

D1.3.1 Barriers for High-Quality Machine Translation

1.5. Analysis of issue types

This update does not attempt to repeat the detailed analysis of issue types seen in the previously submitted version of this deliverable. To some extent this analysis is duplicated through the information provided in the filters described above. It is anticipated that the improved access to annotated data provided with this update will enable researchers to discover additional insights into MT output.

D1.3.1 Barriers for High-Quality Machine Translation

2. Executive Summary

This Deliverable is one of the outcomes of Task 1.3 “Evaluation of existing MT approaches and techniques regarding the barriers”. It contains results of the analysis of translation errors in the corpus of “nearly acceptable translated sentences” collected in D1.2.1. This Deliverable serves as an intermediate step between D1.2.1/D1.2.2 and D1.4.1. The analyses presented in D1.2.2 were partly based on automatic analyses of post-edited data and partly on the small first pilot error corpus D1.2.1. In-line with the recommendations of the reviewers at the first project review, we have increased the size of MQM-annotated data sets across different language pairs by way of an “annotation campaign” coordinated by GALA and involving ten different LSPs. The data generated for this Deliverable was annotated using the translate5 QTLT infrastructure. It will also be used in the Shared Task (WP5) and for testing and further development of QuEst (WP2).

The central goal of this task is the evaluation of the performance of state-of-the-art machine translation systems on nearly acceptable translated sentences with regard to the error types identified in D1.2.1 by:

- Identification of MQM error classes remaining in nearly acceptable translations:
 - common for all automatic approaches
 - specific to statistical systems
 - specific to rule-based systems
 - specific to hybrid systems (where available)
 - specific to human translations
- Estimation of the distribution of different post-editing operations using automatic linguistic analyses of the source segments
- Correlation of the results of these two approaches

These approaches taken together provide insight into the remaining barriers characteristic of “near-miss” translations and show distinct patterns of errors that characterize these sentences. They also show differences between language pairs and systems that can be used to suggest improvements to systems.

Overall findings include the following:

- All of the MT systems examined actually outperformed human “near miss” translations in terms of accuracy to some extent, but were significantly worse in terms of grammaticality.
- Different language pairs show distinct remaining barriers regarding the MQM error set defined in D1.2.1, even when different MT technologies are used. These results suggest that there is no “one-size-fits all” model to improve any class of MT system. Likewise, hybrid solutions will also not be able to rule out all errors. Instead language pair-specific methods will need to be coupled with general methods. For example, German→English MT shows more word order problems than the other pairs examined. This finding is a result of the fact that German is relatively more likely to use a relatively free word order and convey information about syntactic roles through the use of morphological markers than the other languages considered. Addressing this problem in an SMT system would require the introduction of linguistic knowledge rather than simply adding additional training data. (And in fact existing hybrid sys-

D1.3.1 Barriers for High-Quality Machine Translation

tems like Asia Online’s customized Moses system do use syntax-driven reordering rules to obtain improved results (Dion Wiggins, pers. comm.).)

- Inter-annotator agreement (IAA) was lower than anticipated. However, despite being substantially more complex than the Likert scale-based rankings used for WMT evaluations, IAA for MQM was roughly comparable to that of WMT assessment tasks. While there is room for improvement, the comparability of these rates to those found in conceptually simpler tasks suggests that the MQM-based analytic approach has merit.
- As with the analysis of the pilot corpus (D1.2.2), we found good overall agreement between MQM-based assessment tasks and automatic analysis based on post-editing tasks. The similarity of these results shows independent confirmation of each approach.
- Correlation between human quality assessment and the correlation of specific lexical categories in source and target suggests a useful approach to automatically locating problems in translations and categorizing them. The correlation is strongest for verbs and proper nouns, but some other results seem to contradict the results of human annotation and further work is needed to explore the relationship of these measures.

Using translate5 as annotation environment, a substantial database of annotated “near miss” translations for analysis of the “Barriers for high quality machine translation (HQMT)” has been produced. We took the reviewer’s recommendation seriously to measure inter-annotator-agreement in this rather sophisticated task of assigning error tags.

The core results of the analysis presented in more detail in this Deliverable are that the systematic and diagnostic human-centric approach of evaluating translation quality taken by QTLaunchPad leads to new insights that are partly counter-intuitive to wide-spread belief. For example, it turned out that in the high-quality part of translations considered here, the performance of the best SMT and RbMT systems is much closer than one would have expected. At the same time, it is clear that different language combinations show widely divergent patterns of errors that often relate to linguistic properties. For example, word order of nouns is a real issue when translating from German to English while agreement is a severe problem when translating from English to Spanish. In this sense, the database of barriers that comes with this report will be a valuable resource for those MT developers who are aiming at developing substantial solutions for problems of a certain language combination.

D1.3.1 Barriers for High-Quality Machine Translation

3. Experimental setup

3.1. Selection of “near miss” candidates, calibration sets, and additional data

This deliverable required a corpus of “near miss” translations produced by state-of-the-art MT systems for further analysis. Because QTLaunchPad strives to break down the barriers between research and application, two primary options for data selection were considered:

1. Use WMT (Workshop on Machine Translation) data. The advantage of WMT data is that the translations represent the state-of-the art in Machine translation research and are well established in the research community. The disadvantage is that the text type (news data) is rather different from the sorts of technical text typically translated in production MT environments. As a result, using the best-established research data also creates an artificial test. (One major difference between WMT and production environments is the absence of a fixed terminology in WMT data.)
2. Use “customer” from LSPs. The use of domain-specific customer data would represent current production usage of MT better than the translation of general-domain language such as that found in WMT tasks. Production systems would also generally have terminology and other specific linguistic resources (such as training data) and might be expected to deliver better accuracy than general-domain systems. At the same time the project would have little control over the production of data or assurance that best practices were followed.

After an intensive discussion, the QTLP consortium, together with GALA members, decided to use a mix of WMT data—including the best RBMT, SMT, and hybrid (where available) systems—and domain-specific customer data from LSPs for our annotation campaign. We asked the LSPs to provide their own MT data.

To build this corpus, DFKI prepared sets of data from WMT 2013 data (with additional WMT 2012 data used for EN>ES translations). For each language pair the following translations were aligned with the source:

- For EN>ES and EN>DE: human translation (WMT reference translations), top-rated SMT, top-rated RbMT, top-rated hybrid system
- For ES>EN and DE>EN: human translation (WMT reference translations), top-rated SMT, top-rated RbMT (no hybrid system results were available)

Three of these data sets (EN>DE, DE>DE, and EN>ES) were sent to text&form for rating and one (ES>EN) was rated in-house at DFKI. The raters examined both machine and human (reference) translation outputs from WMT and categorized them according to the following scale:

1. “perfect” (i.e., no apparent errors)
2. “almost perfect” or “near miss” (i.e., easy to correct with three or fewer errors)
3. “bad” (more than three errors).

D1.3.1 Barriers for High-Quality Machine Translation

Data set/ranking		Human	SMT	RbMT	Hybrid	Total
DE>EN	<i>Perfect</i>	424	79	40		543
	<i>Near miss</i>	76	250	230		646
	<i>Bad</i>	0	170	229		399
	<i>Total</i>	500	499	499		1498
EN>DE	<i>Perfect</i>	391	15	22	36	464
	<i>Near miss</i>	101	148	237	244	730
	<i>Bad</i>	6	335	239	218	798
	<i>Total</i>	498	498	498	498	1992
EN>ES	<i>Perfect</i>	384	43	30	30	487
	<i>Near miss</i>	216	372	395	402	1385
	<i>Bad</i>	3	188	179	172	542
	<i>Total</i>	603	603	604	604	2414
ES>EN	<i>Perfect</i>	406	129	70		605
	<i>Near miss</i>	91	315	293		699
	<i>Bad</i>	2	55	134		191
	<i>Total</i>	499	499	497		1495

Table 2. WMT rankings by language and system type.

This rating process provided the numbers of sentences shown in **Table 2** (overleaf) for each category (note that a very small number of sentences received unclear ratings, amounting to a total of 9 sentences across all language pairs, and are not included in these figures). Note that more WMT data was annotated for Spanish-to-English because University of Sheffield requested additional data for this language pair for use in the shared task.

After the sentences were rated, 150 near-miss sentences were selected from each data set. Because multiple systems were analyzed, in some cases up to four near-miss translations were available for a single source (SMT, RbMT, hybrid, and human). A pseudo-random selection process was used to select at most one translation for a given source, with the following counts selected from each translation type:

- For EN>ES and EN>DE: 40 from each MT system (120 total MT segments) + 30 human translation segments.
- For DE>EN and ES>EN: 60 from each MT system (120 total MT segments) + 30 human translation segments.

These data sets were then converted to a comma-separated value (CSV) format and loaded into the translate5 system (<http://dfki.translate5.net>) along with an XML file that defined the shared task assessment metric, defined using MQM (see **Section 10.1.4** for more details on this metric). This data was then made available for annotation by LSPs.

3.2. Selection of annotators

For this task, DFKI, in collaboration with GALA, invited a select number of language service providers (LSPs) to be trained in MQM assessment and to analyze texts using MQM. A total of nine LSPs participated in this task, with each LSP analyzing from one to three language

D1.3.1 Barriers for High-Quality Machine Translation

pairs. These LSPs were selected based on previous experience with MT use and quality procedures. Participating LSPs were paid up to €1000 per language pair, with payment depending on whether they were able to provide additional client data for annotation or not.

The following LSPs participated:

- Beo (EN>ES, EN>DE, DE>EN)
- Hermes (EN>ES)
- iDisc (ES>EN)
- Linguaserve (ES>EN)
- Logrus (EN>ES)
- Lucy (DE>EN, ES>EN)
- Rheinschrift (EN>DE)
- text&form (EN>ES, EN>DE, DE>EN)
- Welocalize (EN>DE)

3.3. Training of annotators

Representatives from most of the selected LSPs participated in a webinar on December 12, 2013 that introduced them to the translate5 system and the annotation task. This webinar was recorded and all annotators were instructed to view the webinar online and return any questions to DFKI staff. They were also provided with a set of annotation guidelines (see [Section 10](#)) that they were encouraged to read and consult as needed in the translation task. These guidelines were revised from the pilot corpus task (D1.2.1) to reflect lessons learned in that task.

The calibration data set served as a training set for annotators as well as a set for evaluating inter-annotator agreement.

3.4. Annotation of the calibration set (Phase I)

Each LSP was asked to annotate the calibration data set. LSPs were asked to provide, if possible, double annotation of all data, but in most cases they were able to provide only single annotation. They were instructed to tag all issues found in the calibration data sets and note any problems they found. In the end, the calibration data sets were annotated the following number of times:

- DE>EN: 3 annotations by 3 LSPs
- EN>DE: 5 annotations by 4 LSPs (1 LSP provided double annotation)
- EN>ES: 4 annotations by 4 LSPs
- ES>EN: 4 annotations by 3 LSPs (1 LSP provided double annotation)

During this process a number of questions were raised by the annotators and answered, resulting in proposed changes for the next round of annotation (see [Section 7](#)).

D1.3.1 Barriers for High-Quality Machine Translation

3.5. Selection and annotation of customer data and additional WMT data (Phase II)

For the second phase participating LSPs were requested to provide raw MT output of customer data for annotation. They were asked to rate this data according to the same criteria used to rate WMT data until they had 200 near-miss translations to evaluate. In a few cases the LSPs were not able to obtain clearance to use customer data; in these cases DFKI either provided pre-existing customer data from the TaraXü project or the LSPs were able to obtain open (non-customer, but still domain-specific) data and translate it. They were requested to use whatever production MT system they would normally use and to use the raw output.

In addition, for each combination of an LSP and language pair, DFKI selected an additional 100 sentences from the WMT data. These data sets were selected so as to provide no overlap with the calibration set or with data provided to other annotators. In some cases the same source sentences were repeated, but the annotators were provided with translations they had not previously seen.

The annotators were asked to annotate this data with the same metric and criteria as in the earlier annotation of the calibration data.

3.6. Analysis

After annotation was complete, the annotations were examined from a variety of perspectives. The analysis of this data constitutes the bulk of this Deliverable.

D1.3.1 Barriers for High-Quality Machine Translation

4. Results of the Annotation Task

As noted above, the calibration phase consisted of 150 segments of WMT data annotated multiple times. This multiple annotation allowed us to assess inter-annotator agreement. In addition, in Phase II, additional WMT data was assessed, in addition to customer data. This variety of data, as well as the number of system types (including human translation) allows us to address a variety of factors in assessing quality barriers. The following analysis is ongoing and will be extended as more data becomes available in the Shared Task activities.

4.1. Error profiles

The first analysis was to consider the number of issues found in the various texts to develop error profiles by language and system type. The raw results of this analysis are presented on the following pages in **Table 3**, **Figure 4**, and **Figure 5**. This section addresses some notable points.

4.1.1. Overall issues

Considering all language pairs and translation methods together (see the last columns in **Table 3**), four of the 21 issue types constitute the bulk (59%) of issues:

- Mistranslation: 21%
- Function words: 15%
- Word order: 12%
- Terminology: 11%

The remaining issues all fall below 10% and some of them occur so infrequently as to offer little insight into text quality. For example, although the metric used distinguishes between **Punctuation** and other **Typography** issues, “plain” **Typography** was used so infrequently that the distinction is not important. (As a result this distinction will be eliminated for further Shared Task activities. For more on changes to the metric, see **Section 7**.)

An examination of instances tagged with **Mistranslation** and **Terminology** shows that these two issues are easily confusable and that the distinction is problematic because no terminology resources were available for the assessment of WMT data, making the distinction into guesswork. In many cases the two seemed to be used interchangeably, but an examination of the length of spans tagged with these two issue types reveals that the annotators used **Terminology** for shorter spans than **Mistranslation**, with **Terminology** typically used for single nouns or noun phrases and **Mistranslation** used for larger syntactic units, although both were used for single words as well. The average length of spans tagged for **Mistranslation** was 2.13 words (with a standard deviation of 2.43), versus 1.42 (with a standard deviation of 0.82) for **Terminology**. While a length differential is to be expected, a close examination of actual instances tagged for each category shows that the intended distinction (between general language translation problems for **Mistranslation** and domain-specific language errors for **Terminology**) was not clearly followed by annotators, likely representing a failure to communicate the distinction properly and the general impossibility of determining the category without access to domain-specific terminology.

D1.3.1 Barriers for High-Quality Machine Translation

	de-en			en-de			en-es			es-en			Total			
	WMT	Customer		WMT	Customer		WMT	Customer		WMT	Customer					
Accuracy branch																
Accuracy	0	0%	0	0%	22	1%	3	0%	1	0%	8	0%	2	0%	38	0%
Terminology	102	13%	91	16%	368	9%	72	7%	120	16%	215	11%	208	10%	615	15%
Mistranslation	163	22%	141	26%	1207	28%	355	33%	175	23%	309	15%	390	20%	471	11%
Omission	60	8%	19	3%	218	5%	69	6%	51	7%	97	5%	89	4%	175	4%
Addition	20	3%	8	1%	49	1%	8	1%	23	3%	51	3%	30	2%	51	1%
Untranslated	18	2%	11	2%	41	1%	7	1%	1	0%	53	3%	32	2%	300	7%
Accuracy subtotal	363	48%	270	49%	1905	44%	514	48%	371	49%	733	36%	751	38%	1614	39%
Fluency branch																
Fluency	0	0%	0	0%	2	0%	0	0%	17	2%	151	7%	2	0%	2	0%
Style/register	12	2%	5	1%	273	6%	39	4%	22	3%	77	4%	211	11%	169	4%
Typography	1	0%	0	0%	6	0%	9	1%	5	1%	31	2%	2	0%	53	1%
Punctuation	64	8%	20	4%	346	8%	25	2%	26	3%	79	4%	60	3%	171	4%
Typography subtotal	65	9%	20	4%	352	8%	34	3%	31	4%	110	5%	62	3%	224	5%
Spelling (general)	5	1%	2	0%	18	0%	8	1%	12	2%	27	1%	12	1%	33	1%
Capitalization	9	1%	7	1%	55	1%	7	1%	19	3%	72	4%	2	0%	19	0%
Spelling subtotal	14	2%	9	2%	73	2%	15	1%	31	4%	99	5%	14	1%	52	1%
Grammar (general)	2	0%	50	9%	484	11%	5	0%	6	1%	11	1%	63	3%	127	3%
Morphology (word form)	19	3%	10	2%	168	4%	6	1%	2	0%	7	0%	2	0%	4	0%
Part of speech	10	1%	10	2%	15	0%	19	2%	13	2%	51	3%	39	2%	63	2%
Agreement	21	3%	11	2%	184	4%	94	9%	50	7%	101	5%	67	3%	85	2%
Word order	95	13%	72	13%	361	8%	94	9%	33	4%	202	10%	138	7%	801	19%
Function words	135	18%	68	12%	186	4%	157	15%	105	14%	346	17%	560	28%	757	18%
Tense/aspect/mood	16	2%	27	5%	298	7%	88	8%	77	10%	125	6%	76	4%	237	6%
Grammar subtotal	298	39%	248	45%	1696	39%	463	43%	286	38%	843	41%	945	48%	2074	50%
Unintelligible	6	1%	0	0%	8	0%	16	1%	0	0%	19	1%	2	0%	5	0%
Fluency subtotal	395	52%	282	51%	2404	56%	567	52%	387	51%	1299	64%	1236	62%	2526	61%
TOTAL	758	100%	552	100%	4309	100%	1081	100%	758	100%	2032	100%	1987	100%	4140	100%

Table 3. MQM issues by language pair and data type (WMT or customer data)

D1.3.1 Barriers for High-Quality Machine Translation

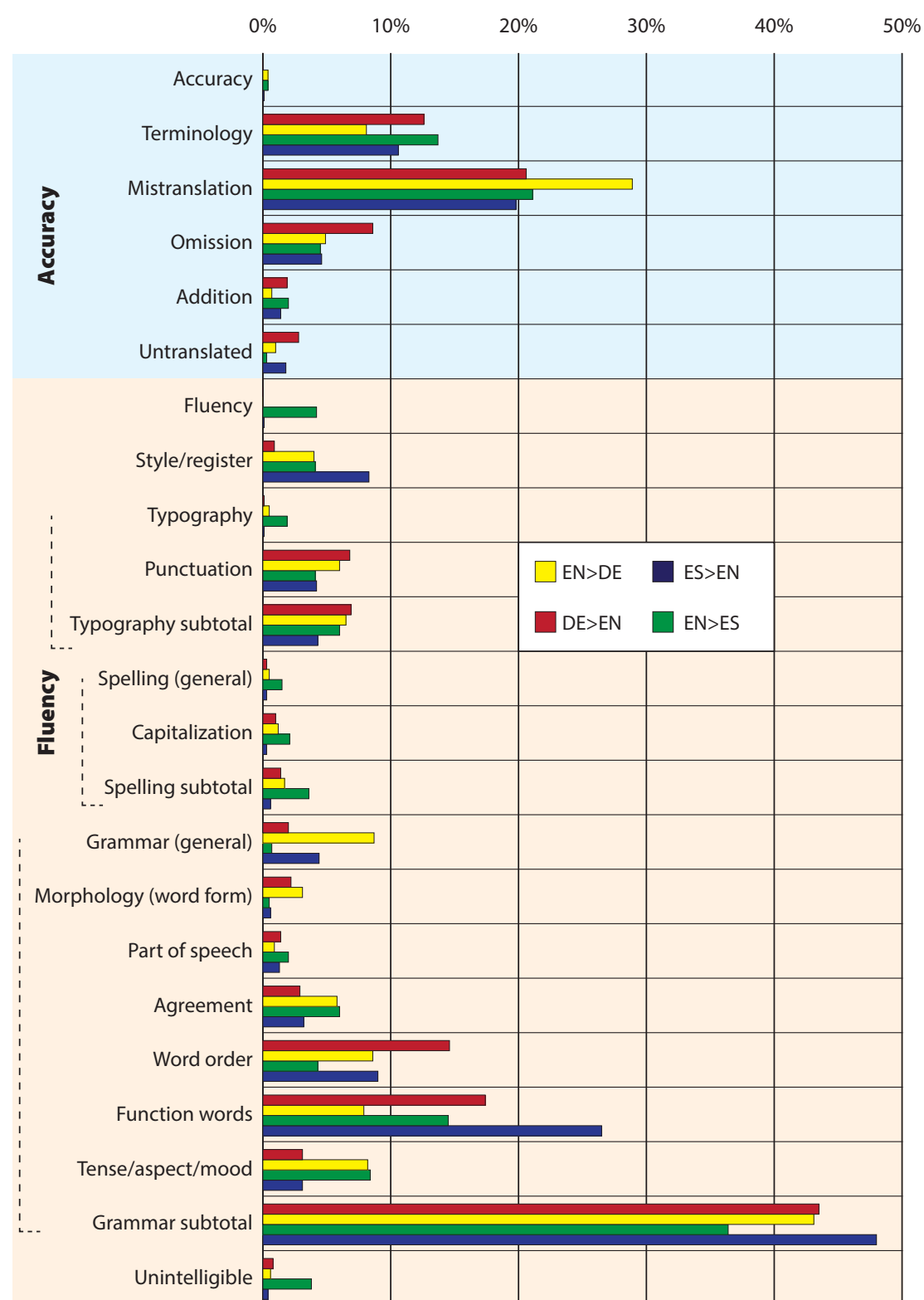


Figure 4. Error rates for all MT results combined, by language pair

D1.3.1 Barriers for High-Quality Machine Translation

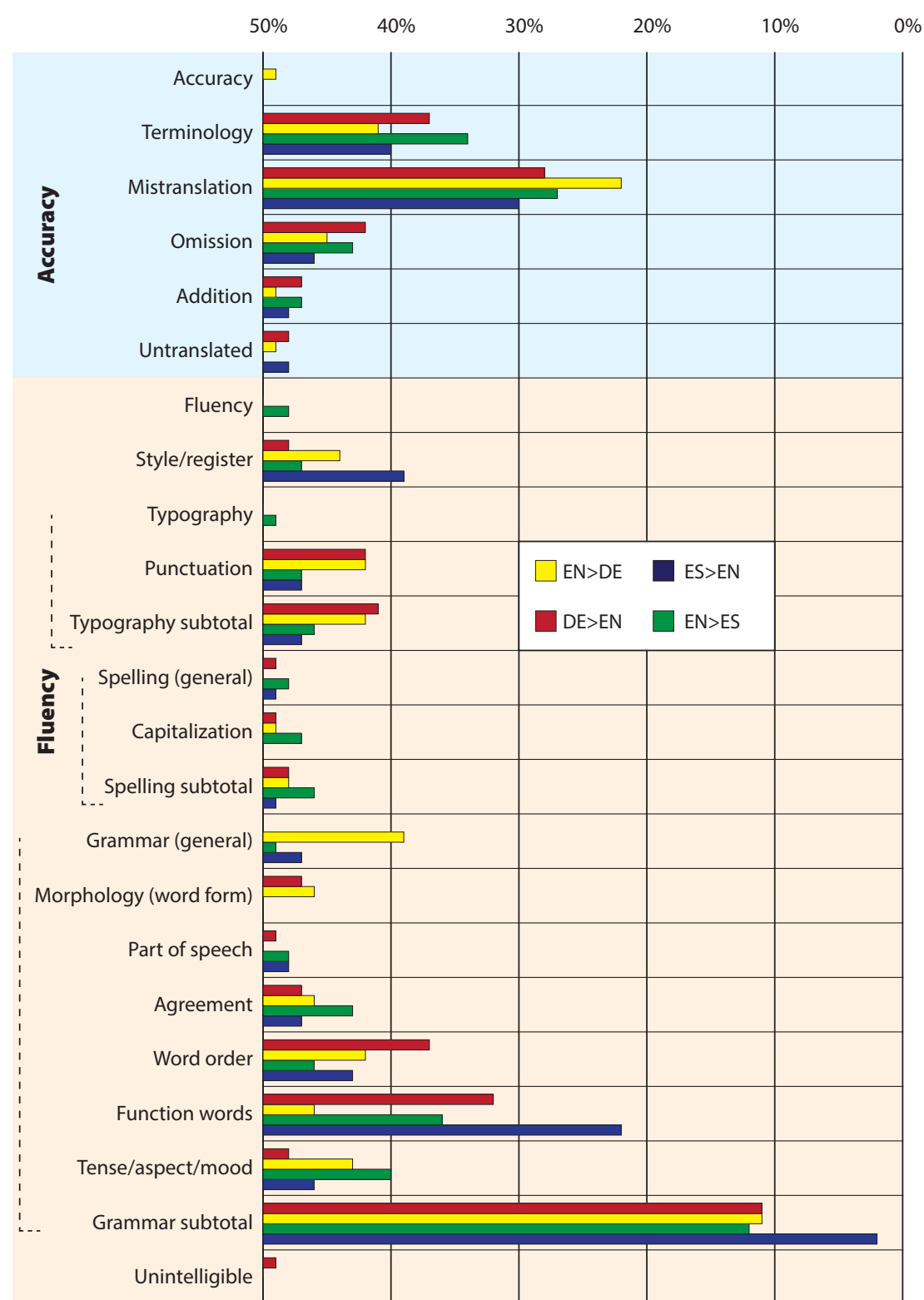


Figure 5. Error distribution for WMT data by language pair.

D1.3.1 Barriers for High-Quality Machine Translation

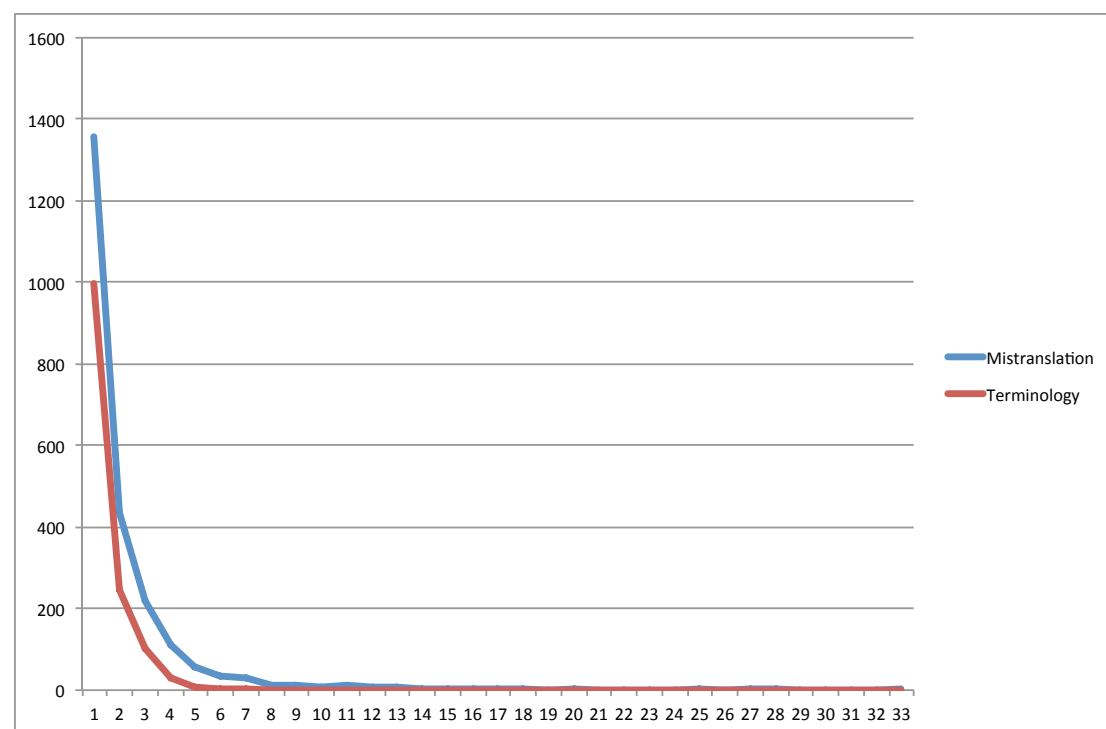


Figure 6. Span lengths for Mistranslation and Terminology.

The longest span tagged for **Mistranslation** was 33 words long, versus 7 for **Terminology**; in the latter case, the mistranslation was an entire phrase, not a term. Overall the two issue types follow a very similar curve (see **Figure 6**), with **Terminology** skewed in its distribution to shorter lengths.

Since one- and two-word **Mistranslations** and **Terminology** errors account for 82% of all errors in these categories, content errors at or near the lexical level thus constitute approximately 25% of all annotated errors (36% for **Terminology** and **Mistranslation** combined multiplied by 82% = 26%). This finding suggests that work on improving MT lexical resources and lexical coverage in training data remains a relatively low-hanging fruit for improving MT output.

4.1.2. Issues by language pair

An examination of the MQM analysis by language pair provides the results depicted in **Figure 7** for MT only. It reveals a few notable results:

1. **Word order** is particularly an issue for German>English translation. German has a relatively free word order and relies on inflectional morphology to carry syntactic load. Thus translations from German that do not rely on linguistic knowledge are likely to have word order errors. In addition, for all language pairs, SMT systems were the most likely to show word order errors (13% of issues overall for SMT) when compared to RbMT (8%) systems hybrid (5%). This result suggests that SMT systems in particular need linguistic knowledge. Anecdotal evidence shows that system developers are moving in the direction of preprocessing (e.g., reordering, combining verbal constituents) for this reason.

D1.3.1 Barriers for High-Quality Machine Translation

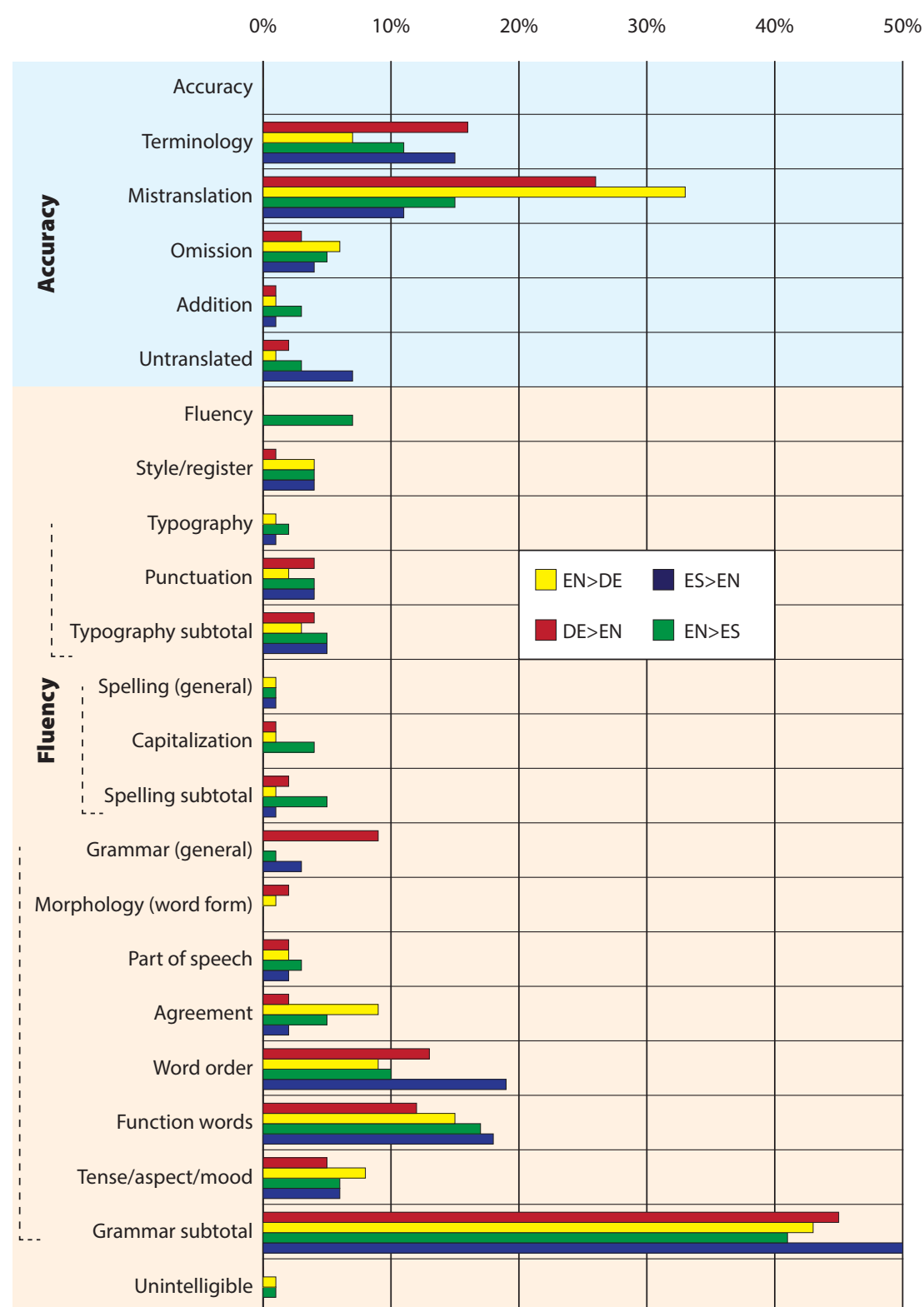


Figure 7. Error distribution for customer data by language pair

D1.3.1 Barriers for High-Quality Machine Translation

2. **Grammar** issues were highest for Spanish>English translation, largely because **Function words** were disproportionately an issue for this language pair. Function words should, in general, be tractable for statistical methods, so additional training data should help improve this situation. However, in some cases it seems that the **Function words** category may have been chosen when **Agreement** may have been a more appropriate choice (e.g., the wrong gender of an article was chosen), in which case linguistic processing rules may be required.
3. Although **Mistranslation** appears to be a particularly problematic issue for English>German, if **Mistranslation** and **Terminology** are lumped together, as discussed above, it appears that the combined categories appear at roughly similar rates to the other language pairs. In part, the English>German evaluators may have adhered more closely to the intended distinction between the two categories than did the other reviewers, since the difference in average length of spans marked with **Mistranslation** (2.83 words) versus those marked with **Terminology** (1.29 words) is much greater than for the other pairs (1.79 words vs. 1.74 for DE>EN, 1.72 vs. 1.55 for EN>ES, and 1.54 vs. 1.33 for ES>EN). Thus the EN>DE pair shows a distinct interpretation of these categories.
4. The type of text translated matters for the language pairs. In particular, for customer data in ES>EN the error distribution is markedly different for customer versus WMT data in some details. These differences may indicate the result of the three specific MT systems used in this pair. Although results for customer data are not broken out, ES>EN used three separate MT systems, two SMT and one RbMT, while all the other pairs used only a single type of system (either SMT or RbMT), which may account for the differences observed for this pair.

4.1.3. Issues by MT type

When the results are broken out by MT type (for WMT data only) in combination with language, other patterns emerge. The raw results are provided in **Table 4** and **Table 5**, and the aggregate for all language pairs combined in **Table 6**.

Reading this data, however, is difficult. A useful way of visualizing the differences to see the strengths and weaknesses of particular translation methods used in the corpus can be seen in **Figure 8**. In this figure the average percentage of issues encountered by a particular method are compared to the average for all other methods. For example, for the **Function words** category the systems received the following scores:

- Human translation: 12.2%
- SMT: 10.2%
- RbMT: 21.3%
- Hybrid: 11.0%

To obtain the figure for human translation shown in **Figure 8**, the rate for human translation given above (12.2%) was subtracted from the average of SMT, RbMT, and Hybrid (14.2%) to yield 2.0%. This number thus indicates that human translation outperforms the average of the other methods by 2% of total errors (i.e., its percentage of errors in this category was 2% lower than the average). Positive bars thus indicate better performance and negative bars worse performance relative to the other options.

Examination of this figure reveals some observations about system types:

D1.3.1 Barriers for High-Quality Machine Translation

	DE-EN				EN-DE			
	HT	SMT	RbMT		HT	SMT	RbMT	Hybrid
Accuracy branch								
Accuracy	0	0%	0	0%	0	0%	3	0%
Terminology	16	13%	36	9%	5	16%	99	11%
Mistranslation	35	28%	66	16%	26	24%	271	31%
Omission	2	2%	58	14%	6	3%	11	1%
Addition	8	6%	11	3%	14	1%	5	1%
Untranslated	0	0%	12	3%	0	3%	5	1%
Accuracy subtotal	61	49%	183	45%	51	49%	394	43%
Fluency branch								
Fluency	0	0%	0	0%	0	0%	1	0%
Style/register	10	8%	3	1%	23	22%	35	4%
Typography	0	0%	0	0%	1	1%	4	0%
Punctuation	17	14%	38	9%	16	15%	45	5%
<i>Typography subtotal</i>	<i>17</i>	<i>14%</i>	<i>38</i>	<i>9%</i>	<i>17</i>	<i>16%</i>	<i>49</i>	<i>6%</i>
Spelling	3	2%	1	0%	0	0%	4	0%
Capitalization	2	2%	2	0%	0	0%	11	1%
<i>Spelling subtotal</i>	<i>5</i>	<i>4%</i>	<i>3</i>	<i>1%</i>	<i>0</i>	<i>0%</i>	<i>15</i>	<i>2%</i>
Grammar	0	0%	6	1%	4	4%	58	7%
Morphology (word form)	2	2%	9	2%	4	4%	17	2%
Part of speech	3	2%	10	2%	0	0%	4	0%
Agreement	2	2%	11	3%	0	0%	30	3%
Word order	10	8%	63	16%	0	0%	60	7%
Function words	10	8%	60	15%	2	2%	89	10%
Tense/aspect/mood	5	4%	14	3%	4	4%	116	13%
<i>Grammar subtotal</i>	<i>32</i>	<i>26%</i>	<i>173</i>	<i>43%</i>	<i>14</i>	<i>13%</i>	<i>374</i>	<i>43%</i>
Unintelligible	0	0%	5	1%	0	0%	4	0%
Fluency subtotal	64	51%	222	55%	54	51%	478	57%
TOTAL	125	100%	405	100%	105	100%	872	100%

Table 4. MQM issue types for WMT data by system type (DE>EN and EN>DE). This table is continued in Table 5.

D1.3.1 Barriers for High-Quality Machine Translation

	EN-ES				ES-EN			Total		
	HT	SMT	RbMT	Hybrid	HT	SMT	RbMT			
Accuracy branch										
Accuracy	0	0%	1	0%	3	1%	1	0%	14	0%
Terminology	11	12%	41	9%	83	16%	64	16%	10	9%
Mistranslation	23	26%	89	20%	105	20%	97	24%	17	15%
Omission	5	6%	30	7%	22	4%	10	2%	7	6%
Addition	7	8%	8	2%	12	2%	7	2%	1	1%
Untranslated	0	0%	2	0%	2	0%	0	0%	0	0%
Accuracy subtotal	46	51%	171	39%	227	43%	179	44%	35	31%
Fluency branch										
Fluency	2	2%	19	4%	27	5%	12	3%	0	0%
Style/register	2	2%	15	3%	19	4%	23	6%	18	16%
Typography	1	1%	5	1%	11	2%	10	2%	0	0%
Punctuation	4	4%	16	4%	23	4%	18	4%	8	7%
Typography subtotal	5	6%	21	5%	34	6%	28	7%	8	7%
Spelling	1	1%	5	1%	12	2%	3	1%	1	1%
Capitalization	1	1%	4	1%	10	2%	15	4%	0	0%
Spelling subtotal	2	2%	9	2%	22	4%	18	4%	1	1%
Grammar	0	0%	6	1%	1	0%	3	1%	1	1%
Morphology (word form)	0	0%	5	1%	0	0%	2	0%	0	0%
Part of speech	0	0%	12	3%	11	2%	4	1%	3	3%
Agreement	3	3%	38	9%	26	5%	19	5%	4	4%
Word order	4	4%	23	5%	21	4%	15	4%	9	8%
Function words	12	13%	47	11%	89	17%	64	16%	29	25%
Tense/aspect/mood	14	16%	55	12%	32	6%	28	7%	6	5%
Grammar subtotal	33	37%	186	42%	180	34%	135	33%	52	46%
Unintelligible	0	0%	23	5%	21	4%	8	2%	0	0%
Fluency subtotal	44	49%	273	61%	303	57%	224	56%	79	69%
TOTAL	90	100%	444	100%	530	100%	403	100%	114	100%

Table 5. MQM issue types for WMT data by system type (EN>ES and ES>EN) and overall. (This table is a continuation of Table 4.)

D1.3.1 Barriers for High-Quality Machine Translation

	HT		SMT		RbMT		Hybrid	
Accuracy branch								
Accuracy	0	0%	3	0%	7	0%	4	0%
Terminology	42	10%	130	7%	426	12%	131	10%
Mistranslation	101	23%	333	19%	810	24%	365	29%
Omission	20	5%	221	13%	82	2%	35	3%
Addition	30	7%	38	2%	39	1%	10	1%
Untranslated	0	0%	32	2%	42	1%	11	1%
Accuracy subtotal	193	44%	757	43%	1406	41%	556	44%
Fluency branch								
Fluency	2	0%	19	1%	29	1%	12	1%
Style/register	53	12%	30	2%	219	6%	69	5%
Typography	2	0%	6	0%	18	1%	16	1%
Punctuation	45	10%	140	8%	133	4%	61	5%
Typography subtotal	47	11%	146	8%	151	4%	77	6%
Spelling	5	1%	8	0%	23	1%	9	1%
Capitalization	3	1%	9	1%	32	1%	31	2%
Spelling subtotal	8	2%	17	1%	55	2%	40	3%
Grammar	5	1%	74	4%	150	4%	85	7%
Morphology (word form)	6	1%	39	2%	30	1%	38	3%
Part of speech	6	1%	36	2%	33	1%	17	1%
Agreement	9	2%	98	6%	121	4%	81	6%
Word order	23	5%	225	13%	257	8%	74	6%
Function words	53	12%	180	10%	726	21%	140	11%
Tense/aspect/mood	29	7%	109	6%	207	6%	70	6%
Grammar subtotal	131	30%	761	43%	1524	45%	505	40%
Unintelligible	0	0%	38	2%	29	1%	13	1%
Fluency subtotal	241	56%	1011	57%	2007	59%	716	56%
TOTAL	434	100%	1768	100%	3413	100%	1272	100%

Table 6. Comparison of average MQM results by system

- SMT
 - SMT does particularly well in the **Terminology/Mistranslation** area when compared to other methods.
 - SMT substantially outperformed all other translation methods, including human translators in **Style/register**.
 - SMT is the most likely to drop content (**Omission**); without this tendency, it would be the most accurate translation method considered. This impact is the second largest for any issue type and method combination examined.
 - SMT is the most likely method to get **Function words** right.
 - By contrast, SMT is weak in **Grammar**, largely because it does poorly with **Word order** (and not just when dealing with German).
- RbMT
 - RbMT showed the worst results for **Terminology**, but was the least likely method to omit content (**Omission**). Indeed, RbMT had the highest overall **Accuracy**.
 - RbMT had the lowest rating for **Grammar**, almost entirely because it performed very weakly with respect to **Function words**. Since function words are one of the most idiosyncratic aspects of human language, rule-based approaches have difficulty with them. Here statistical enhancements (moving RbMT more in the

D1.3.1 Barriers for High-Quality Machine Translation

(Positive numbers indicate better performance)

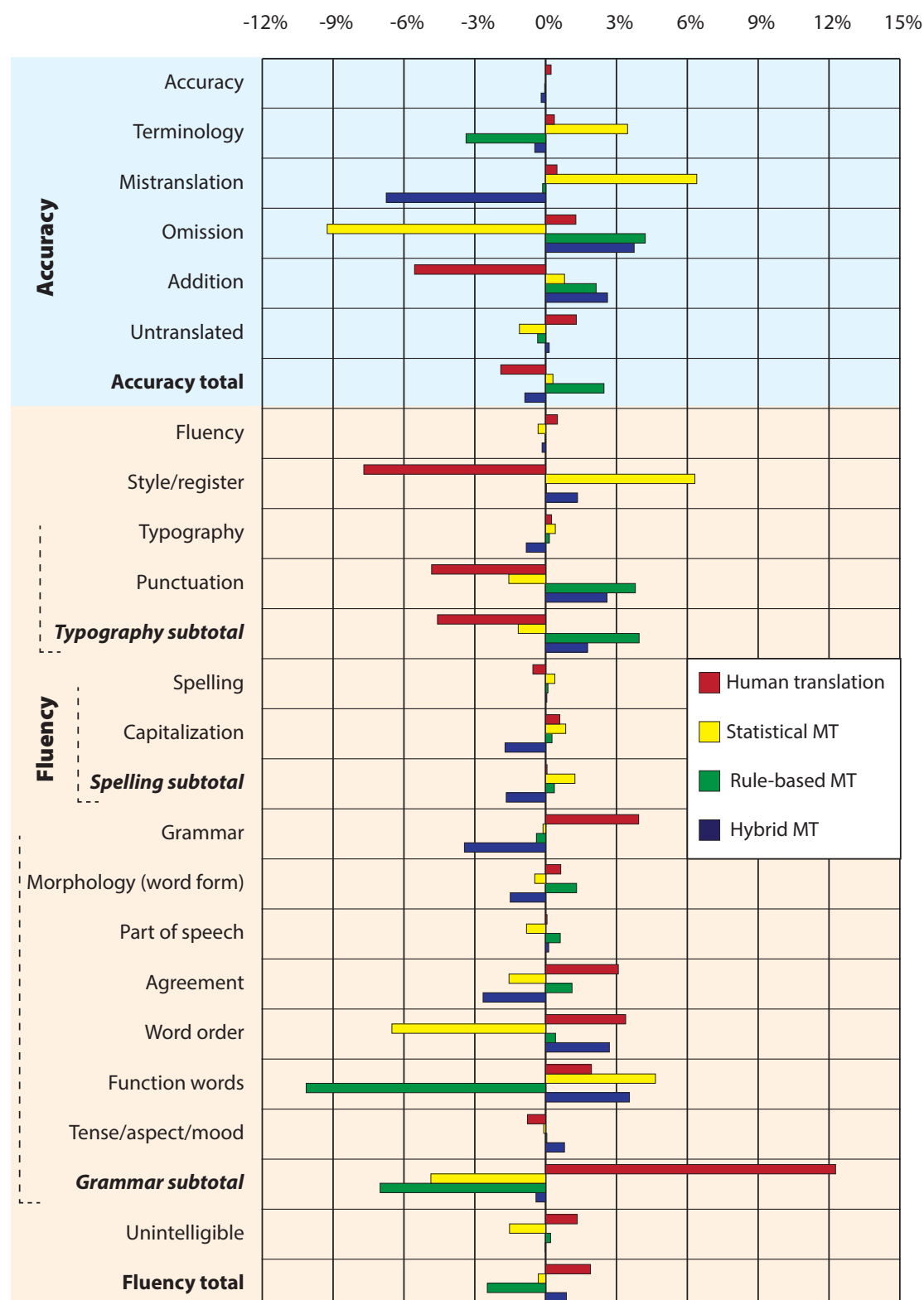


Figure 8. Comparison of MQM issue types by system compared to average of other systems.

D1.3.1 Barriers for High-Quality Machine Translation

- direction of hybrid systems) would offer considerable potential to improve output.
- The impact of **Function words** is so extreme that it not only drags down the overall Grammar performance, but also results in the lowest overall **Fluency** scores, even though it would otherwise outperform the other methods, including human translation in this aspect. (Revising scores to remove **Function words** results in the following figures: human translation: 1.0%; SMT: -3.0%; RbMT: 3.2%; Hybrid: -1.1%). The impact here is the largest of any issue type and method combination considered in this Deliverable. Note as well that RbMT does not include specific modules to correct for fluency (in contrast to SMT, which uses language models), which means that this problem is likely to remain particularly troublesome for RbMT.
 - Hybrid:
 - Because the hybrid system results were available only for EN>DE and EN>ES, it is impossible to say how representative the results would be for cases with a non-English source, preventing full comparison with the other systems. As a result, any general statements here must be understood to be tentative.
 - Not surprisingly, the hybrid MT system evaluated tends to perform in between the SMT and RbMT systems evaluated in most respects. It is, however, by far the most likely method to return mistranslated texts and seems to perform the worst on lexical matters overall.
 - As the subtotals for **Accuracy**, **Fluency**, and **Grammar** show, the hybrid system tends to hew more closely to the average than the other systems.

When compared to the results of human translation assessment, it is apparent that all of the near-miss MT methods are somewhat more accurate than near-miss human translation and significantly less grammatical. Humans are far more likely to make typographic errors, but otherwise are much more fluent. Note as well that humans are more likely to add information to translations than MT systems, perhaps in an effort to render texts more accessible. Thus, despite substantial differences, all of the MT systems are overall more similar to each other than they are to human translation. However, when one considers that a far greater proportion of human translation sentences were in the “perfect” category and a far lower proportion in the “bad” category, and that these comparisons focus only on the “near miss sentences,” it is apparent that outside of the context of this comparison, human translation still maintains a much higher level of **Accuracy** and **Fluency**. (In addition, a number of the annotators commented on the poor level of translation evident in the WMT human translations, indicating that the difference for high-end human translation output may be even more marked.)

4.2. Inter-Annotator Agreement

In the Y1 review, the reviewers requested that the project consider Inter-Annotator Agreement (IAA) as part of assessing whether MQM is a fair, valid, and reliable method for assessing translation quality. Using the results of the calibration task we had between three and five independent annotations of the calibration data for each language pair.

After examining the data, we found that there was no obviously applicable measure for IAA that would account for all variables. Previous work in IAA for MT evaluation had tended to focus on much simpler measures (such as Likert scales used in the quality measurement

D1.3.1 Barriers for High-Quality Machine Translation

tasks with WMT data). By contrast, MQM annotation is significantly more complex because of the following issues:

- It operates at the span level, and different annotators may disagree on what constitutes a “minimal span,” leading to different decisions as to what should be marked. Should differences in span scope be considered inter-annotator disagreements?
- Annotators must agree on whether or not an error exists at all. Our analysis shows that annotators frequently disagreed on whether a particular span constitutes an error, with multiple annotators seeing an error in one location and other annotators seeing no error. Such disagreements may be attributable to the next point.
- In some cases errors can be analyzed in multiple ways. For example, if there is an agreement error in a text, annotators might disagree about which word(s) constitute the error and thus annotate different spans even while agreeing on the fundamental nature of the error. Thus annotators may superficially appear to disagree on location or even existence of errors due to the need to restrict errors to specific spans.
- Despite having training materials, annotators may find them unclear and differ in their understanding of the meaning of specific issues, leading to disagreement about how to classify individual instances.
- Because issues exist in a hierarchy, different annotators might choose to annotate issues at different levels in the hierarchy.
- Certain issues may be confusable. For instance, **Mistranslation** and **Terminology** were frequently confused, partially because the WMT data used had no defined terminology, leaving it up to the translator to determine whether a word was a term or not based on context and experience.
- The heterogeneous nature of the material (even within WMT data) made it impossible to specify the job requirements very precisely. This left a grey area where one translator might, e.g., find a given sentence “good enough for the job” while another one marks an dubious formulation as Mistranslation.

Thus an ideal IAA measure would need to account for the positioning of marked spans, notation at various levels in the hierarchy, lack of clarity, and unclear cause. **Function words** were noted as particularly problematic, since reviewers often wished to indicate whether they were omitted, added, or incorrect, thus blurring the boundaries between **Fluency** and **Accuracy**.

There is ongoing discussion about the most appropriate method for evaluating IAA. The various approaches considered have generally yielded results that could best be described as “fair”, but not good. To facilitate comparison with the results from other WMT tasks, we opted to use Cohen’s Kappa coefficients,² which have previously been used to assess IAA for the WMT rating task.

² The coefficient is calculated as follows: $\kappa = \frac{P(a) - P(e)}{1 - P(e)}$

where $P(a)$ is probability of actual agreement, i.e., $\sum_k p(a1 = k, a2 = k)$

and $P(e)$ is probability of agreement by chance, i.e., $\sum_k p(a1 = k) * p(a2 = k)$

where k denotes class (in this case the error tag) and $a1$ and $a2$ refer to the two annotators.

D1.3.1 Barriers for High-Quality Machine Translation

Source: *While she's a U.S. citizen, she sees herself as a part of two countries.*

		Während	sie	ein	US-Bürger,	sie	sieht	sich	selbst	als	Teil	der	beiden	Länder.
A1	Mistranslation													
	Word Order													

		Während	sie	ein	US-Bürger,	sie	sieht	sich	selbst	als	Teil	der	beiden	Länder.
A2	Mistranslation													
	Word order													

A1 / A2 = .85 (11/13)




Figure 9. Assessing IAA for an English>German translation (absolute agreement = .85, Kappa IAA = .72)

Source: *A first year PPE student, who, ironically, had been to Eton, said: "You're the daughter of a fascist pig."*

		Un	primer	año	estudiante	de	PPE,	que,	irónicamente,	había	sido	a	Eton,	dijo:	"Es	hija	de	un	cerdo	fascista".
A1	Word order																			
	Mistranslation																			
	Agreement																			

		Un	primer	año	estudiante	de	PPE,	que,	irónicamente,	había	sido	a	Eton,	dijo:	"Es	hija	de	un	cerdo	fascista".
A2	Word order																			
	Mistranslation																			
	Agreement																			

		Un	primer	año	estudiante	de	PPE,	que,	irónicamente,	había	sido	a	Eton,	dijo:	"Es	hija	de	un	cerdo	fascista".
A3	Word order																			
	Mistranslation																			
	Agreement																			

A1 / A2 = .79 (15/19)																				
A1 / A3 = .95 (18/19)																				
A2 / A3 = .84 (16/19)																				

Figure 10. IAA for an English>Spanish translation (absolute agreement average = .86, Kappa IAA = .66)

For this evaluation we considered word-level agreement for MQM classes. If two annotators marked the same word with a class, they were deemed to be in agreement for that word; if they used a different class or one annotator marked a word with an issue class and another marked it as having no issue, they were deemed to be in disagreement.

Figure 9 shows an example from the corpus in which two annotators agreed on the presence and scope of a **Mistranslation** but disagreed on a **Word order** error. Their assessment agrees for 11 of 13 positions, yielding an absolute agreement rate of 0.85. For Kappa coefficients, an agreement by chance is also taken into account. **Mistranslation** is assigned to 4 of 13 words by both raters; **Word order** is assigned to 2 of 13 words by the first rater and is not assigned to any of words by the second rater; and the first rater considered 7 words as correct whereas the second rater considered 9 words as correct; this leads to an agreement by chance of $(4/13 \times 4/13) + (2/13 \times 0) + (7/13 \times 9/13) = 0.46$. Kappa IAA is then obtained as $(0.85 - 0.46) / (1 - 0.46) = 0.72$.

Figure 10 provides a more complex example with three annotators. In this example three different IAA figures are assessed, one for each of the three possible pair-wise comparisons. In this example, Rater 1 and Rater 3 are quite similar with Kappa IAA of 0.89 while Rater 2

D1.3.1 Barriers for High-Quality Machine Translation

	de-en	es-en	en-de	en-es
a1-a2	0.23	0.30	0.36	0.35
a1-a3	0.36	0.18	0.28	0.36
a2-a3	0.29	0.19	0.33	0.28
a1-a4		0.25	0.30	0.33
a2-a4		0.26	0.34	0.36
a3-a4		0.34	0.30	0.35
Average	0.29	0.25	0.32	0.34

Table 7. Kappa coefficients measuring inter-annotator agreement for MQM error annotation

	de-en	es-en	en-de	en-es
WMT 2011	0.32	0.38	0.49	0.37
WMT 2012	0.38	0.36	0.30	0.25
WMT 2013	0.44	0.46	0.42	0.33
Average	0.38	0.40	0.40	0.32

Table 8. Kappa coefficients measuring inter-annotator agreement for WMT ranking task

differs from both of them with IAA of 0.57 with Rater 1 and 0.53 with Rater 3. Although both Rater 2 and Rater 3 identified the same types of errors (and were alike in not identifying the Agreement error identified by Rater 1), they disagreed on the precise spans for those errors, leading to lower IAA.

Note that if a simpler segment-level measure that counts only whether the same issue classes were identified for each segment were used instead, the results would be rather different. In that case the example in **Figure 9** would yield a figure of 0.5 (there would be a total of two issues for the segment and the annotators would agree on one). For the example in **Figure 10**, by contrast, Rater 1 would show the same agreement with Raters 2 and 3 (.67) while Rater 2 and Rater 3 would show perfect agreement (1.0) since they identified the same issues, even though they disagreed on the scope.

It is not clear what is the best measure for IAA in such cases. The word-level analysis may magnify cognitively insignificant differences in scope while ignoring more profound disagreement about the types of errors present. On the other hand, using segment-level analysis that looks simply for the presence or absence of issues between raters may understate agreement about scope. Both approaches also have their advantages: the word-level analysis provides fine-grained information on whether raters agree precisely on the location and nature of errors and the segment-level analysis provides a view on whether raters agree on the presence or absence of types of errors in a segment. In some cases, such as complex **Agreement** errors, the word-level analysis may hide fundamental agreement that the segment-level analysis would find. Conversely, segment-level counts may conceal disagreement if the same issue class is applied to different problems in the text that do not correspond. For example one rater might mark a portion as a **Mistranslation** that a second reviewer does not, while the second rater marks a completely separate portion as a **Mistranslation** that the first rater does not mark at all; this situation would yield 100% agreement at the segment level and 0% agreement at the word level.

D1.3.1 Barriers for High-Quality Machine Translation

Table 7 presents Cohen’s Kappa coefficients for each language pair and each pair of annotators, based on the word-level assessment outlined above. Four annotators were involved in each language pair except for DE>EN, where only three annotators performed the task. (In the case of EN>DE, one calibration set was received after this experiment was completed and this set was not considered.) The results lie between 0.2 and 0.4 and are considered to be “fair”. The overall average is 0.30.

The coefficients for the WMT rating task can be seen in **Table 8**. The overall average is 0.38. Although the results for the WMT ranking task are somewhat higher, MQM error annotation is a considerably more complex task. For the WMT ranking task, intra-annotator agreement (i.e., self consistency of annotators) was calculated as well, and the coefficients did not exceed 0.65, indicating a possible upper bound for IAA in this area. (We were not able to calculate intra-annotator agreement for MQM annotation.)

This result shows that the coefficients for a very complex task are not an obstacle for estimating error distributions of translation systems and language pairs.

D1.3.1 Barriers for High-Quality Machine Translation

5. Comparison Data from Automatic Error Analysis

In order to provide points of comparison/confirmation with the MQM-based assessment, we performed an automatic analysis of WMT data (going beyond what we reported in D1.2.2). While less fine-grained than MQM assessment, this data provides an independent comparison for MQM annotation from similar data and complementary insights at the same time.

The first step of the error analysis was carried out on the French-to-English and English-to-Spanish translation outputs described in Specia (2011) as well as English-to-Spanish data used for training in the Quality Estimation shared task 2013 (Callison-Burch et al. 2012). For each sentence in the 2011 corpora, a human annotator assigned one of four quality levels, one of them being “almost acceptable”. In addition, a portion of the best German-to-English and English-to-German statistical and rule-based translations obtained in the shared task 2011 (Callison-Burch et al. 2011) is used. Corpora statistics can be seen in **Table 9**.

5.1. Edit types

Post-editing operations were analyzed using the Hjerseon (Popović 2011) automatic tool for error analysis, which provides a rough categorisation into five classes:

- word form (agreement errors, some capitalization, and part of speech)
- word order
- omission
- addition
- lexical error (mistranslations, terminology errors, style, punctuation, and any changes to wording)

The results are available in the form of edit rates for each category, i.e., the raw count of errors normalized over the total number of words in the given translation output.

Automatic classification of post-editing operations was performed on all almost acceptable sentences and results in the form of edit rates are presented in **Table 10**. These figures show the relative proportion of each edit type for each language pair and show differences in the overall distribution of edits.

As discussed in D.1.2.2, the dominant edit operation in all cases for almost acceptable translations is correction of lexical choice, and it is especially frequent for English-to-German translation which is generally the most difficult case. The English-German lexical edit rate reaches almost 14%, and lies between 6% and 9% for other translation directions, meaning that for each group of 100 words, between 6 and 9 of them (14 for the German output) have to be post-edited lexically.

For German-to-English translation, the percentage of words where word order is corrected is almost as high (about 8%), indicating that for this translation direction even almost-acceptable translations often contain syntactic problems. For English-to-German translation this edit rate is also relatively high, over 5%, but not significantly higher than other types of edits (which are all above 4%).

Correcting word form seems to be very rare in English outputs, which can be expected due to the relatively impoverished morphology of the English language. For Spanish and Ger-

D1.3.1 Barriers for High-Quality Machine Translation

data set	almost-acceptable sentences	running words
fr-en 2011	1559	40994
en-es 2011	399	8401
en-es 2012	548	13922
de-en 2011	778	18050
en-de 2011	955	18566

Table 9. Data sets used for automatic analysis of post-edit operations.

edit rate (%)	fr-en 2011	en-es 2011	en-es 2012	de-en 2011	en-de 2011
form	1.0	4.2	3.6	1.5	4.1
order	2.4	3.3	3.7	7.6	5.4
omission	3.0	4.2	3.8	5.2	4.2
addition	2.2	3.1	2.7	3.0	4.1
lexical error	6.0	8.9	8.0	8.7	13.8

Table 10. Edit rates for each data set; edit rate is defined as percentage of edited words over the total number of words.

edit rate (%)	fr-en 2011	en-es 2011	en-es 2012	de-en 2011	en-de 2011
local ($r \leq 4$)	1.7	1.9	2.1	2.4	2.8
long ($4 < r < 15$)	0.6	1.2	1.3	4.3	2.2
distant ($r \geq 15$)	0.1	0.2	0.3	0.9	0.4

Table 11. Reordering distances.

man translation outputs, however, this edit rate is higher, being comparable with correcting omissions and deleting added text.

5.2. Reordering distances

In order to better understand word order errors, an analysis of distance between reorderings has been carried out. Reordering intervals are divided into three types: local (less than four positions), long (between four and fifteen positions) and distant (more than fifteen positions).

The corresponding edit rates are shown in **Table 11**. The first observation is that distant reorderings are very rare for all language pairs. Local reorderings are dominant for the French-to-English translations, and slightly more frequent than long ones for English-Spanish and English-German translation. On the other hand, for the German-to-English translation, long-distance reorderings are most frequent.

These results can be presented graphically as shown in **Figure 11**. They reveal that reordering is particularly an issue for cases involving German, especially when German serves as a source. Because German has a relatively free word order, it is especially likely to cause issues with word ordering. The results also show that long-distance reordering (i.e., reordering

D1.3.1 Barriers for High-Quality Machine Translation

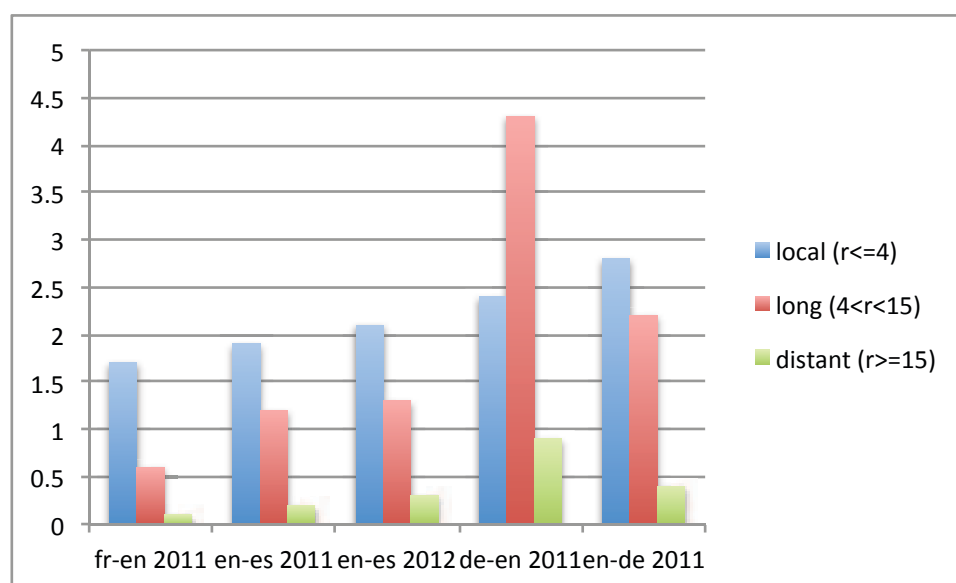


Figure 11. Percentage of reordering edits by distance.

error group edit rate (%)	fr-en 2011	en-es 2011	en-es 2012	de-en 2011	en-de 2011
lex-lex	1.0	2.1	1.5	1.7	3.4
order-order	<1	<1	<1	3.4	1.8
lex-order	<1	<1	<1	1.1	1.1
order-lex	<1	<1	<1	1.2	1.4
miss-miss	<1	<1	<1	1.4	<1
lex-lex-lex	<1	<1	<1	<1	1.0
order-order-order	<1	<1	<1	1.6	<1

Table 12. Groups of errors

across sentence constituent boundaries rather than within them) is particularly likely when German is involved.

5.3. Edit clustering

The results were also examined to see the extent to which different edit types correlated with each other in this data set. The results presented in **Table 12** are word-based, i.e. there is no indication if all the errors are isolated or some of them appear in groups/clusters. Therefore an analysis of bigram and trigram error groups was carried out, and the most frequent groups with their percentages are shown in **Table 12**. The error group rate is defined as the percentage of the particular error group normalized over the total number of the given word groups.

It can be seen that there is no significant clustering in almost acceptable translations (contrary to what is seen in low quality translations). The most frequent error bigrams are two lexical errors in English-to-German and two order errors in German-to-English translation outputs (more than 3%). A certain amount of lexical-order pairs can be found for the German-English language pair in both directions. As for trigrams, one percent in English-to-

D1.3.1 Barriers for High-Quality Machine Translation

German outputs consists of three lexical errors, and 1.6% in German-to-English outputs consists of three reordering errors. All other bigrams and trigrams occur less than 1% and are therefore considered unimportant.

An additional experiment was carried out in order to confirm the absence of error clusters in almost acceptable translations: the six-grams containing at least one edit operation are analyzed. The results showed that the most frequent six-grams for all language pairs are those containing five correct words and one lexical error (between 2 and 3% of all six-grams). In addition, about 1% of German-to-English six-grams contain one order error. This confirms the observation that the majority of editing operations in almost acceptable translations are isolated rather than linked.

6. Relating Translation Quality Barriers to Source-Text Properties

In order to relate the errors in the MT output to linguistic properties of the source texts, we used DELiC4MT (www.computing.dcu.ie/~atoral/delic4mt/) (Toral et al., 2012), an open-source toolkit for diagnostic MT evaluation. It was developed within the CoSyne EU FP7 STREP project, in which DCU was a partner. DELiC4MT's diagnostic dimension derives from its ability to focus on user-defined linguistic checkpoints, i.e., phenomena of the source language that the user decides to focus on for the evaluation of the quality of the MT output. Linguistic checkpoints can correspond to interesting or difficult lexical items and/or grammatical constructions for which a specific translation quality assessment is required. They can be defined at any level of specificity desired by the user, considering lexical, morphological, syntactic and/or semantic information related to the source language. See the Appendix for more information on DELiC4MT.

Instead of being used on its own as was the case in previous work, in this report the diagnostic MT evaluation based on DELiC4MT supplements the other types of analyses of translation quality barriers: we therefore extend the applications of DELiC4MT to new evaluation scenarios in combination with other approaches which focus on error types, quality annotation, and state-of-the-art MT evaluation metrics. Methodologically, this is an appealing addition to the applications and evaluation scenarios on which DELiC4MT has already been tested.

Since the focus of this report is on translation quality barriers, the emphasis of the analysis based on DELiC4MT is on the weaknesses displayed by the MT systems, especially insofar as they contribute to quality boundaries between poor, medium and good translations. Although one could argue that identifying aspects hindering quality is the main purpose of diagnostic MT evaluation, comparative MT evaluation studies in particular tend to focus on aspects contributing to successful MT performance, rather than on problematic aspects.

On the other hand, by concentrating on the unsatisfactory treatment of PoS-based linguistic checkpoints for the various language combinations with different MT systems, and considering the variation between quality rankings, we put forward a strong case for the relevance of diagnostic evaluation to the removal of translation quality barriers. In this respect, this is a first attempt to test the contribution that DELiC4MT can make to the broader effort of identifying and breaking down translation quality barriers, in combination with other approaches to MT quality evaluation.

The diagnostic dimensions of MT quality investigated with DELiC4MT are considered according to two main variables:

- **MT system type.** This variable enables us to explore the performance of the different types of MT software available, namely statistical (SMT), rule-based (RBMT) and hybrid, on each of the source-language checkpoints. For each language pair, we are therefore able to have a clear view of the quality of the various MT systems, broken down according to a range of checkpoints as salient linguistically-motivated morphosyntactic units of evaluation.
- **Quality ranking.** This variable concerns the rankings assigned by human evaluators to the output of each MT system, whereby each sentence was rated as perfect (rank 1), "near miss" (rank 2), or poor (rank 3). DELiC4MT was previously used to evaluate overall output from MT systems, but in this report it is used in conjunction with

D1.3.1 Barriers for High-Quality Machine Translation

human ranking scores to evaluate the performance of MT systems on each checkpoint separately for each quality band. Both of these variables lend themselves to comparative qualitative evaluations, which will be investigated in what follows.

Following some preliminary tests, we decided to restrict this part of the analysis to linguistic checkpoints consisting of individual part-of-speech (PoS) items. This level of detail for the diagnostic evaluation was deemed sufficiently fine-grained to obtain interesting and useful information that could also be combined with the rest of the analyses, e.g., the errors found in the MT output and analyzed with MQM. It should also be noted that using individual PoS items for the linguistic checkpoints avoided the data sparseness problems that we would have run into using more elaborate and specific linguistic checkpoints, given the limited amount of data available (see [Section 6.4](#)).

On the basis of these considerations and some preliminary explorations of the potential possibilities, we identified 9 suitable linguistic checkpoints, consisting of the following individual PoS items: adjective (ADJ), adverb (ADV), determiner (DET), noun-common (NOC), noun-proper (NOP), noun (NOU, combining NOC and NOP), particle (PAR), pronoun (PRO) and verb (VER). For a given checkpoint, the scores given below express the ratio between all the instances of the checkpoint detected on the source side and those that were translated correctly by the MT system. Thus, the higher the score for a checkpoint, the higher the quality of the translations in the MT output for the words that correspond to that linguistic phenomenon (in this study, PoS class) in the input. In addition, for each language pair, we provide an overall score (AVG), by averaging the scores of 7 checkpoints (NOC and NOP are not considered to calculate the average as they overlap with NOU), as a general indication of MT quality.

6.1. Data and processing

In order to pre-process the data for DELiC4MT, we PoS tagged the source and target sides of the reference. Freeling³ was used to this end for English and Spanish, while TreeTagger⁴ was used for German. Finally, the source and target sides of the reference were word aligned with GIZA++. As the reference datasets are rather small for word alignment, in order to obtain alignments of higher quality, they were appended to a bigger corpus of the same domain (news commentary), before performing word alignment. Once the alignment was ready, we extracted the subset that corresponds to the sentences of the reference set.

Given the set-up required to perform diagnostic MT evaluation with DELiC4MT, we were able to conduct this analysis only on the data sets for which human reference translations were available, i.e. the WMT 2012 data set from which the calibration set for MQM evaluation was drawn. **Table 13** shows the data used for the evaluation, detailing the number of sentences and the types of MT systems available for each language direction. For each of the four translation directions, the diagnostic evaluation concerns the very same input when comparisons of MT systems take place on the whole input data. On the other hand, this is not the case for the evaluations considering the MT output in each quality rankings because the DELiC4MT-based evaluation was run separately on a subset of the input, corresponding to the quality evaluation of the raters. For example, the subset of rank 1 sentences translated

³ <http://nlp.lsi.upc.edu/freeling/>

⁴ <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

D1.3.1 Barriers for High-Quality Machine Translation

Language Direction	Number of sentences	MT systems
EN>ES	500	SMT, RBMT, hybrid
ES>EN	203	SMT, RBMT
EN>DE	500	SMT, RBMT, hybrid
DE>EN	500	SMT, RBMT

Table 13. Datasets used for the diagnostic evaluation

with system 1 for EN>ES is different from (although partially overlapping with) the subset of the same rank and language direction for the RbMT system.

The sections that follow focus on each language pair in turn, providing scores for the individual linguistic checkpoints, which have a more fine-grained diagnostic meaning, and then their aggregate average value as a global indicator of overall MT quality. After going through the specific analysis for each language pair, we conclude with a brief comparison between the results obtained using DELiC4MT and the findings yielded by the other evaluation methods presented in this report.

It is important to note that the following analysis compares sentences *within* each quality band, regardless of how many are in that quality band. As a result, even if one system ranks higher than another in all checkpoints for rank 1 (the highest quality band), the other system may have more sentences in the quality band overall, thus indicating an overall higher quality. The following figures can be used only to compare linguistic features of sentences within a given quality band, not to make overall predictions about quality.

6.2. Diagnostic Evaluation Results

6.2.1. Diagnostic Evaluation Results for EN>ES

(See **Figure 12** through **Figure 15**.)

One valuable finding worth emphasizing for the EN>ES language pair is that the SMT system is the best overall, even though it receives (virtually) the same scores as the Hybrid system for the PAR, PRO and VER linguistic checkpoints. Overall, looking at the average scores, the SMT system is the clear winner over (in this order) the hybrid and RbMT systems.

Considering only the top-ranking translations (rank 1), the SMT and hybrid systems perform best for different linguistic checkpoints (except for NOC, where there is a tie), and RbMT is on par with SMT only for ADJ and NOC; otherwise it lags clearly behind the other two MT systems. It is particularly striking that for the rank 1 translations the has a much higher score than SMT for the VER checkpoint, corresponding roughly to a 10% improvement in relative terms; the difference is even more marked between the hybrid and RbMT systems, which has the worst performance on VER for the top-quality tier translations.

⁵ At the time this test was run, the ES>EN data had only been partially rated, resulting in a smaller number of data points for this language pair

D1.3.1 Barriers for High-Quality Machine Translation

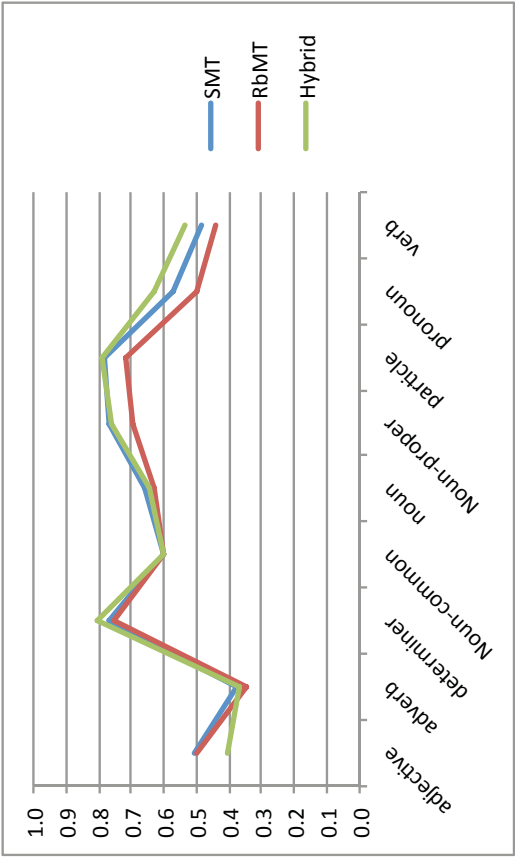


Figure 13. EN>ES results for rank 1

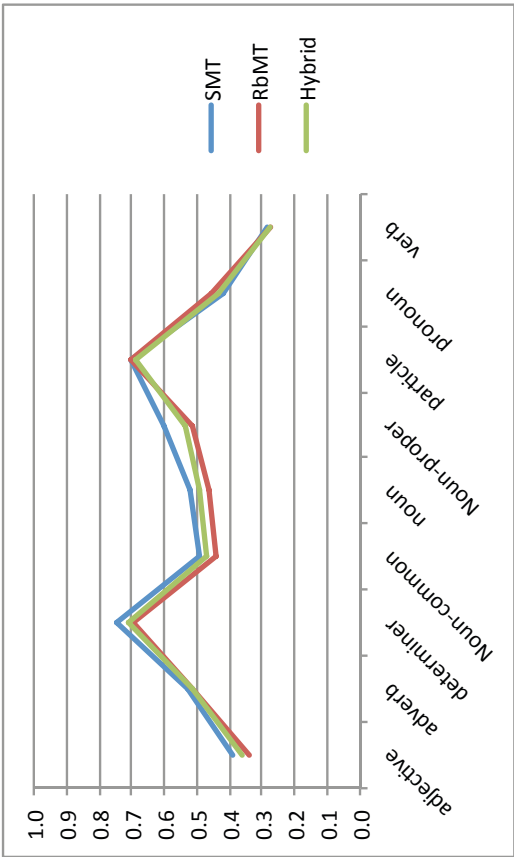


Figure 15. EN>ES results for rank 3

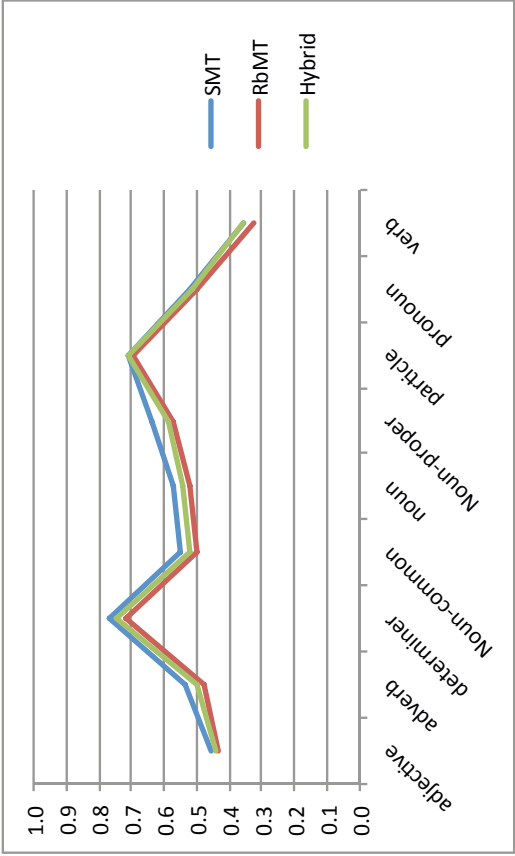


Figure 12. EN>ES results (overall)

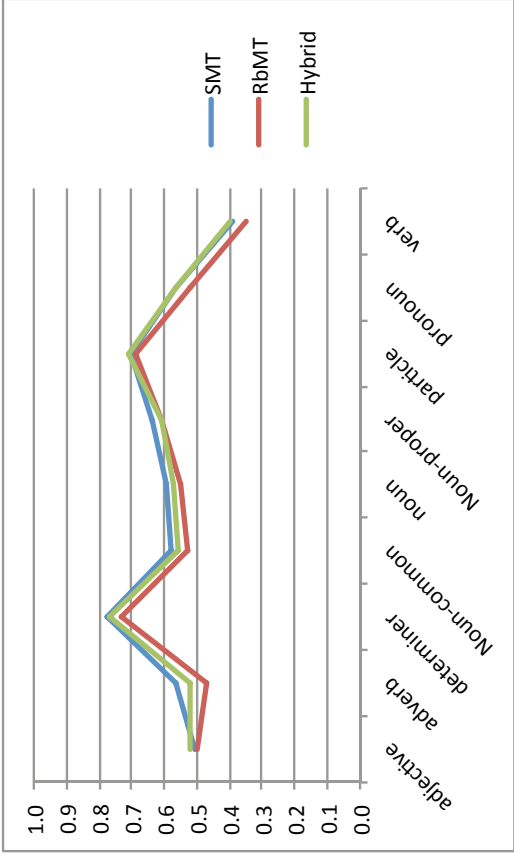


Figure 14. EN>ES results for rank 2

D1.3.1 Barriers for High-Quality Machine Translation

Rank 2 translations show similar results for all the three systems, with the RbMT system lagging slightly behind, especially for ADV, NOC and VER. Similar trends are observed for rank 3 translations, with the SMT system obtaining a wider advantage on DET, NOU and NOP. For these three checkpoints the hybrid system comes second and the RbMT system last. The results by the three systems are rather similar for the remaining checkpoints.

6.2.2. Diagnostic Evaluation Results for ES>EN

(See **Figure 16** through **Figure 19**.)

For the ES>EN translation direction, the performance of the SMT system is consistently better than that of the RbMT system for all the 9 linguistic checkpoints, with approximately a 10% difference in the respective DELiC4MT scores. Particularly severe quality barriers for the RbMT system seem to be adverbs, common nouns, pronouns and verbs. For rank 1 translations (bearing in mind the comparatively small numbers of checkpoint instances, with respect to the other two quality bands), the performance of the RbMT system is particularly modest for common nouns, verbs, particles and nouns. On the other hand, the SMT system and the RbMT system have very similar performances for adverbs and proper names in rank 1 translations, showing that these two areas are not specifically affected by quality differences of the two different MT systems.

The rank 2, i.e. medium-quality, translations show that the SMT system outperforms the RbMT system by a similar margin across all the linguistic checkpoints. As a result, in this case, the breakdown into the linguistic checkpoints does not allow us to gain particularly useful insights, simply showing a rather similar difference in quality. The situation is more interesting for the rank 3, i.e. poor-quality, translations, where both the SMT system and the RbMT system show specific weaknesses in the translation of adverbs, pronouns and verbs. Interestingly, although these three checkpoints show the lowest DELiC4MT scores, they are also the ones where the RbMT system performs better than the SMT system, by a significant margin. For the remaining six checkpoints in the rank 3 translations, the SMT system's output shows higher scores, with a difference of approximately 10% in value at times.

6.2.3. Diagnostic Evaluation Results for EN>DE

(See **Figure 20** through **Figure 23**.)

For the EN>DE translation direction in overall terms the performance of the three systems is very similar, showing that the SMT system gives slightly better scores than the RbMT system for all the 9 checkpoints, and that it also beats the hybrid system most of the time, except for particles (where there is a tie), pronouns and verbs. As a result, it is difficult to identify prominent barriers from this aggregate analysis, except for a comparatively poor performance of the RbMT system particularly for adjectives, common nouns, nouns and particles.

Looking at the results by ranking, on the other hand, gives a more interesting picture. For rank 1 translation, the SMT system shows a particularly disappointing performance for common nouns and nouns, while it is the top system for proper names and particles. The RbMT system receives the poorest score of the three systems for the adjective checkpoint, where the hybrid system also performs particularly badly. The rank 2 translations show a consistent trend, with the SMT system showing the best results for all the checkpoints (there is

D1.3.1 Barriers for High-Quality Machine Translation

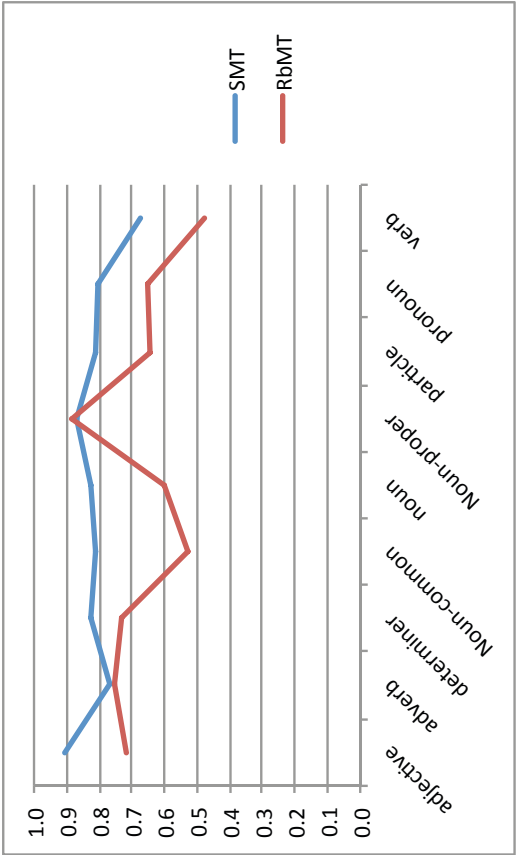


Figure 17. ES>EN results for rank 1

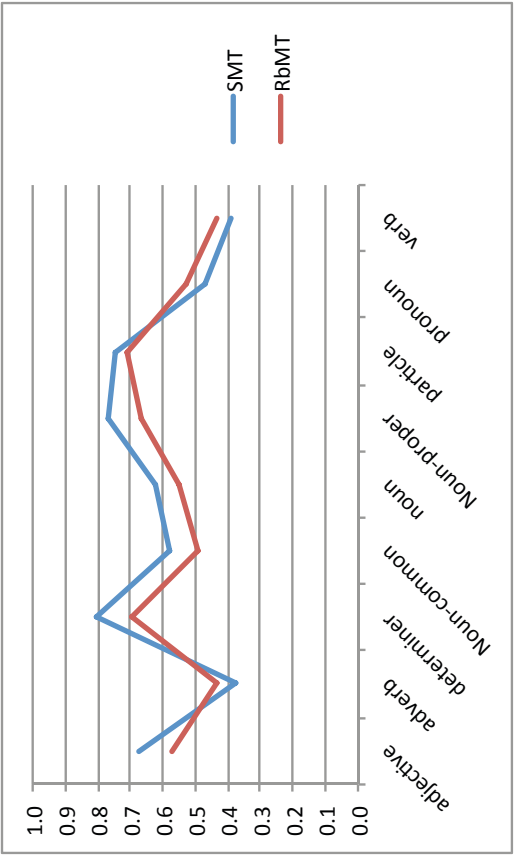


Figure 19. ES>EN results for rank 3

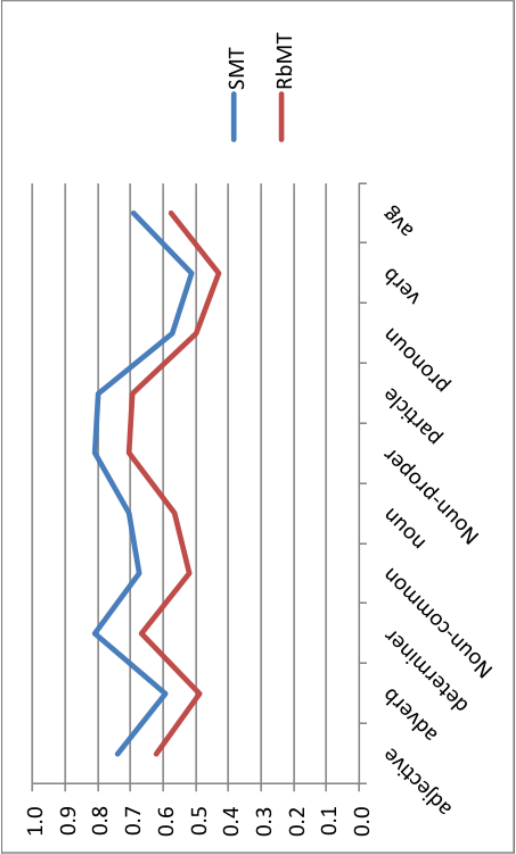


Figure 16. ES>EN results (overall)

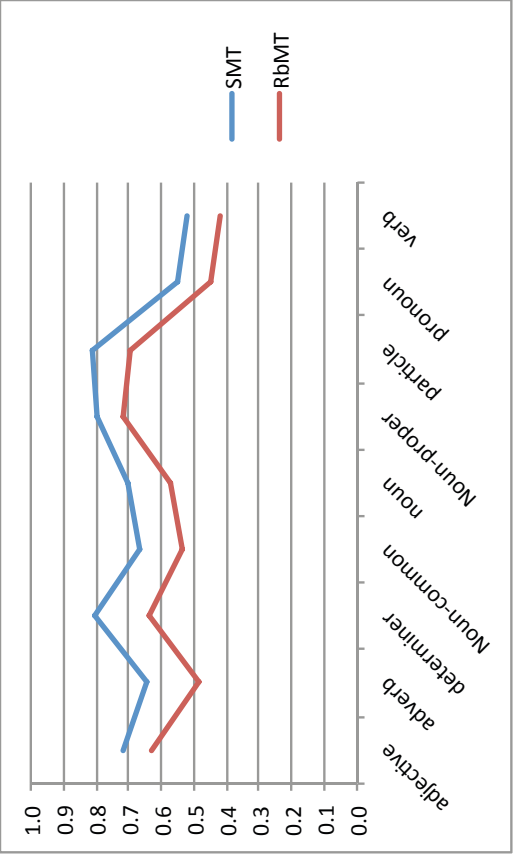


Figure 18. ES>EN results for rank 2

D1.3.1 Barriers for High-Quality Machine Translation

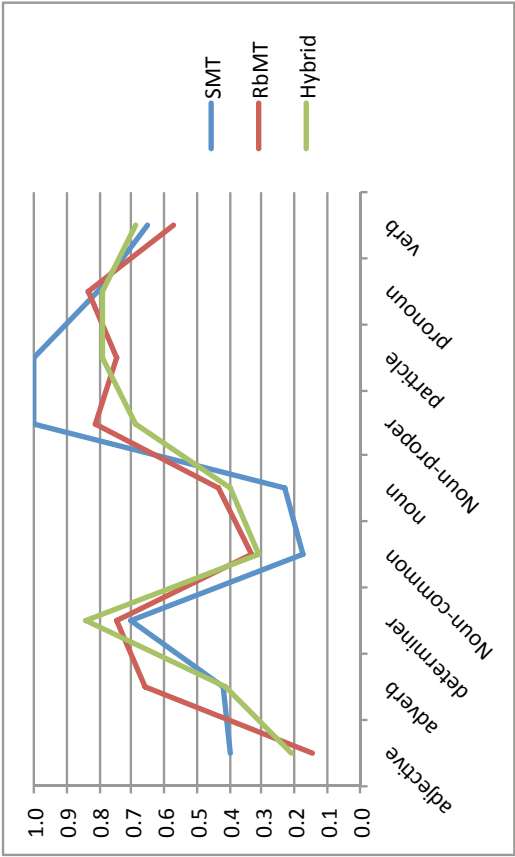


Figure 21. EN>DE results for rank 1

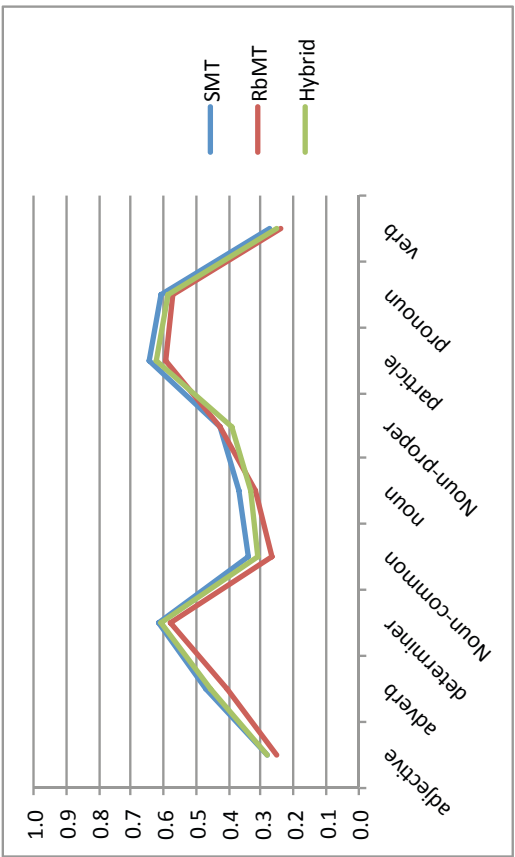


Figure 23. EN>DE results for rank 3

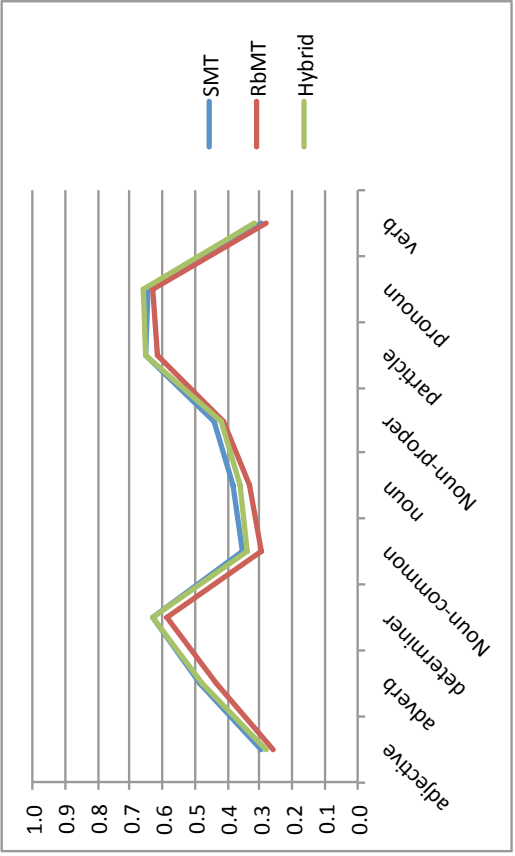


Figure 20. EN>DE results (overall)

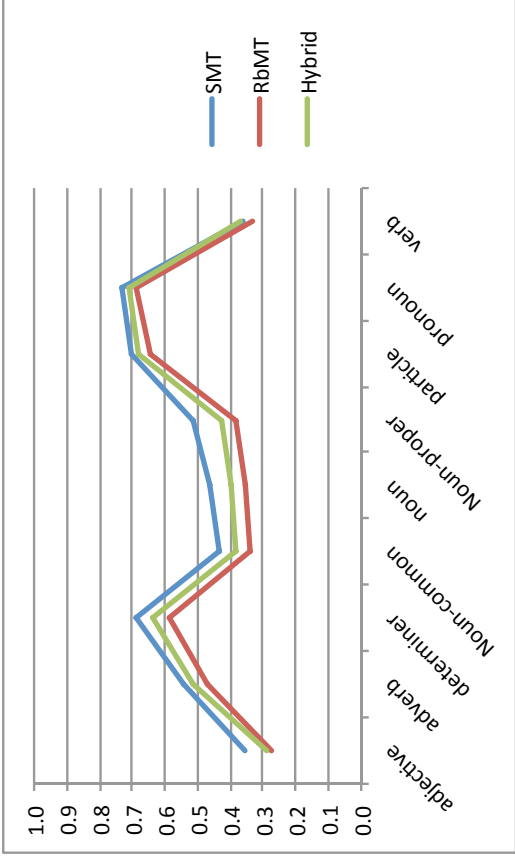


Figure 22. EN>DE results for rank 2

D1.3.1 Barriers for High-Quality Machine Translation

a tie with the hybrid system for verbs), and RbMT lagging behind. Finally, looking at rank 3 translations for EN>DE, all three MT systems find adjectives and verbs similarly problematic to translate, whereas the RbMT system runs into noticeably more difficulties with common nouns. For the remaining checkpoints the scores of the three MT systems do not show clear differences, hence we cannot identify other particularly severe or interesting translation quality barriers for translations of modest quality in the EN>DE translation direction.

6.2.4. Diagnostic Evaluation Results for DE>EN

(See **Figure 24** through **Figure 27**.)

Finally, coming to the DE>EN translation combination, the overall picture indicates that both the SMT system and the RbMT system encounter specific difficulties with the translation of adjectives, common nouns and nouns checkpoints, with similarly low performances (the scores of the RbMT system are slightly lower in all these three cases). On the other hand, DELiC4MT reveals that there are no particular problems for the translation of determiners, where both systems perform very well. The other checkpoints show very similar scores, but the RbMT system is comparatively weaker, especially for adverbs and pronouns.

With regard to the translation quality ranks, the RbMT system receives distinctly lower scores with DELiC4MT for adverbs, proper names, particles and pronouns for the rank 1 translations. On the other hand, the performance of the SMT system is particularly bad for adjectives (20% lower than the RbMT system), thus pointing to a clear quality barrier within the better translations. The RbMT system also gets a better evaluation than the SMT system for determiners, where the RbMT system translates correctly 97.6% of them. As far as rank 2 translations are concerned, the performance of the SMT system and the RbMT system is very similar across all the checkpoints: some are handled slightly better by the SMT system (e.g. adjectives, common nouns and pronouns), while in particular for proper names the DELiC4MT score of the RbMT system is higher. It is perhaps not particularly surprising that for rank 2, i.e. middle-quality, translations it is difficult to clearly differentiate the performance of the two systems in terms of quality barriers.

Finally, for rank 3 translation the performance tends to be equivalent again for most checkpoints, but the RbMT system struggles more with adverbs, common nouns and nouns in general. On the other hand, for low-quality translations the SMT system seems to find more serious barriers in the translation of verbs, for which the RbMT system receives a DELiC4MT score that is 5% higher.

6.2.5. Overall Diagnostic Evaluation Results and Analysis

Across all the translation directions and systems there tend to be comparatively few rank 1 translations (i.e. those rated as high-quality by the human judges). This considerably reduces the number of checkpoints detected in the input for that sub-set of the data, thus making it particularly difficult to draw reliable generalizations in such circumstances, due to data sparseness. To help put the DELiC4MT scores and our results in perspective, the number of instances of each checkpoint detected on the source/input side of the various data subsets appear in **Section 6.4**.

D1.3.1 Barriers for High-Quality Machine Translation

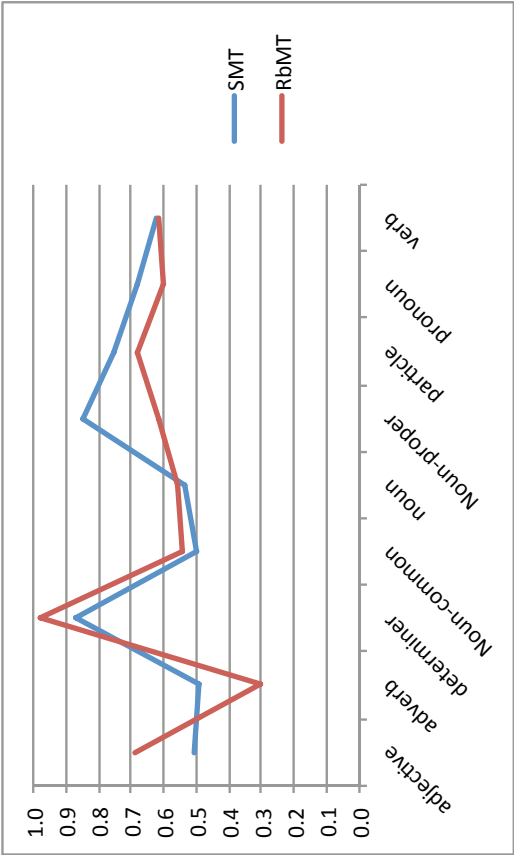


Figure 25. DE>EN results for rank 1

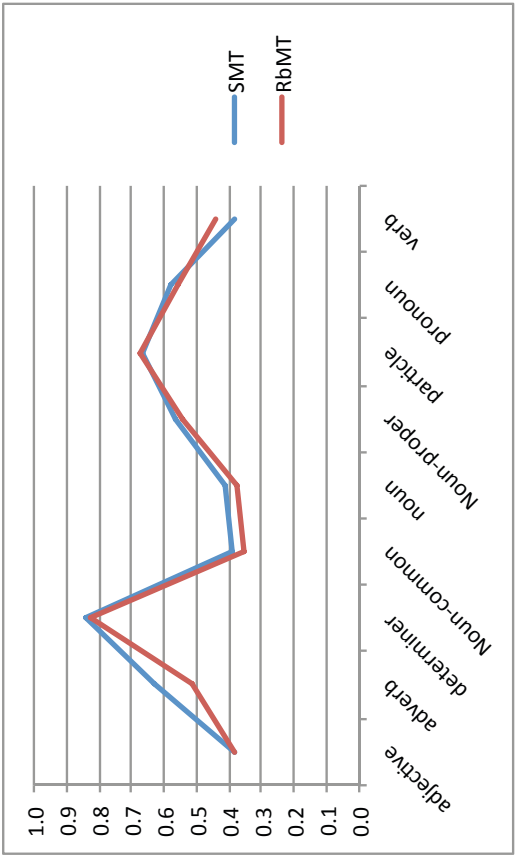


Figure 27. DE>EN results for rank 3

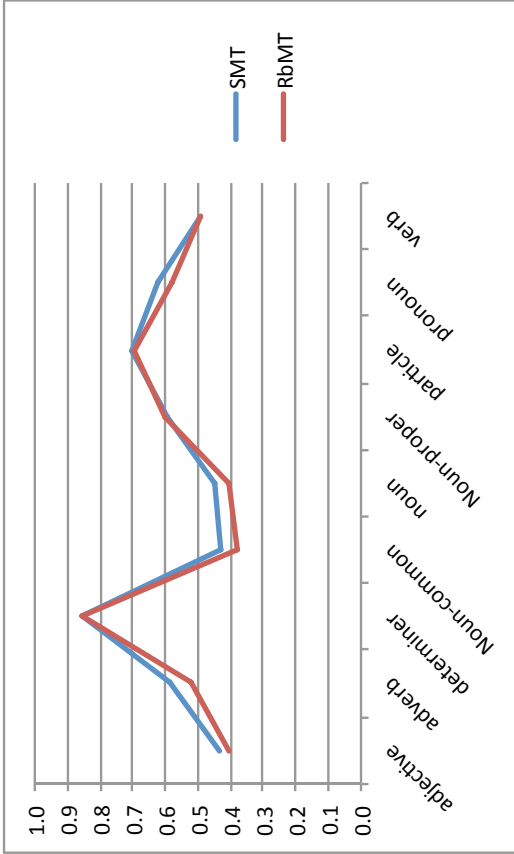


Figure 24. DE>EN results (overall)

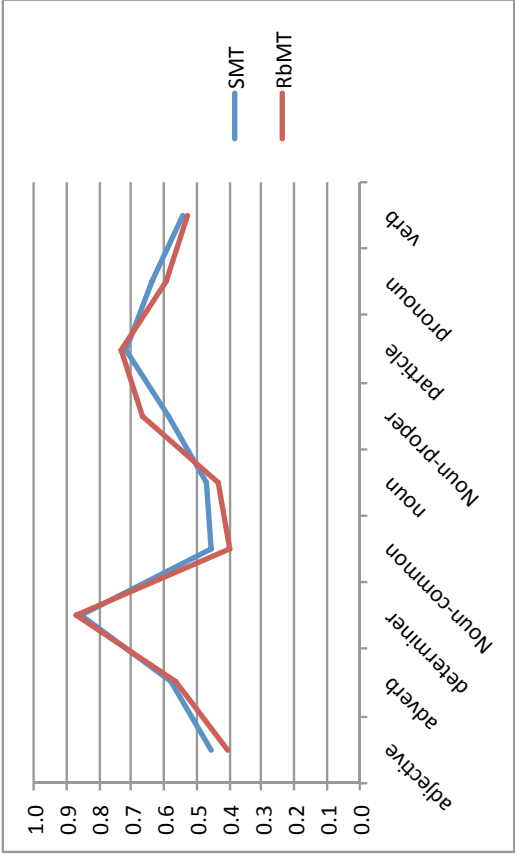


Figure 26. DE>EN results for rank 2

D1.3.1 Barriers for High-Quality Machine Translation

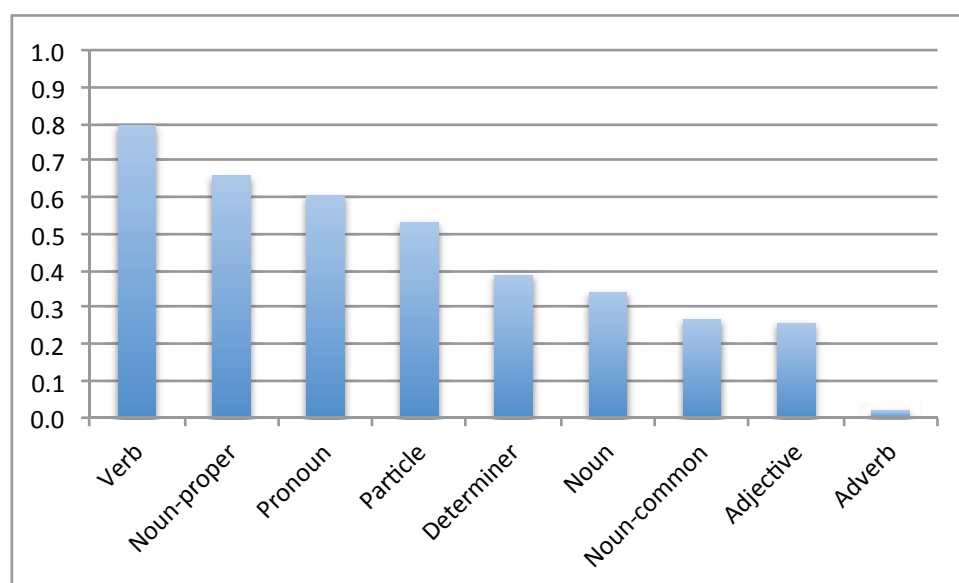


Figure 28. Pearson's r correlation for DELiC4MT scores and human quality ratings

To ascertain the correlation between the DELiC4MT scores and the human evaluations, we calculated correlation (Pearson's r) scores for the DELiC4MT scores and the human rating. Normalizing the results to positive values (since the rating is an inverse scale) gives the results shown in **Figure 28**.

As seen in this figure, quality levels correlated with all of the checkpoints, but many of the correlations are weak at best (and essentially nonexistent for *Adverb*). Assuming significance over the approximately 4400 translations examined, it seems the presence of verbal, proper noun, and pronominal checkpoints are the best indicators of human quality assessment.

Seemingly contradictory to the observations found in the MQM evaluation, however, *Particles* and *Determiners* also exhibited reasonable correlation (see **Section 6.4** below for more details), although these items proved particularly problematic for human annotators. Part of the reason for this, however, is that the values clustered tightly, indicating a high correlation, but the overall difference in these categories between quality bands was quite low, as can be seen in **Figure 29**. When broken out by individual category, aggregating all data, the distribution of scores and trend-lines for each category are as shown in **Figure 29**.

(Note that the trend lines in **Figure 29** slant down, which would indicate a negative r value, but reversing the values for the X axis so that 1 is the highest quality and 3 the lowest provides the r values cited above.)

6.3. Comparison of DELiC4MT Results with Other Findings

DELiC4MT takes a rather different approach from the qualitative, analytic MQM evaluation and from the Hjerson-based approach described above. DELiC4MT is particularly good at identifying specific classes of linguistic difficulties at a finer level than the other methods.

The hope with this comparison is that the individual checkpoints could be linked to features of the output of the MT systems, as presented in **Section 4**. However, this comparison runs

D1.3.1 Barriers for High-Quality Machine Translation

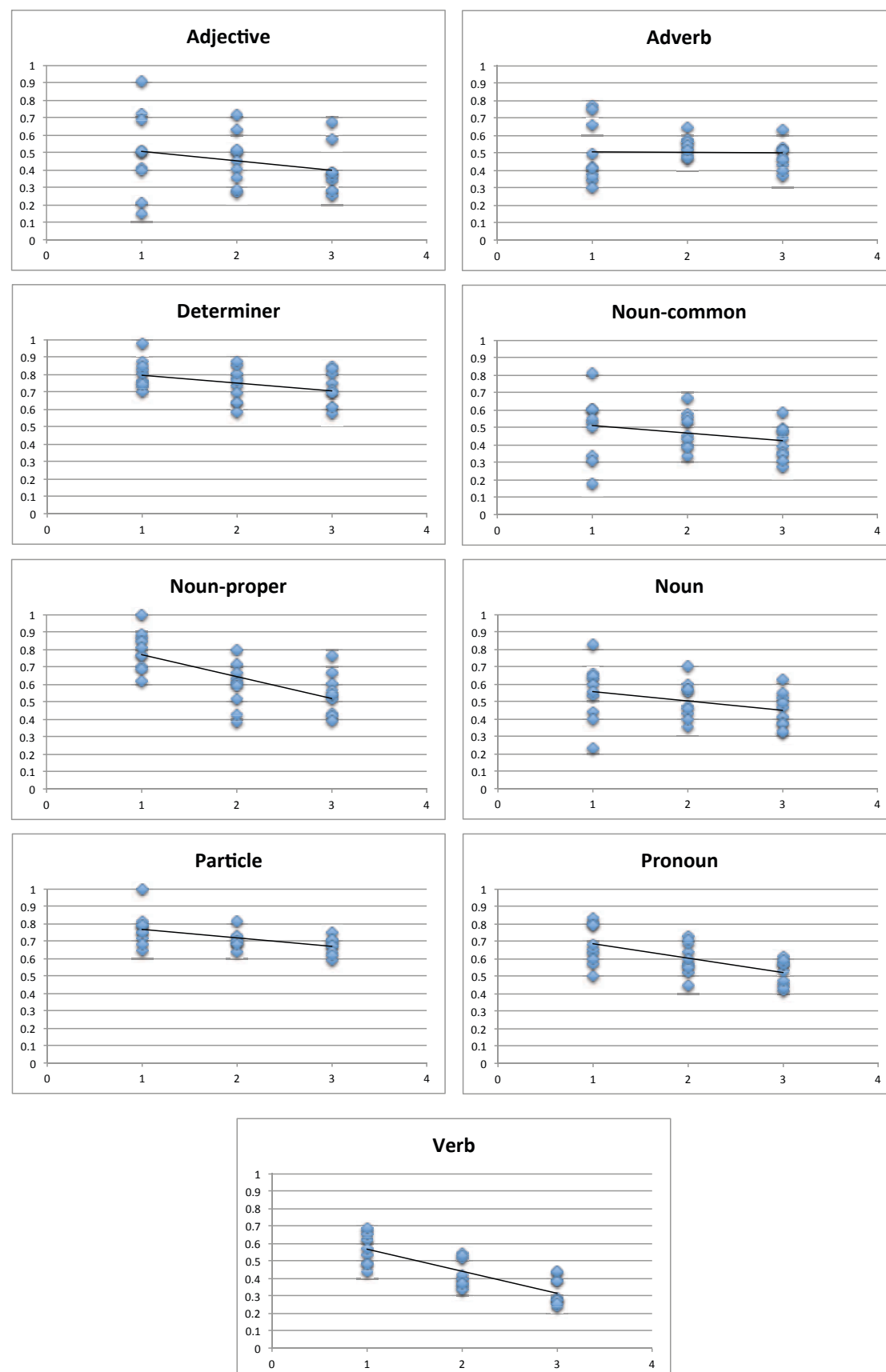


Figure 29. Distribution of DELiC4MT scores for each category by ranking class with trend lines.

D1.3.1 Barriers for High-Quality Machine Translation

into difficulty because the overall performances of the systems for “near miss” translations within each language pair, with the exception of ES>EN, were very close to each other for most checkpoints and showed similar distributions of scores between categories. Unexpectedly, the strongest differentiators between systems in DELiC4MT were for rank 1 (perfect) translations, which, by definition, showed no MQM errors. This result is almost certainly due to the low number of segments in this category, which created problems of data sparseness and “noise”; as a result the systems-specific scores for rank 1 should be interpreted with caution.

What the DELiC4MT analysis does show, however, is that the presence or absence of checkpoint correspondences by themselves does not correspond to human judgment concerning issues. In particular, some of the areas where the MT systems did well overall in the DELiC4MT evaluations, such as determiners and particles, were actually areas of weakness for MT overall from a human perspective. Determiners and particles would generally fall under **Function words** in MQM. For all quality bands and language pairs these checkpoints showed relatively high numbers. However, the presence of matching checkpoints in these categories may not be a good thing overall since this is an area where there is considerable variation between languages. Thus, finding a matching word in two languages may or may not be a good thing.

This finding indicates that high DELiC4MT results on lexical-level checkpoints may not be indicative of high-quality translation, but may instead point to overly literal renderings in some cases. By contrast, low scores may also not be a problem: in many cases the highest-rated translations received relatively low scores for noun-related checkpoints, even though one might expect that nominals should generally correspond. Some of the low scores may be related to word ordering issues or other factors that lead to low alignment success.

The next MQM evaluation round will expand on the analysis of **Function words** by allowing annotators to indicate whether a **Function words** issue is related to a missing word, an unneeded word, or an incorrect word. Based on the above observations, comparing the segment-level MQM results with DELiC4MT results should help clarify the relationship between the two approaches.

6.4. Tabular Data for Diagnostic Evaluation

This section contains the results for the diagnostic evaluation ([Section 6.2](#)) in tabular format. These tables complement the figures shown in [Section 6.2](#). The tables contain the scores for the different checkpoints as well as the number of instances per checkpoints. We present tables with results by rank and overall for all the four language directions.

D1.3.1 Barriers for High-Quality Machine Translation

6.4.1. EN>ES

		SMT		RBMT		HYBRID	
		score	instances	score	instances	score	instances
RANK1	adjective	0.5085	59	0.5000	44	0.4085	71
	adverb	0.3768	69	0.3455	55	0.3659	82
	determiner	0.7647	68	0.7568	37	0.8030	66
	Noun-common	0.6041	197	0.5979	97	0.6011	183
	noun	0.6612	304	0.6333	150	0.6473	258
	Noun-proper	0.7664	107	0.6981	53	0.7600	75
	particle	0.7818	110	0.7164	67	0.7900	100
	pronoun	0.5714	56	0.5000	48	0.6286	70
	verb	0.4870	193	0.4412	102	0.5354	198
	avg	0.6135		0.5766		0.6155	
RANK2	adjective	0.5081	492	0.5000	496	0.5179	448
	adverb	0.5621	443	0.4673	443	0.5228	394
	determiner	0.7768	569	0.7339	575	0.7647	527
	Noun-common	0.5798	1723	0.5299	1772	0.5559	1655
	noun	0.5969	2508	0.5530	2499	0.5711	2390
	Noun-proper	0.6344	785	0.6094	727	0.6054	735
	particle	0.7007	1156	0.6904	1153	0.7120	1073
	pronoun	0.5664	512	0.5216	510	0.5622	450
	verb	0.3892	1449	0.3492	1515	0.3984	1373
	avg	0.5905		0.5505		0.5789	
RANK3	adjective	0.3868	380	0.3422	412	0.3649	433
	adverb	0.5298	302	0.5168	327	0.5114	352
	determiner	0.7471	340	0.6917	373	0.7074	393
	Noun-common	0.4942	1202	0.4408	1275	0.4717	1308
	noun	0.5233	1649	0.4630	1840	0.4914	1852
	Noun-proper	0.6018	447	0.5133	565	0.5386	544
	particle	0.7037	729	0.7011	793	0.6888	845
	pronoun	0.4197	274	0.4533	289	0.4316	329
	verb	0.2844	1016	0.2722	1069	0.2757	1117
	avg	0.5212		0.4883		0.4979	
OVERALL	adjective	0.4586	931	0.4317	952	0.4401	952
	adverb	0.5344	814	0.4788	825	0.5024	828
	determiner	0.7656	977	0.7188	985	0.7444	986
	Noun-common	0.5484	3122	0.4959	3144	0.5235	3146
	noun	0.5741	4461	0.5188	4489	0.5427	4500
	Noun-proper	0.6341	1339	0.5725	1345	0.5871	1354
	particle	0.7063	1995	0.6955	2013	0.7061	2018
	pronoun	0.5190	842	0.4970	847	0.5171	849
	verb	0.3563	2658	0.3220	2686	0.3575	2688
	avg	0.5663		0.5257		0.5468	

Table 14. EN>ES results

D1.3.1 Barriers for High-Quality Machine Translation

6.4.2. ES>EN

		SMT		RBMT	
		score	instances	score	instances
RANK1	adjective	0.9063	53	0.7195	27
	adverb	0.7705	24	0.7503	17
	determiner	0.8243	119	0.7305	67
	Noun-common	0.8093	193	0.5304	100
	noun	0.8240	254	0.5989	124
	Noun-proper	0.8695	61	0.8861	24
	particle	0.8118	134	0.6455	76
	pronoun	0.8049	41	0.6516	29
	verb	0.6738	177	0.4769	157
	avg	0.8105		0.6655	
RANK2	adjective	0.7143	196	0.6333	196
	adverb	0.6471	119	0.4837	109
	determiner	0.8030	335	0.6389	327
	Noun-common	0.6682	639	0.5326	662
	noun	0.7034	853	0.5691	823
	Noun-proper	0.8001	214	0.7175	161
	particle	0.8091	482	0.6968	459
	pronoun	0.5520	125	0.4450	127
	verb	0.5178	786	0.4168	687
	avg	0.6906		0.5704	
RANK3	adjective	0.6714	70	0.5735	93
	adverb	0.3725	51	0.4335	68
	determiner	0.8024	132	0.6961	188
	Noun-common	0.5804	286	0.4900	354
	noun	0.6226	371	0.5478	528
	Noun-proper	0.7647	85	0.6664	174
	particle	0.7485	163	0.7082	240
	pronoun	0.4722	72	0.5269	81
	verb	0.3883	273	0.4339	388
	avg	0.6026		0.5640	
OVERALL	adjective	0.7398	319	0.6220	316
	adverb	0.5928	194	0.4899	194
	determiner	0.8089	586	0.6676	582
	Noun-common	0.6717	1118	0.5188	1116
	noun	0.7057	1478	0.5640	1475
	Noun-proper	0.8103	360	0.7051	359
	particle	0.7985	779	0.6950	775
	pronoun	0.5714	238	0.4976	237
	verb	0.5121	1236	0.4297	1232
	avg	0.6901		0.5766	

Table 15. ES>EN results

D1.3.1 Barriers for High-Quality Machine Translation

6.4.3. EN>DE

		SMT		RBMT		HYBRID	
		score	instances	score	instances	score	instances
RANK1	adjective	0.4000	5	0.1462	20	0.2105	38
	adverb	0.4167	12	0.6619	16	0.4103	38
	determiner	0.7000	10	0.7429	13	0.8433	29
	Noun-common	0.1750	40	0.3337	44	0.3071	140
	noun	0.2326	43	0.4362	56	0.4000	185
	Noun-proper	1.0000	3	0.8118	12	0.6889	45
	particle	1.0000	5	0.7486	16	0.7917	48
	pronoun	0.8059	16	0.8333	18	0.7893	30
	verb	0.6500	40	0.5702	42	0.6850	75
	avg	0.5183		0.5443		0.5307	
RANK2	adjective	0.3556	180	0.2740	360	0.2864	419
	adverb	0.5449	156	0.4710	252	0.5127	275
	determiner	0.6911	191	0.5832	373	0.6382	434
	Noun-common	0.4317	593	0.3364	1201	0.3829	1312
	noun	0.4597	905	0.3539	1878	0.3985	1995
	Noun-proper	0.5128	312	0.3846	677	0.4281	683
	particle	0.7034	290	0.6414	678	0.6815	743
	pronoun	0.7305	167	0.6887	317	0.7083	336
	verb	0.3631	493	0.3330	896	0.3679	1086
	avg	0.5046		0.4315		0.4684	
RANK3	adjective	0.2811	708	0.2493	512	0.2821	436
	adverb	0.4693	488	0.4048	388	0.4564	343
	determiner	0.6142	749	0.5819	560	0.6063	487
	Noun-common	0.3418	2551	0.2684	1930	0.3073	1732
	noun	0.3689	3771	0.3151	2766	0.3342	2539
	Noun-proper	0.4254	1220	0.4243	836	0.3879	807
	particle	0.6411	1435	0.5902	1034	0.6224	939
	pronoun	0.6089	509	0.5721	355	0.5854	326
	verb	0.2731	1801	0.2400	1396	0.2493	1173
	avg	0.4229		0.3752		0.3990	
OVERALL	adjective	0.2968	893	0.2576	892	0.2811	893
	adverb	0.4863	656	0.4374	656	0.4782	656
	determiner	0.6305	950	0.5854	946	0.6284	950
	Noun-common	0.3565	3184	0.2954	3175	0.3388	3184
	noun	0.3850	4719	0.3327	4700	0.3645	4719
	Noun-proper	0.4443	1535	0.4107	1525	0.4169	1535
	particle	0.6526	1730	0.6129	1728	0.6529	1730
	pronoun	0.6431	692	0.6307	690	0.6557	692
	verb	0.2986	2334	0.2816	2334	0.3184	2334
	avg	0.4402		0.4006		0.4350	

Table 16. EN>DE results (by rank)

D1.3.1 Barriers for High-Quality Machine Translation

6.4.4. DE>EN

		SMT		RBMT	
		score	instances	score	instances
RANK1	adjective	0.5087	173	0.6889	45
	adverb	0.4921	63	0.3000	20
	determiner	0.8725	102	0.9762	42
	Noun-common	0.5000	356	0.5461	152
	noun	0.5354	396	0.5562	178
	Noun-proper	0.8462	39	0.6154	26
	particle	0.7500	152	0.6809	47
	pronoun	0.6782	87	0.6000	35
	verb	0.6257	179	0.6180	89
	avg	0.6120		0.6250	
RANK2	adjective	0.4552	591	0.4020	587
	adverb	0.5764	203	0.5641	195
	determiner	0.8559	479	0.8721	438
	Noun-common	0.4543	2023	0.3988	1655
	noun	0.4704	2294	0.4352	1921
	Noun-proper	0.5889	270	0.6629	264
	particle	0.7192	673	0.7281	581
	pronoun	0.6376	298	0.5971	278
	verb	0.5441	680	0.5297	640
	avg	0.5594		0.5371	
RANK3	adjective	0.3843	536	0.3834	673
	adverb	0.6305	203	0.5118	254
	determiner	0.8437	403	0.8294	504
	Noun-common	0.3874	1737	0.3541	2299
	noun	0.4100	1995	0.3745	2574
	Noun-proper	0.5620	258	0.5455	275
	particle	0.6661	581	0.6748	778
	pronoun	0.5777	251	0.5604	323
	verb	0.3806	578	0.4393	708
	avg	0.4935		0.4794	
OVERALL	adjective	0.4331	1300	0.4023	1305
	adverb	0.5885	469	0.5245	469
	determiner	0.8526	984	0.8547	984
	Noun-common	0.4300	4116	0.3792	4106
	noun	0.4502	4685	0.4064	4673
	Noun-proper	0.5944	567	0.6035	565
	particle	0.7006	1406	0.6970	1406
	pronoun	0.6195	636	0.5786	636
	verb	0.4885	1437	0.4906	1437
	avg	0.5375		0.5100	

Table 17. DE>EN results

D1.3.1 Barriers for High-Quality Machine Translation

7. Lessons Learned in the Annotation Task

This section addresses some of the lessons learned from an examination of the MQM annotation described in [Section 4](#), with a special emphasis on ways to improve inter-annotator agreement (IAA). Although IAA does not appear to be a barrier to the present analytic task, the QTLaunchPad project has undertaken the following actions to improve the consistency of forthcoming annotations:

- We restructured certain aspects of the metric used. These changes will allow annotators to focus on distinctions that are important and are clear. The specific changes made are the following:
 - The distinction between **Mistranslation** and **Terminology** was eliminated, even though these were the two most frequently used categories. For future analysis we will automatically distinguish between word- and phrase-level mistranslations. (Although **Terminology** will remain in MQM, it will not be used in the Shared Task Metric; in MQM itself it will be moved to become a daughter of **Mistranslation**)
 - The **Style/Register** category was eliminated since stylistic and register expectations were unclear.
 - **Punctuation** and **Typography** were unified under the parent node, **Typography**. Although **Typography** appeared infrequently in the existing corpus, all instances of **Punctuation** are also instances of **Typography**.
 - **Capitalization**, which was infrequently encountered was merged into its parent, **Misspelling**.
 - The **Morphology (word form)** category was renamed **Word form** and **Part of Speech, Agreement**, and **Tense/mood/aspect** were moved to become its children. (This change will be made to the full MQM structure as well.)
 - Three custom subtypes were added to **Function words**:
 - **Extraneous**, for cases where unneeded function words appear
 - **Missing**, for cases where a required function word does not appear
 - **Incorrect**, for cases where the wrong function word is used
 This change was made because annotators reported confusion on whether to categorize specific instances as **Addition** or **Omission** (thus crossing branches into **Accuracy**). Adding these categories allows annotators to be more specific about these frequent errors and help prevent confusion between **Accuracy** and **Fluency** issues.
- We created a decision tree for training purposes. This decision tree should assist annotators to be more consistent. Previously they had a list of issue types and definitions with examples, but no tools to guide them in the decision-making process.
- We updated the annotation guidelines to reflect the new category structure and decision tree.
- We are planning to prepare another introduction video to show annotators the new categories and explain them.

D1.3.1 Barriers for High-Quality Machine Translation

8. References

- Balyan, Renu, Sudip Kumar Naskar, Antonio Toral and Niladri Chatterjee (2012) "A Diagnostic Evaluation Approach Targeting MT Systems for Indian Languages". *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), COLING 2012, Mumbai, India, December 2012*, pp. 61–72.
- Balyan, Renu, Sudip Kumar Naskar, Antonio Toral and Niladri Chatterjee (2013) "A Diagnostic Evaluation Approach for English to Hindi MT Using Linguistic Checkpoints and Error Rates". In A. Gelbukh (ed) *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013)*, pp. 285–96. LNCS 7817. Berlin: Springer.
- Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut and Lucia Specia. "Findings of the 2012 Workshop on Statistical Machine Translation". In *Proceedings of the Seventh Workshop on Statistical Machine Translation (WMT 12)*, pp. 10–51.
- Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar Zaidan, 2011. "Findings of the 2011 Workshop on Statistical Machine Translation". In *Proceedings of Workshop on Statistical Machine Translation (WMT11)*, pp 22–64.
- Naskar, Sudip Kumar, Antonio Toral, Federico Gaspari and Andy Way (2011) "A Framework for Diagnostic Evaluation of MT Based on Linguistic Checkpoints". *Proceedings of Machine Translation Summit XIII, Xiamen, China, 19-23 September 2011*, pp. 529–36.
- Naskar, Sudip Kumar, Antonio Toral, Federico Gaspari and Declan Groves (2013) "Meta-Evaluation of a Diagnostic Quality Metric for Machine Translation". Khalil Sima'an, Mikel L. Forcada, Daniel Grasmick, Heidi Depraetere and Andy Way (eds) *Proceedings of the XIV Machine Translation Summit. Nice, France, 2-6 September 2013*, pp. 135–42. Allschwil: The European Association for Machine Translation.
- Popović, Maja. 2011. "Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output". *Prague Bulletin of Mathematical Linguistics* 96:59–68.
- Specia, Lucia. 2011. "Exploiting Objective Annotations for Measuring Translation Post-editing Effort". In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT 11), Leuven, Belgium, May 2011*, pp. 73–80.
- Toral, Antonio, Sudip Kumar Naskar, Federico Gaspari and Declan Groves (2012) "DELic4MT: A Tool for Diagnostic MT Evaluation over User-defined Linguistic Phenomena". *The Prague Bulletin of Mathematical Linguistics* 98(1):121–31. DOI: 10.2478/v10108-012-0014-9.
- Toral, Antonio, Sudip Kumar Naskar, Joris Vreeke, Federico Gaspari and Declan Groves (2013) "A Web Application for the Diagnostic Evaluation of Machine Translation over Specific Linguistic Phenomena". Chris Dyer and Derrick Higgins (eds) *Proceedings of the 2013 NAACL HLT Conference - Demonstration Session. Atlanta, GA, USA. 10-12 June 2013*, pp. 20–23 Stroudsburg, PA: Association for Computational Linguistics.

D1.3.1 Barriers for High-Quality Machine Translation

9. Appendix: Description of DELiC4MT Functionality

For illustration purposes and to explain how the analysis performed by DELiC4MT works, here we show three examples on the segment level (all of them for the language direction Spanish to English and for the checkpoint verb). Given a checkpoint instance, these examples show the reference (source and target sides), the alignments (words between “<” and “>”), the MT output and the *n*-gram matches.

Source ref: *Y aún así, <es> una estrella .*
 Target ref: *And yet, he <is> a star .*
 MT output: *And still like this, <is> a star .*
 ngram matches: *is (1/1)*

The first example concerns a correct translation, scored successfully by DELiC4MT. The form *es* (3rd person present of *ser* ‘to be’) is correctly translated to its equivalent in English, *is*.

Source ref: *“Fue un regalo que me <hizo> él ”*
 Target ref: *“It was a gift he <gave> me ”*
 MT output: *“It was a gift that did me he ”*
 ngram matches: *– (0/1)*

The second example regards a verb literally (and incorrectly) translated. The source *hizo* (3rd person past of *hacer* ‘to make’) would normally correspond to *gave* in English. However, in the expression *hacer un regalo* it corresponds to *make a present*. The evaluation tool correctly identifies this as a mistake.

Source ref: *Anto tiene asma, <respira> con dificultad [...]*
 Target ref: *Anto has asthma, <he> <has> difficulty breathing [...]*
 MT output: *Anto <has> asthma, it breathes with difficulty [...]*
 ngram matches: *has (1/3)*

Finally, the third example concerns a correct translation, which the evaluation tool fails to assess positively. The verb *respira* (3rd person present of *respirar* ‘to breathe’) is correctly translated as *breathing*. However, due to a wrong word alignment (*respira* is wrongly aligned to *he has* instead of to *breathing*), the score is not 1/1, but 1/3.

9.1. General Presentation of DELiC4MT

DELiC4MT [www.computing.dcu.ie/~atoral/delic4mt/] (Toral et al., 2012) is an open-source toolkit for diagnostic machine translation (MT) evaluation. Its diagnostic dimension derives from its ability to focus on user-defined linguistic checkpoints, i.e. phenomena of the source language that the user decides to focus on for the evaluation of the quality of the MT output. Linguistic checkpoints can correspond to interesting or difficult lexical items and/or grammatical constructions for which a specific translation quality assessment is required. They can be defined at any level of specificity desired by the user, considering lexical, morphological, syntactic and/or semantic information related to the source language.

Linguistic checkpoints can consist, for example, of: individual grammatical or content words (literal tokens or lemmas), or a (discontinuous) series thereof; one or more words, identified on the basis of their part of speech (PoS); grammatical features (e.g. plural nouns, verbs in

D1.3.1 Barriers for High-Quality Machine Translation

the simple past tense); named entities. Any of these levels of description of source-language phenomena can be combined to create linguistic checkpoints of variable composition, ranging from very basic and generic (e.g. focusing on any noun found in the input) to very complex and specific (e.g. all the word sequences in the source text composed by: a determiner, followed by any singular noun, followed by the literal word “of”, followed by any plural noun, followed by a finite form of the verb ‘go’, etc.). The only constraint on the design of linguistic checkpoints for DELiC4MT is that they should consist of features supported by the language resources and processing tools previously used to annotate the data sets (most notably the PoS tagger); clearly, some languages are better served than others in this respect.

In addition to being able to support user-defined linguistic checkpoints of varying complexity and potentially combining several levels of description of source-language phenomena, DELiC4MT is language-independent and can be easily adapted to any language pair – it has, for example, been successfully applied to the diagnostic evaluation of MT quality for European language pairs (e.g. Naskar et al., 2011; Naskar et al., 2013), as well as for English in combination with Indian languages (Balyan et al., 2012; Balyan et al., 2013) on a range of checkpoints specific to the respective source languages. DELiC4MT was released in 2012 as an open-source toolkit and was built adopting best practices, combining cutting-edge software components and integrating well-established formats developed in a variety of recent high-profile projects and initiatives. Toral et al. (2012) describe the different modules that make up the toolkit and present a step-by-step case study of how it can be applied to a specific language pair for an illustrative linguistic checkpoint defined by the user. A step-by-step tutorial is also available, showing how the toolkit works, applying it to a specific language pair, test set and linguistic checkpoint (https://github.com/antot/DELiC4MT/blob/master/doc/tutorial/delic4mt_tutorial.pdf). DELiC4MT is also available via a web application and a web service, which are more convenient for users who wish to avoid the burden of installing, configuring and maintaining the software (Toral et al., 2013).

In order to operate, DELiC4MT requires three pre-processed data sets: the source-language text (i.e. the input in language A), the MT output (in the target language B) and a reference translation (again in language B). The input needs to be PoS-tagged and word-aligned to the reference translation. The linguistic checkpoint (defined by the user in the source language A) is detected on the source side, and linked to its corresponding word(s) in the reference translation in the target language B. At this stage, DELiC4MT checks if the MT output in the target language B matches the aligned translation in the reference for the given source-language checkpoint. Once this analysis has been completed for the whole data set, the toolkit produces a score, indicating how many of the relevant checkpoints detected on the source side were translated correctly by the MT system; the user can then also analyze the occurrences of the checkpoints individually in detail, if required, comparing the source sentence, its corresponding MT output and the reference translation, where the linguistic checkpoints and their translations are highlighted. Given its design, DELiC4MT allows users to investigate several linguistic checkpoints on the same data sets, and also to diagnostically evaluate the same linguistic checkpoint(s) across different MT systems or for different data sets in the same language pair, in a comparative perspective.

9.2. Novelty of the Analysis with DELiC4MT

DELiC4MT was developed to respond to the widespread needs of MT users requiring a strong diagnostic dimension in MT quality evaluation. While being relatively simple to use and requiring a limited amount of time and resources to run, the toolkit supports more

D1.3.1 Barriers for High-Quality Machine Translation

flexible, transparent and delicate evaluation than the standard automatic MT evaluation metrics, whose scores are difficult to interpret and do not help one to understand the actual strengths and weaknesses of an MT system. DELiC4MT, on the other hand, can provide information at different levels of detail on the linguistic phenomena of specific interest to the users, showing which ones are handled more or less successfully by the MT system. This diagnostic feedback can then be incorporated into the further development, fine-tuning and customization of the MT software to optimize its performance. In fact, nothing prevents users from using DELiC4MT in combination with state-of-the-art MT evaluation metrics, to obtain a more complete picture of the quality achieved by a certain MT system and/or on a specific data set for a given language pair.

D1.3.1 Barriers for High-Quality Machine Translation

10. Appendix: Annotation Guidelines

This section includes the original Annotation Guidelines distributed to annotators and the revised Guidelines (as of 2014 February 4). Note that the Guidelines are still under discussion and are likely to be revised further.

10.1. Original MQM Issue Annotation Guidelines

Prepared by Kim Harris (text&form) and Arle Lommel (DFKI).

Version 1.0 (2013-12-17). Adapted from QTLaunchPad Deliverable 1.2.1

10.1.1. Objective

The purpose of the error annotation exercise is to accurately categorize errors found in “near miss” translations output by various machine translation systems and system types. It is of the utmost importance that these errors be systematically and correctly annotated without prejudice or preference, as the annotation will be performed by various people and, ultimately, for various language pairs, and must therefore remain as objective as possible to ensure consistent and reliable results.

10.1.2. Approach

You will be given unedited bilingual machine translation output for **review** and **error annotation**. The first step, **review**, will determine whether or not the translation qualifies as a “near miss”.⁶ If it does, the second step, **error annotation**, must be performed on the respective translation. If, however, the translation is either perfect or contains too many errors, annotation will not be performed.

10.1.3. Procedure

10.1.3.1. What is an error?

An error represents any issue you may find with the translated text that either does not correspond to the source or is considered incorrect in the target language. The list of language issues upon which you are to base your annotation is described in detail below and provides a range of examples.

The list is divided into two main issue categories, **Accuracy** and **Fluency**, each of which contains relevant, more detailed subcategories. Whenever possible, the correct subcategory should be chosen; however, if in doubt, please **do not guess**. In this case, select the category level you are most certain in order to avoid inconsistencies in the results.

Example

The German term *Zoomfaktor* was incorrectly translated as *zoom shot factor* by one of the systems, and you are unsure whether this represents a terminology error or an addition.

⁶ Does not apply to the calibration task since the sentences have been preselected.

D1.3.1 Barriers for High-Quality Machine Translation

In this case, categorize the error as an **Accuracy** error.

10.1.3.2. When one error is “right” if the other is fixed

At first glance it may appear as if more than one error has been made, when in fact this may not be the case. An agreement error, for example, will only have one issue, not two, so it is important to understand what the actual underlying error is before rejecting a potential “near miss” translation that could be used for annotation.

10.1.3.3. Examples

Source: *Importfilter werden geladen*
 Translation: *Import filter are being loaded*
 Correct: *Import filters are being loaded*

In this example, the only error is the translation of *filter* in the singular rather than the plural. Had the subject been translated properly, agreement would have been correct. Agreement is therefore **not** an error in this case. Only *Import filter* would be tagged in this sentence.

Source: *im Dialog Exportieren*
 Translation: *in the dialog export*
 Correct: *in the Export dialog*

In this example, only terminology is incorrect, as a software term would be considered a terminology issue. While word order and capitalization would be considered errors in other contexts, this would not be the case here, as these two words constitute one term.

10.1.3.4. The Review Process

Review precedes the annotation phase as a pre-processing step to filter the “near miss” translations that you will subsequently annotate and to categorize both perfect and very bad translations for integrity purposes. Sentences in the latter two categories will not be annotated any further.

1. Make sure you understand how to establish what should be considered an error before beginning this task.
2. Read the source and target language sentences carefully.
3. Determine whether or not the target sentence contains any translation errors. The visual analysis required for this task would be similar to that required to actually edit the translation.
 - a. If the translation contains no errors, select **Perfect**.
 - b. If the translations contains one to three errors, select **Near Miss**.
 - c. If the translation contains more than three errors, select **Reject**.
4. You do not need to tag or highlight any of the issues at this stage.

D1.3.1 Barriers for High-Quality Machine Translation

10.1.3.5. The Annotation Process

Once you have completed the review process, you should have a set of “near miss” translation to annotate. Please follow these rules when selecting errors and tagging the respective text in the translations:

1. If you do not understand what types of errors belong to a particular category, please see the documentation, which contains examples.
2. In many cases multiple issue types may describe a *single* actual issue/error in the text. For example, if a term (domain-specific word with specialized meaning) is translated incorrectly, this could be considered an example of both the *Terminology* and *Mistranslation* issue types. In such cases issues should not be counted multiple times but instead *one* issue type should be selected based on the following principles:
 - a. If it is possible to assign an issue to a more specific type (e.g., *Part of speech*) rather than a more general one (e.g., *Grammar*), the more specific one should be used. (But see item #6 below: do not use a category that does not clearly apply. If no category at a certain level is precise, use a more general one.)
 - b. General types should be used for cases in which the issue is of a more general nature or where the specific problem does not have a precise type in the hierarchy. For example *He slept the baby* uses a direct object with an intransitive verb (a problem technically known as a *valency error*), but as there is no specific type for this error, it should be assigned to *Grammar*.
 - c. If multiple issue types apply at the same level in the hierarchy, the one that appears first in the list should be used. For example, if both *Terminology* and *Mistranslation* apply, *Terminology* will be used because it appears first in the hierarchical listing.
3. **Less is more.** Only tag the relevant text. Do not tag the entire sentence, as this will cause errors to occur in the analysis of the actual error itself and potentially render the results unacceptable. For example, if a single word is wrong in a phrase, tag only the single word rather than the entire phrase. (See the section on “minimal markup” below.)
4. Always look for the predominant issue. If that issue is fixed, do the others still exist?
5. One word in a sentence may contain two errors. If this is the case, enter both errors separately and mark the respective word in both cases.
6. If in doubt, choose a more general category.

10.1.3.6. Tricky cases

- **Hyphenation:** Hyphenation issues are spelling issues or represent part of an untranslated term.

Example: XML-files represents a spelling error Pocket-PC represents an untranslated term

- **Number** (plural vs. singular) is a mistranslation
- **Mistranslation vs. terminology:** A mistranslation is the random selection of a translation for a word, whereas a terminology error is the selection of the wrong term, even if it may still convey the meaning.

D1.3.1 Barriers for High-Quality Machine Translation

Example: *Zoom shot factor* for *Zoomfaktor* (should be *zoom factor*) is a Terminology error. Terminology errors also include software terms, named entities and other compound words.

Example: An English translation uses the term *thumb drive* to translate the German *USB Speicherkarte*. This translation is intelligible, but if the translation mandated in specifications or a relevant termbase is *USB memory stick*, the use of *thumb drive* constitutes a Terminology error.

- **Unintelligible:** If you cannot understand the sentence because a verb is missing or a word has been incorrectly translated, or perhaps both of these together, the sentence would not be considered unintelligible. A sentence that can easily be fixed should be categorized accordingly. An Unintelligible sentence is generally one whose word order issues and omissions are relatively complicated to remedy and where the meaning is unclear.

Example: In the sentence “You can also you can use this tab to precision, with the colors are described as well as the PostScript Level.” there are enough errors that the meaning is unclear and the precise nature of the errors that lead to its unintelligibility cannot be easily determined.

- **Agreement:** This category generally refers to agreement between subject and predicate or gender and case.

Example: A boy plays with *his* train (not *her* train)

Example: I *am* at work (not I *is* at work)

- **Untranslated:** Many words may look as if they have been translated and simply forgotten to apply proper capitalization or hyphenations rules. In most, cases, this would represent an untranslated term and not a capitalization or spelling issue. If the target word or phrase is identical to the source word or phrase, it should be treated as Untranslated, even if a spelling or capitalization error could also account for the problem.

10.1.3.7. Minimal markup

It is vital in creating error markup that errors be marked up with the shortest possible spans. Markup should identify only that area needed to identify the problem. In some cases, spans may be discontinuous. Examples include:

Sentence	Incorrect markup	Problem	Correct minimal markup
Double click on the number faded in the status bar.	Double click on the number faded in the status bar. [Marked for Mistranslation]	Only the single word <i>faded</i> is problematic, but the markup indicates that <i>number faded in</i> is incorrect.	Double click on the number faded in the status bar.

D1.3.1 Barriers for High-Quality Machine Translation

Sentence	Incorrect markup	Problem	Correct minimal markup
Source: Die Standardschriftgröße für Dialoge ist dabei 12pt, die einem Maßstab von 100 % entspricht. Translation: The standard font size for dialogs is 12pt, which corresponds to a standard of 100%.	The standard font size for dialogs is 12pt, which corresponds to a standard of 100% . [Marked for Terminology]	Only the term <i>Maßstab</i> has been translated incorrectly. The larger span indicates that text that is perfectly fine has a problem.	The standard font size for dialogs is 12pt, which corresponds to a standard of 100%.
The man who they saw on Friday night at the store were very big.	The man who they saw on Friday night at the store were very big. [Marked for Agreement]	Here the highlighted portion identifies only a single word, insufficient to identify the agreement problem. The correct version highlights the two words that do not agree with each other.	The man who they saw on Friday night at the store were very big.
Buttons must not be pressed until the system is ready.	Buttons must not be pressed until the system is ready. [Marked for Style]	Here the entire sentence is marked, making it impossible to tell which portion was problematic. Since on the first (passive) clause was problematic, only it should be marked.	Buttons must not be pressed until the system is ready.
The in 1938 nascent leader with flair divined %temp_name eating lonely.	The in 1938 nascent leader with flair divined %temp_name eating lonely. [Marked for Unintelligible]	The entire sentence is Unintelligible and should be marked as such.	The in 1938 nascent leader with flair divined %temp_name eating lonely.

In the event of questions about the scope of markup that should be used, utilize the note field to make a query or explain your choice.

10.1.4. Appendix B: Issue categories

The error corpus uses the following issue categories:

- **Accuracy.** Accuracy addresses the extent to which the target text accurately renders the meaning of the source text. For example, if a translated text tells the user to push a button when the source tells the user *not* to push it, there is an accuracy issue.
- **Terminology.** A term is translated with a term other than the one expected for the domain or otherwise specified.

Example: The English musicological term *dog* is translated (literally) into German as *Hund* instead of as *Schnarre*, as specified in a terminology database.

Note(s): (1) If a term is not translated, it should be listed as *Untranslated* rather than as *Terminology*. (2) *Terminology* applies only to mistranslation of words or phrases that are specified in a terminology database or that are specific to a given topic field. For example, if a musicological term base stated that English *dog* is to be translated as German *Schnarre*, but it is instead rendered as *Hund*, the issue is a terminology error

D1.3.1 Barriers for High-Quality Machine Translation

since the general language equivalent is *Hund*. If a word is translated incorrectly independent of the subject field, it should be classified as *Mistranslation*.

- **Mistranslation.** *The target content does not accurately represent the source content.*

Example: A source text states that a medicine should not be administered in doses greater than 200 mg, but the translation states that it should not be administered in doses *less than* 200 mg.

Note(s): *Mistranslation* can be used for single words if the target text word does not represent a contextually plausible translation of the corresponding source word(s). It should not be used for *terms* that are mistranslated (but which could be appropriate translations for the words in question in another context). At the word level, an example would be if the German word *konsequent* is translated as English *consequently*, rather than the proper *consistently*.
- **Omission.** *Content is missing from the translation that is present in the source.*

Example: A source text refers to a “mouse pointer” but the translation does not mention it.

Note(s): *Omission* should be reserved for those cases where content present in the source and essential to its meaning is not found in the target text.
- **Addition.** *The target text includes text not present in the source.*

Example: A translation includes portions of another translation that were inadvertently pasted into the document.
- **Untranslated.** *Content that should have been translated has been left untranslated.*

Example: A sentence in a Japanese document translated into English is left in Japanese.

Note(s): As noted above, if a term is passed through untranslated, it should be classified as *Untranslated* rather than as *Terminology*.
- **Fluency.** Fluency relates to the monolingual qualities of the source or target text, relative to agreed-upon specifications, but independent of relationship between source and target. In other words, fluency issues can be assessed without regard to whether the text is a translation or not. For example, a spelling error or a problem with register remain issues regardless of whether the text is translated or not.
- **Style.** *The text has stylistic problems, other than those related to language register.*

Example: A text uses a confusing style with long sentences that are difficult to understand.

Note(s): *Style* includes many factors where meaning is preserved accurately and grammar is correct, but where the text is considered

D1.3.1 Barriers for High-Quality Machine Translation

problematic. For example, IT style guides might dictate that writers and translators should avoid the use of the passive voice, but the passive voice appears in the text.

- **Spelling.** *Issues related to spelling of words*
Example: The German word *Zustellung* is spelled *Zustetlugn*.
- **Capitalization.** *Issues related to capitalization*
Example: The name John Smith is written as “john smith”.
- **Typography.** *Issues related to the mechanical presentation of text. This category should be used for any typographical errors other than spelling.*
Example: Extra, unneeded carriage returns are present in a translated text.
Note(s): For punctuation-related issues, always use the more specific *Punctuation*. Issues related to spelling should be listed under *Spelling*.
- **Punctuation.** *Punctuation is used incorrectly for the locale or style⁷*
Example: An English text uses a semicolon where a comma should be used.
- **Grammar.** *Issues related to the grammar or syntax of the text, other than spelling and orthography.*
Example: An English text reads “The man was in seeing the his wife.”
Note(s): Use *Grammar* only if no subtype accurately describes the issue.
- **Morphology (word form).** *There is a problem in the internal construction of a word*
Example: An English text has *comed* instead of *came*.
Note(s): *Morphology* applies *only* to cases where the problem is found in the internal construction of a word and that word would be incorrect for its intended meaning. It should not be used for agreement errors or other errors where the individual word form is correct.
- **Part of speech.** *A word is the wrong part of speech*
Example: A text reads “Read these instructions careful” instead of “Read these instructions carefully.”
- **Agreement.** *Two or more words do not agree with respect to case, number, person, or other grammatical features*
Example: A text reads “They was expecting a report.”

⁷ For cases of systematic uses of punctuation from the wrong locale, use *Quote marks format* under *Locale convention* instead.

D1.3.1 Barriers for High-Quality Machine Translation

Note(s): *Agreement* can be distinguished from *Morphology* in that the individual words in an *Agreement* error would be fine in isolation, but the combination of them does not work.

- **Word order.** *The word order is incorrect*

Example: A German text reads “Er hat gesehen den Mann” instead of “Er hat den Mann gesehen.”

Note(s): For issues related to word order where the sentence is grammatical and correct, but awkward, use *Style* instead.

- **Function words.** *Linguistic function words such as prepositions, particles, and pronouns are used incorrectly*

Example: An English text reads “He beat him around” instead of “he beat him up.”

Note(s): *Function words* is used for cases where individual words with a grammatical function are used incorrectly. The most common problems will have to do with prepositions, and particles. For languages where verbal prefixes play a significant role in meaning (as in German), they should be included here, even if they are not independent words.

- **Tense/aspect/mood.** *A verbal form inappropriate for the context is used*

Example: An English text reads “Yesterday he sees his friend” instead of “Yesterday he saw his friend”; an English text reads “The button must press” instead of “The button must be pressed”.

Note(s): In the case of subject-verb agreement errors, use *Agreement* instead

- **Unintelligible.** *The exact nature of the error cannot be determined. Indicates a major break down in fluency.*

Example: The following text appears in an English translation of a German automotive manual: “The brake from whe this दुदरु सि S149235 part numbr,,»

Note(s): If an issue is categorized as *Unintelligible* no further categorization is required. *Unintelligible* can refer to texts where a significant number of issues combine to create a text for which no further determination of error type can be made or where the relationship of target to source is entirely unclear.

D1.3.1 Barriers for High-Quality Machine Translation

10.2. Revised Annotation Guidelines

The document which follows this page comprises the latest version of the Annotation Guidelines (as of 2014 February 6). It includes the decision tree for selecting issue types from the revised list of issues in the Shared Task metric.

Guide to selecting MQM issues for the MT Evaluation Metric

Selecting issues can be a complex task. In order to assist annotators, the QTLaunchPad project has prepared a decision tree that helps annotators select appropriate issues. While the primary purpose of this decision tree is for training annotators, they are encouraged to keep it at hand while annotating data for cases that may not be clear.

To use the decision tree, annotators should start at the upper left corner and answer questions to find appropriate issues. The decision tree is organized differently than the hierarchy in translate5 in that it eliminates more specific issue types before arriving at general issue types. In practical terms annotators may stop at any point in the hierarchy if they know that the more specific issue types do not apply based on their annotation experience (see the discussion of selecting issue types in **Section 3** below). Evaluators are encouraged to add notes in translate5 to explain any decisions that may not be entirely clear or to provide information needed to understand issues, such as notes about what has been omitted in a translation.

In addition to the decision tree, evaluators are encouraged to consider the following guidelines. For questions that are not answered, reviewers should send inquiries to info@qt21.eu.

1. What is an error?

An error represents any issue you may find with the translated text that either does not correspond to the source or is considered incorrect in the target language. The list of language issues upon which you are to base your annotation is described in detail below and provides a range of examples.

The list is divided into two main issue categories, **Accuracy** and **Fluency**, each of which contains relevant, more detailed subcategories. Whenever possible, the correct subcategory should be chosen; however, if in doubt, please do not guess. Instead, select the category level about which you are most certain in order to avoid inconsistencies in the results.

Example: The German term *Zoomfaktor* was incorrectly translated as *zoom shot factor*, and you are unsure whether this represents a Mistranslation or an Addition. In this case, categorize the error as an Accuracy error since it is unclear whether content has been added or a term mistranslated.

2. When one error is “right” if the other is fixed

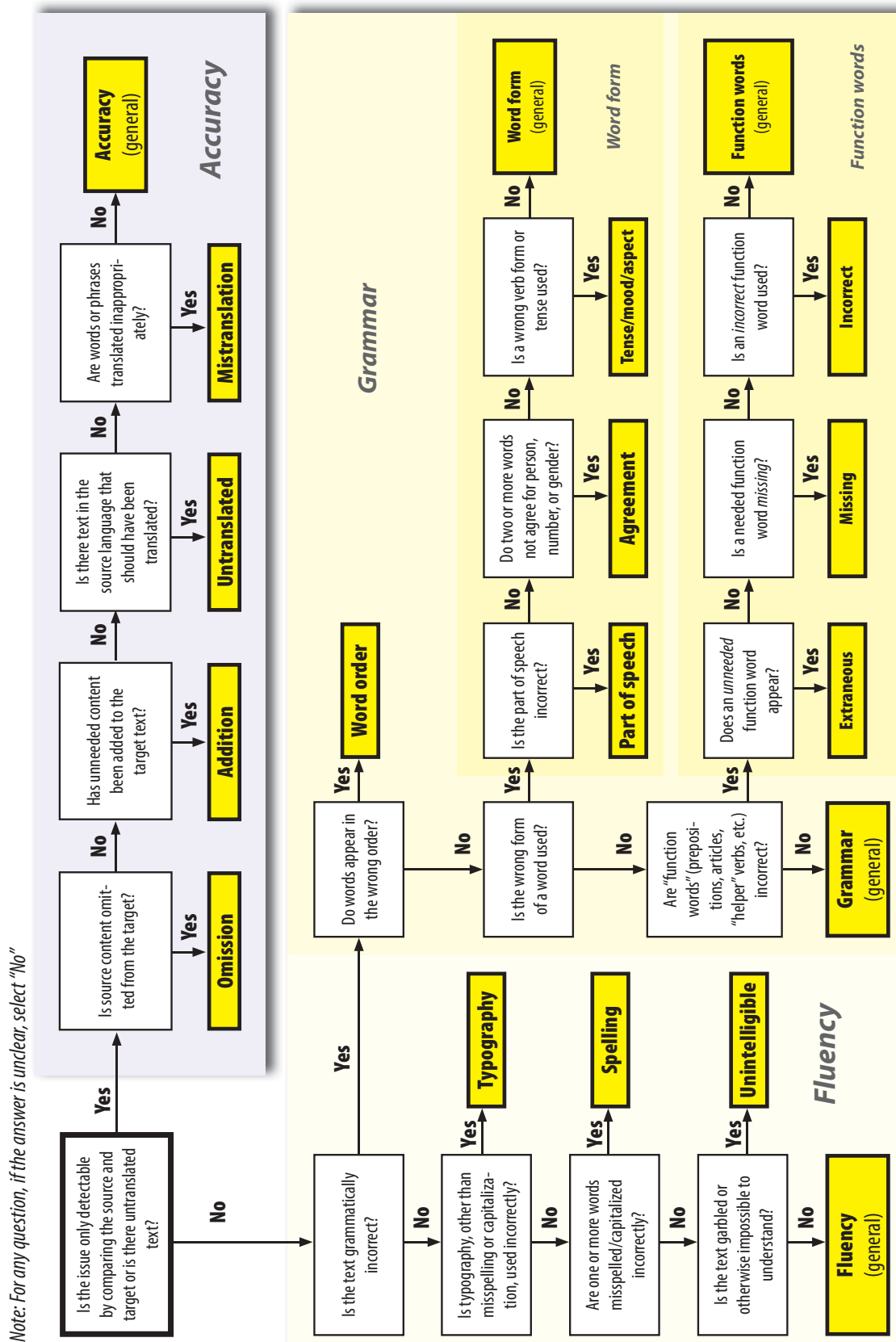
At first glance it may appear as if more than one error has been made, when in fact this may not be the case. An **Agreement** error, for example, will only have one issue, not two, so it is important to understand what the actual underlying error is before rejecting a potential “near miss” translation that could be used for annotation.

Examples

Source: Importfilter werden geladen
 Translation: Import filter are being loaded
 Correct: Import filters are being loaded

In this example, the only error is the translation of *filter* in the singular rather than the plural. Had the subject been translated properly, agreement would have been correct. **Agreement** is

D1.3.1 Barriers for High-Quality Machine Translation



D1.3.1 Barriers for High-Quality Machine Translation

therefore not an error in this case. Only *Import filter* would be tagged in this sentence as a **Mis-translation**.

Source: im Dialog Exportieren
 Translation: in the dialog export
 Correct: in the Export dialog

In this example, only **Mistranslation** should be marked. While **Word order** and **Spelling** (capitalization) would be considered errors in other contexts, this would not be the case here, as these two words constitute one term.

3. The Annotation Process

The translations you annotate should be a set of “near miss” (i.e., “almost perfect”) translations to annotate. Please follow these rules when selecting errors and tagging the respective text in the translations:

5. If you do not understand what types of errors belong to a particular category, please see the examples in this documentation.
6. In many cases multiple issue types may describe a single actual issue/error in the text. For example, if a text displays incorrect word order, it could be listed as both a **Grammar** error and a **Word order** error. In such cases issues should be tagged with only one issue type based on the following principles:
 - a. If it is possible to assign an issue to a more specific type (e.g., **Part of speech**) rather than a more general one (e.g., **Grammar**), the more specific one should be used. (But see item #6 below: do not use a category that does not clearly apply. If no category at a certain level is precise, use a more general one.)
 - b. General types should be used for cases in which the issue is of a more general nature or where the specific problem does not have a precise type in the hierarchy. For example *He slept the baby* uses a direct object with an intransitive verb (a problem technically known as a *valency error*), but as there is no specific type for this error, it should be assigned to **Grammar**.
 - c. If multiple issue types apply at the same level in the hierarchy, they should be selected according to the priority in the decision tree.
7. **Less is more.** Only tag the relevant text. Do not tag the entire sentence, as this will cause errors to occur in the analysis of the actual error itself and potentially render the results unacceptable. For example, if a single word is wrong in a phrase, tag only the single word rather than the entire phrase. (See the section on “minimal markup” below.)
8. Always look for the predominant issue. If that issue is fixed, do the others still exist? For example, if correcting an **Agreement** error would fix other related issues that derive from it, tag only the **Agreement** error, not the resulting errors.
9. One word in a sentence may contain two errors (e.g., it could be a **Misspelling** and an **Extraneous function word**). If this is the case, enter both errors separately and mark the respective word in both cases.
10. If in doubt, choose a more general category. The categories **Accuracy** and **Fluency** can be used if the nature of an error is unclear. In such cases, providing notes to explain the problem will assist the QTLaunchPad team in its research.

D1.3.1 Barriers for High-Quality Machine Translation

4. Tricky cases

- **Hyphenation:** Hyphenation issues are spelling issues or represent part of an untranslated term.
Example: *XML-files* represents a spelling error (**Misspelling**)
Pocket-PC represents an untranslated term (**Untranslated**)
- **Number** (plural vs. singular) is a **Mistranslation**
- **Terminology:** Inappropriate use of terms is classified as **Mistranslation**.
Example: An English translation uses the term *thumb drive* to translate the German *USB Speicherkarte*. This translation is intelligible, but if the translation mandated in specifications or a relevant termbase is *USB memory stick*, the use of *thumb drive* constitutes a **Mistranslation**, even if *thumb drive* would be acceptable in everyday usage.
- **Unintelligible:** If you cannot understand the sentence because a verb is missing or a word has been incorrectly translated, or perhaps both of these together, the sentence would not be considered unintelligible. A sentence that can easily be fixed should be categorized accordingly. An Unintelligible sentence is generally one whose word order issues and omissions are relatively complicated to remedy and where the meaning is unclear.
Example: In the sentence “You can also you can use this tab to precision, with the colors are described as well as the PostScript Level,” there are enough errors that the meaning is unclear and the precise nature of the errors that lead to its unintelligibility cannot be easily determined.
- **Agreement:** This category generally refers to agreement between subject and predicate or gender and case.
Examples: A boy plays with his train (not her train)
 I am at work (not I is at work)
- **Untranslated:** Many words may look as if they have been translated and simply forgotten to apply proper capitalization or hyphenations rules. In most, cases, this would represent an untranslated term and not a **Misspelling**. If the target word or phrase is identical to the source word or phrase, it should be treated as **Untranslated**, even if a **Misspelling** error could also account for the problem.

5. Minimal markup

It is vital in creating error markup that errors be marked up with the shortest possible spans. Markup should identify only that area needed to identify the problem. In some cases, ideal spans would be discontinuous, in which case the whole span that includes the issue should be marked.

Agreement poses special challenges because portions that disagree may be widely separated. To select appropriate minimal spans, consider the following guidelines:

- If two items disagree and it is readily apparent which should be fixed, mark just the portion that needs to be fixed. E.g., in “The man and its companion were business partners” it is readily apparent that *its* should be *his* and the wrong grammatical gender has been used, so only *its* should be marked.
- If two items disagree and it is not clear which portion is incorrect, mark the *entire span* containing the items in question, as shown in the example in the table above.

The following examples help clarify the general principles:

D1.3.1 Barriers for High-Quality Machine Translation

Incorrect markup	Problem	Correct minimal markup
Double click on the number faded in the status bar. [Mistranslation]	Only the single word <i>faded</i> is problematic, but the markup indicates that number faded in is incorrect.	Double click on the number faded in the status bar.
The standard font size for dialogs is 12pt, which corresponds to a standard of 100%. [Mistranslation]	Only the term <i>Maßstab</i> has been translated incorrectly. The larger span indicates that text that is perfectly fine has a problem.	The standard font size for dialogs is 12pt, which corresponds to a standard of 100%.
The man and its companion were business partners.	In this example, it is clear that <i>its</i> is the problematic portion, and that <i>man</i> is correct, so only <i>its</i> should be marked.	The man and its companion were business partners.
The man whom they saw on Friday night at the store were very big. [Agreement]	In this example it is not clear whether <i>man</i> or <i>were</i> is the error since there is nothing to indicate whether singular or plural is intended. Here the highlighted portion identifies only a single word, insufficient to identify the agreement problem. The correct version highlights the entire span containing the agreement error.	The man whom they saw on Friday night at the store were very big. [Agreement]
The in 1938 nascent leader with flair divined %temp_name eating lonely. [Marked for Unintelligible]	The entire sentence is Unintelligible and should be marked as such.	The in 1938 nascent leader with flair divined %temp_name eating lonely.

In the event of questions about the scope of markup that should be used, utilize the note field to make a query or explain your choice.

A. Issue categories

The error corpus uses the following issue categories:

- **Accuracy.** **Accuracy** addresses the extent to which the target text accurately renders the meaning of the source text. For example, if a translated text tells the user to push a button when the source tell the user not to push it, there is an accuracy issue.
- **Mistranslation.** The target content does not accurately represent the source content.
 - Example:** A source text states that a medicine should not be administered in doses greater than 200 mg, but the translation states that it should *not* be administered in doses less than 200 mg.
 - Note(s):** **Mistranslation** can be used for both words and phrases.
- **Omission.** Content is missing from the translation that is present in the source.
 - Example:** A source text refers to a “mouse pointer” but the translation does not mention it.
 - Note(s):** **Omission** should be reserved for those cases where content present in the source and essential to its meaning is not found in the target text.
- **Addition.** The target text includes text not present in the source.
 - Example:** A translation includes portions of another translation that were inadvertently pasted into the document.

D1.3.1 Barriers for High-Quality Machine Translation

- **Untranslated.** Content that should have been translated has been left untranslated.
 - Example:** A sentence in a Japanese document translated into English is left in Japanese.
 - Note(s):** As noted above, if a term is passed through untranslated, it should be classified as **Untranslated** rather than as **Mistranslation**.
- **Fluency.** Fluency relates to the monolingual qualities of the source or target text, relative to agreed-upon specifications, but independent of relationship between source and target. In other words, fluency issues can be assessed without regard to whether the text is a translation or not. For example, a spelling error or a problem with register remain issues regardless of whether the text is translated or not.
- **Misspelling.** Issues related to spelling of words (including capitalization)
 - Examples:** The German word *Zustellung* is spelled *Zustetlughn*.
The name *John Smith* is written as “john smith”.
- **Typography.** Issues related to the mechanical presentation of text. This category should be used for any typographical errors other than spelling.
 - Examples:** Extra, unneeded carriage returns are present in a text.
A semicolon is used in place of a comma.
- **Grammar.** Issues related to the grammar or syntax of the text, other than spelling and orthography.
 - Example:** An English text reads “The man was in seeing the his wife.”
 - Note(s):** Use **Grammar** only if no subtype accurately describes the issue.
- **Word form.** The wrong form of a word is used. Subtypes should be used when possible.
 - Example:** An English text has *comed* instead of *came*.
- **Part of speech. A word is the wrong part of speech**
 - Example:** A text reads “Read these instructions *careful*” instead of “Read these instructions *carefully*.”
- **Agreement. Two or more words do not agree with respect to case, number, person, or other grammatical features**
 - Example:** A text reads “*They was* expecting a report.”
- **Tense/aspect/mood. A verbal form inappropriate for the context is used**
 - Example:** An English text reads “Yesterday he *sees* his friend” instead of “Yesterday he *saw* his friend”; an English text reads “The button must be *pressing*” instead of “The button must be *pressed*”.
- **Word order.** The word order is incorrect
 - Example:** A German text reads “Er hat gesehen den Mann” instead of “Er hat den Mann gesehen.”

D1.3.1 Barriers for High-Quality Machine Translation

- **Function words.** Linguistic function words such as prepositions, particles, and pronouns are used incorrectly

Example: An English text reads “He beat him around” instead of “he beat him up.”

Note(s): Function words is used for cases where individual words with a grammatical function are used incorrectly. The most common problems will have to do with prepositions, and particles. For languages where verbal prefixes play a significant role in meaning (as in German), they should be included here, even if they are not independent words.

There are three subtypes of *Function words*. These are used to indicate whether an unneeded function word is present (**Extraneous**), a needed function word is missing (**Missing**), or a incorrect function word is used (**Incorrect**). Annotators should use the note field to specify details for missing function words.

- **Unintelligible.** The exact nature of the error cannot be determined. Indicates a major break down in fluency.

Example: The following text appears in an English translation of a German automotive manual: “The brake from whe this दुतरो सि S149235 part numbr,,”

Note(s): Use this category sparingly for cases where further analysis is too uncertain to be useful. If an issue is categorized as **Unintelligible** no further categorization is required. Unintelligible can refer to texts where a significant number of issues combine to create a text for which no further determination of error type can be made or where the relationship of target to source is entirely unclear.

D1.3.1 Barriers for High-Quality Machine Translation

11. Revised training materials

This section includes the revised training materials used in the second annotation round described in Section 1 of this revised version.

D1.3.1 Barriers for High-Quality Machine Translation

Guide to selecting MQM issues for the MT Evaluation Metric

version 1.3 (2014 June 11)

Selecting issues can be a complex task. In order to assist annotators, the QTLaunchPad project has prepared a decision tree that helps annotators select appropriate issues. **Use the decision tree not only for learning about MQM issues, but to guide your annotation efforts and resolve any questions or concerns you may have.**

Start at the upper left corner of the decision tree and then answer the questions and follow the arrows to find appropriate issues. The decision tree is organized a bit differently than the hierarchy in translate5 because it eliminates specific issue types before moving to general ones, so you familiarize yourself with how issues are organized in translate5 before beginning annotation.

Add notes in translate5 to explain any decisions that you feel need clarification, to ask questions, or to provide information needed to understand issues, such as notes about what has been omitted in a translation.

In addition to using the decision tree, please understand and follow the guidelines in this document. Email us at info@qt21.eu if you have questions that the decision tree and other content in this document do not address.

1. What is an error?

An error represents any issue you may find with the translated text that either does not correspond to the source or is considered incorrect in the target language. The list of language issues upon which you are to base your annotation is described in detail below and provides a range of examples.

The list is divided into two main issue categories, **Accuracy** and **Fluency**, each of which contains relevant, more detailed subcategories. Whenever possible, the correct subcategory should be chosen; however, if in doubt, please do not guess. Instead, select the category level about which you are most certain in order to avoid inconsistencies in the results.

Example: The German term *Zoomfaktor* was incorrectly translated as *zoom shot factor*, and you are unsure whether this represents a **Mistranslation** or an **Addition**. In this case, categorize the error as an **Accuracy** error since it is unclear whether content has been added or a term mistranslated.

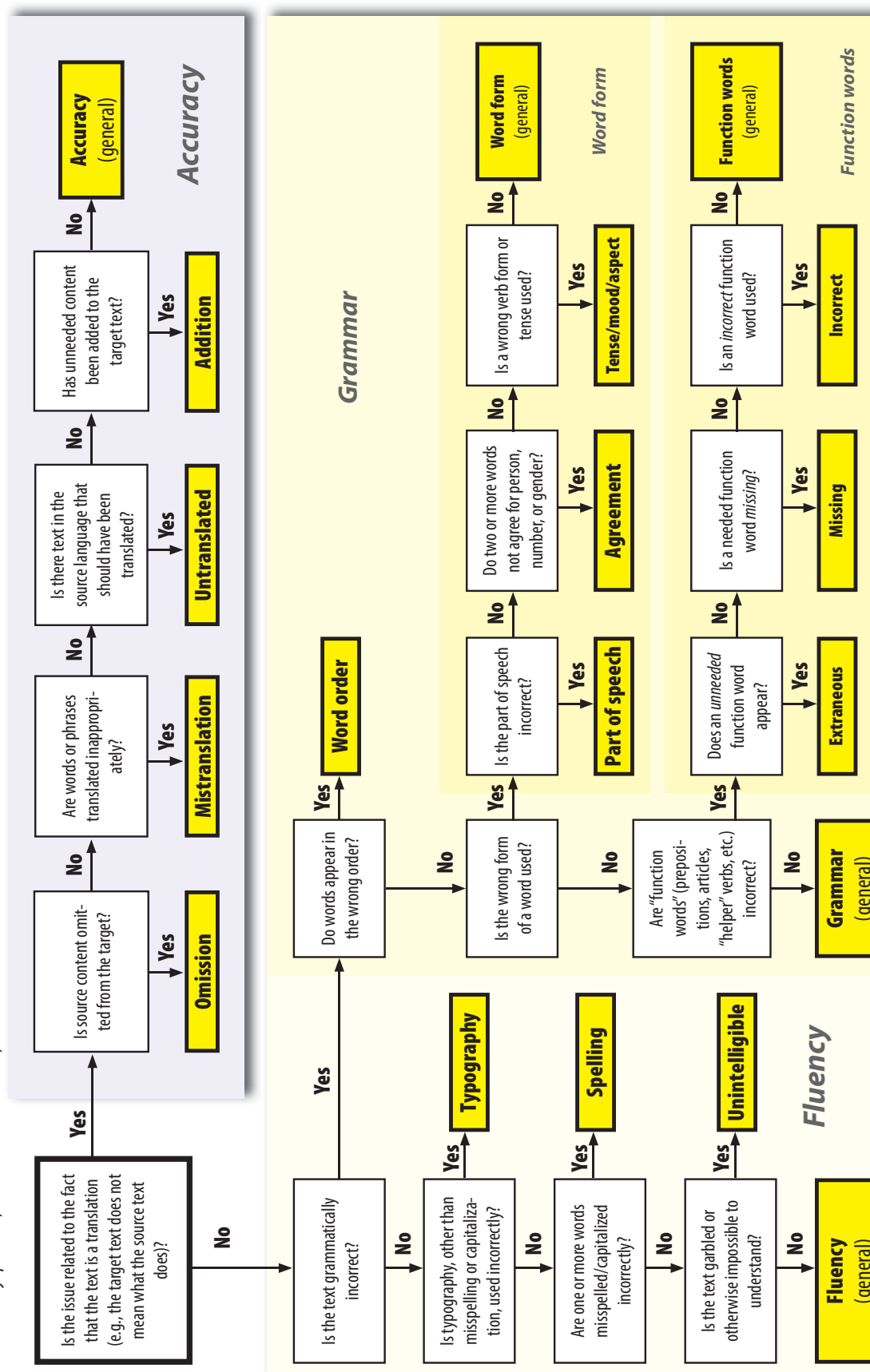
2. The Annotation Process

The translations you annotate should be a set of “near miss” (i.e., “almost perfect”) translations to annotate. Please follow these rules when selecting errors and tagging the respective text in the translations:

1. Use the examples in this documentation to understand specific classes.
2. If multiple types could be used to describe an issue (e.g., **Agreement**, **Word form**, **Grammar**, and **Fluency**), select the first one that the decision tree guides you to. The tree is organized along the following principles:
 - a. It prefers more specific types (e.g., **Part of speech**) to general ones (e.g., **Grammar**). However, if a specific type does not apply, it guides you to use the general type.
 - b. General types are used where the problem is of a general nature *or where the specific problem does not have a precise type*. For example *He slept the baby* exhibits what is technically known as a *valency error*, but because there is no specific type for this error available, it is assigned to **Grammar**.
3. **Less is more.** Only tag the relevant text. For example, if a single word is wrong in a phrase, tag only the single word rather than the entire phrase. If two words, separated by other words, constitute an error, mark only those two words separately. (See the section on “minimal markup” below.)
4. If correcting one error would take care of others, tag *only* that error. For example, if fixing an **Agreement** error would fix other related issues that derive from it, tag only the **Agreement** error, not the errors that result from it.

D1.3.1 Barriers for High-Quality Machine Translation

Note: For any question, if the answer is unclear, select "No"



D1.3.1 Barriers for High-Quality Machine Translation

Examples

Source: Importfilter werden geladen
 Translation: Import filter are being loaded
 Correct: Import filters are being loaded

In this example, the only error is the translation of *filter* in the singular rather than the plural (as made clear by the verb form in the source text). This case should be classified as **Mistranslation**, even though it shows problems with agreement: if the subject had been translated properly the agreement problem would be resolved. In this case only *filter* should be tagged as a **Mistranslation**.

Source: im Dialog Exportieren
 Translation: in the dialog export
 Correct: in the Export dialog

In this example, only **Mistranslation** should be marked. While **Word order** and **Spelling** (capitalization) would be considered errors in other contexts, this would not be the case here, as these two words constitute one term that has been incorrectly translated.

5. If one word contains two errors (e.g., it has a **Spelling** issue and is also an **Extraneous function word**), enter both errors separately and mark the respective word in both cases.
6. If in doubt, choose a more general category. The categories **Accuracy** and **Fluency** can be used if the nature of an error is unclear. In such cases, providing notes to explain the problem will assist the QTLaunchPad team in its research.

3. Tricky cases

The following examples are ones that have been encountered in practice and that we wish to clarify.

- **Function words:** In some cases issues related to function words break the accuracy/fluency division seen in the decision tree because they are listed under *Fluency* even though they may impact meaning. Despite this issue, please categorize them as the appropriate class under *Function words*.

Example: *The ejector may be found **with** the external case* (should be **on** in this case). Even though this error changes the meaning, it should be classified as **Function words: incorrect** in the **Fluency** branch.

- **Word order:** Word order problems often affect long spans of text. When encountering word orders, mark the smallest possible portion that could be moved to correct the problem.

Example: *He has the man with the telescope seen.* Here only *seen* should be marked as moving this one word would fix the problem.

- **Hyphenation:** Hyphenation issues sometimes occur in untranslated content and should be classified as such. Otherwise they should be classified as **Spelling**.

Example: *Load the XML-files* (**Spelling**)
Nützen Sie die macro-lens (**Untranslated**, if the source has *macro-lens* as well)

- **Number** (plural vs. singular) is a **Mistranslation**.
- **Terminology:** Inappropriate use of terms is classified as **Mistranslation**.

Example: An English translation uses the term *thumb drive* to translate the German *USB Speicherkarte*. This translation is intelligible, but if the translation mandated in specifications or a relevant termbase is *USB memory stick*, the use of *thumb drive* constitutes a **Mistranslation**, even if *thumb drive* would be acceptable in everyday usage.

D1.3.1 Barriers for High-Quality Machine Translation

- **Unintelligible:** Use **Unintelligible** if content cannot be understood and the reason cannot be analyzed according to the decision tree. *This category is used as a last resort for text where the nature of the problem is not clear at all.*

Example: In the sentence “You can also you can use this tab to precision, with the colours are described as well as the PostScript Level,” there are enough errors that the meaning is unclear and the precise nature of the errors that lead to its unintelligibility cannot be easily determined.

- **Agreement:** This category generally refers to agreement between subject and predicate or gender and case.

Examples: The boy was playing with *her* own train
I *is* at work

- **Untranslated:** Many words may look as if they have been translated and simply forgotten to apply proper capitalization or hyphenations rules. In most, cases, this would represent an untranslated term and not a **Spelling**. If the target word or phrase is identical to the source word or phrase, it should be treated as **Untranslated**, even if a **Spelling** error could also account for the problem.

4. Minimal markup

It is vital in creating error markup that errors be marked up with the shortest possible spans. Markup must identify *only* that area needed to specify the problem. In some cases this requirement means that two separate spans must be identified.

The following examples help clarify the general principles:

Incorrect markup	Problem	Correct minimal markup
Double click on the number faded in the status bar. [Mistranslation]	Only the single word <i>faded</i> is problematic, but the markup indicates that number faded in is incorrect.	Double click on the number faded in the status bar.
The standard font size for dialogs is 12pt, which corresponds to a standard of 100% . [Mistranslation]	Only the term <i>Maßstab</i> has been translated incorrectly. The larger span indicates that text that is perfectly fine has a problem.	The standard font size for dialogs is 12pt, which corresponds to a standard of 100%.
The in 1938 nascent leader with flair divined %temp_name eating lonely. [Unintelligible]	The entire sentence is Unintelligible and should be marked as such.	The in 1938 nascent leader with flair divined %temp_name eating lonely.

As noted above, **Word order** can be problematic because it is often unclear what portion(s) of the text should be marked. In cases of word order, mark the *shortest* portion of text (in number of words) that could be moved to fix the problem. If two portions of the text could resolve the problem and are equal in length, mark the one that occurs first in the text. The following examples provide guidance:

Incorrect markup	Problem	Correct minimal markup
The telescope big observed the operation	Moving the word <i>telescope</i> would solve the problem and only this word should be marked (since it occurs first in the text).	The telescope big observed the operation
The eruption by many instruments was recorded.	Although this entire portion shows word order problems, moving <i>was recorded</i> would resolve the problem (and is the shortest span that would resolve the problem).	The eruption by many instruments was recorded.
The given policy in the manual user states that this action voids the warranty.	This example actually has two separate issues that should be marked separately.	The given policy in the manual user states that this action voids the warranty.

D1.3.1 Barriers for High-Quality Machine Translation

Agreement poses special challenges because portions that disagree may be widely separated. To select appropriate minimal spans, consider the following guidelines:

- If two items disagree and it is readily apparent which should be fixed, mark *only* the portion that needs to be fixed. E.g., in “The man and its companion were business partners” it is readily apparent that *its* should be *his* and the wrong grammatical gender has been used, so only *its* should be marked.
- If two items disagree and it is not clear which portion is incorrect, mark the both items and mark them for **Agreement**, as shown in the example in the table below.

The following examples demonstrate how to mark **Agreement**:

Incorrect markup	Problem	Correct minimal markup
The man and its companion were business partners. [Agreement]	In this example, it is clear that <i>its</i> is the problematic portion, and that <i>man</i> is correct, so only <i>its</i> should be marked.	The man and its companion were business partners.
The man whom they saw on Friday night at the store were very big. [Agreement]	In this example it is not clear whether <i>man</i> or <i>were</i> is the error since there is nothing to indicate whether singular or plural is intended. Here the highlighted portion identifies only a single word, insufficient to identify the agreement problem. The correct version highlights both words as separate issues. In such cases use the Notes field to explain the decision.	The man whom they saw on Friday night at the store were very big. [Agreement]

In the event of questions about the scope of markup that should be used, utilize the Notes field to make a query or explain your choice.

D1.3.1 Barriers for High-Quality Machine Translation

A. Issue categories

The error corpus uses the following issue categories:

- **Accuracy.** **Accuracy** addresses the extent to which the target text accurately renders the meaning of the source text. For example, if a translated text tells the user to push a button when the source tell the user not to push it, there is an accuracy issue.
- **Mistranslation.** The target content does not accurately represent the source content.
 - Example:** A source text states that a medicine should not be administered in doses greater than 200 mg, but the translation states that it should *not* be administered in doses less than 200 mg.
 - Note(s):** **Mistranslation** can be used for both words and phrases.
- **Omission.** Content is missing from the translation that is present in the source.
 - Example:** A source text refers to a “mouse pointer” but the translation does not mention it.
 - Note(s):** **Omission** should be reserved for those cases where content present in the source and essential to its meaning is not found in the target text.
- **Addition.** The target text includes text not present in the source.
 - Example:** A translation includes portions of another translation that were inadvertently pasted into the document.
- **Untranslated.** Content that should have been translated has been left untranslated.
 - Example:** A sentence in a Japanese document translated into English is left in Japanese.
 - Note(s):** As noted above, if a term is passed through untranslated, it should be classified as **Untranslated** rather than as **Mistranslation**.
- **Fluency.** Fluency relates to the monolingual qualities of the source or target text, relative to agreed-upon specifications, but independent of relationship between source and target. In other words, fluency issues can be assessed without regard to whether the text is a translation or not. For example, a spelling error or a problem with register remain issues regardless of whether the text is translated or not.
- **Spelling.** Issues related to spelling of words (including capitalization)
 - Examples:** The German word *Zustellung* is spelled *Zustetlugn*.
The name *John Smith* is written as “john smith”.
- **Typography.** Issues related to the mechanical presentation of text. This category should be used for any typographical errors other than spelling.
 - Examples:** Extra, unneeded carriage returns are present in a text.
A semicolon is used in place of a comma.
- **Grammar.** Issues related to the grammar or syntax of the text, other than spelling and orthography.
 - Example:** An English text reads “The man was in seeing the his wife.”
 - Note(s):** Use **Grammar** only if no subtype accurately describes the issue.
- **Word form.** The wrong form of a word is used. Subtypes should be used when possible.
 - Example:** An English text has *comed* instead of *came*.
- **Part of speech.** A word is the wrong part of speech
 - Example:** A text reads “Read these instructions *careful*” instead of “Read these instructions *carefully*”.

D1.3.1 Barriers for High-Quality Machine Translation

- **Agreement.** Two or more words do not agree with respect to case, number, person, or other grammatical features

Example: A text reads "*They was* expecting a report."

- **Tense/aspect/mood.** A verbal form inappropriate for the context is used

Example: An English text reads "Yesterday he *sees* his friend" instead of "Yesterday he *saw* his friend"; an English text reads "The button must be *pressing*" instead of "The button must be *pressed*".

- **Word order.** The word order is incorrect

Example: A German text reads "Er hat gesehen den Mann" instead of "Er hat den Mann gesehen."

- **Function words.** Linguistic function words such as prepositions, particles, and pronouns are used incorrectly

Example: An English text reads "He beat him around" instead of "he beat him up."

Note(s): Function words is used for cases where individual words with a grammatical function are used incorrectly. The most common problems will have to do with prepositions, and particles. For languages where verbal prefixes play a significant role in meaning (as in German), they should be included here, even if they are not independent words.

There are three subtypes of *Function words*. These are used to indicate whether an unneeded function word is present (**Extraneous**), a needed function word is missing (**Missing**), or an incorrect function word is used (**Incorrect**). Annotators should use the note field to specify details for missing function words.

- **Unintelligible.** The exact nature of the error cannot be determined. Indicates a major break down in fluency.

Example: The following text appears in an English translation of a German automotive manual: "The brake from whe this कृतारो सि S149235 part numbr,,"

Note(s): Use this category sparingly for cases where further analysis is too uncertain to be useful. If an issue is categorized as **Unintelligible** no further categorization is required. Unintelligible can refer to texts where a significant number of issues combine to create a text for which no further determination of error type can be made or where the relationship of target to source is entirely unclear.

D1.3.1 Barriers for High-Quality Machine Translation

12. Change log

12.1. July 2014

- Added section 1, Updated Results, including information on inter-annotator agreement, data structures, the revised issue set, annotators, and the online presentation of data

12.2. December 2014

- Fixed incorrect figure captions
- Added section 1.1.2 on annotation costs
- Added link to supplemental report on practical aspects of annotation
- Information about the systems used was clarified. However, in most cases we know little about the specific components of the systems used (either because the translations were done by LSPs or because WMT did not provide details of all systems).
- Created a separate report (at <http://qt21.eu/downloads/MQM-usage-guidelines.pdf>) that addresses costs, tool selection, and other practical aspects.