Relating Translation Quality Barriers to Source-Text Properties

Federico Gaspari[*], Antonio Toral[*], Arle Lommel[^], Stephen Doherty[°], Josef van Genabith[§], Andy Way[*]

[*] School of Computing
Dublin City University
Glasnevin
Dublin 9
Ireland

[^] DFKI GmbH
Language Technology
Alt Moabit 91c
D-10559 Berlin
Germany

[°] School of Humanities & Languages
University of New South Wales
Sydney 2052
Australia

[§] DFKI GmbH
Language Technology
Campus D3 2
D-66123 Saarbrücken
Germany

E-mail: {fgaspari, atoral, away}@computing.dcu.ie, arle.lommel@dfki.de,
s.doherty@unsw.edu.au, josef.van_genabith@dfki.de

Abstract

This paper aims to automatically identify which linguistic phenomena represent barriers to better MT quality. We focus on the translation of news data for two bidirectional language pairs: EN↔ES and EN↔DE. Using the diagnostic MT evaluation toolkit DELiC4MT and a set of human reference translations, we relate translation quality barriers to a selection of 9 source-side PoS-based linguistic checkpoints. Using output from the winning SMT, RbMT, and hybrid systems of the WMT 2013 shared task, translation quality barriers are investigated (in relation to the selected linguistic checkpoints) according to two main variables: (i) the type of the MT approach, i.e. statistical, rule-based or hybrid, and (ii) the human evaluation of MT output, ranked into three quality groups corresponding to good, near miss and poor. We show that the combination of manual quality ranking and automatic diagnostic evaluation on a set of PoS-based linguistic checkpoints is able to identify the specific quality barriers of different MT system types across the four translation directions under consideration.

Keywords: MT quality barriers, diagnostic evaluation, statistical/rule-based/hybrid MT, linguistic features

1. Introduction

This study was conducted as part of the European Commission-funded project QTLaunchPad (Seventh Framework Programme (FP7), grant number: 296347), preparing groundwork for major developments in translation technology, with a special focus on identifying and overcoming barriers to translation quality.[1] Key goals of the project include providing test suites and tools for translation quality assessment, creating a shared quality metric for human and machine translation (MT), and improving automatic translation quality estimation. The project involves key research and industrial stakeholders interested in improving translation technology.

This paper presents work on the identification of translation quality barriers, one of the central objectives of QTLaunchPad. Given the widely perceived need to enhance MT quality and the reliability of MT evaluation for real-life applications, which has been confirmed further by QTLaunchPad surveys,[2] this study is of potential interest to a variety of MT users and developers. The main motivation behind the research is to systematically tackle quality barriers in MT, investigating closely the relationship between different types of MT systems, the overall quality of their output and the properties of the input. A key part of the work conducted in QTLaunchPad addresses this problem, with the goal of improving MT performance and extending its applicability.

Our study focuses on identifying the source-side linguistic properties that pose MT quality barriers for specific types of MT systems (statistical, rule-based and hybrid) and for output representative of different quality levels (poor-, medium- and high-quality) in four translation combinations, considering English to and from Spanish and German. Many commentators say that developers of SMT systems (in particular) are not able to predict which linguistic phenomena their systems are capable of handling. In this paper, on the contrary, we demonstrate the potential of combining manual MT quality ranking and DELiC4MT (an automatic diagnostic MT evaluation toolkit focusing on source-side linguistic phenomena that is described in more detail in Section 2.1) to identify translation quality barriers.

The remainder of the paper is organised as follows. After this introduction, Section 2 presents DELiC4MT, focusing on the novelty of its application to the discovery of translation quality barriers. Section 3 covers the evaluation, including the experimental set-up, the results for each of the four translation directions (paying special attention to the identified translation quality barriers in relation to the MT system types and to the quality rankings assigned to their output) and further correlation analysis. Finally, Section 4 summarises the main findings of the study and outlines possibilities for future work.

[1] www.qt21.eu/launchpad

[2] www.qt21.eu/launchpad/sites/default/files/QTLP_Survey2i.pdf

2. DELiC4MT for the Analysis of Translation Quality Barriers

2.1 DELiC4MT: an Open-Source Toolkit for Diagnostic MT Evaluation

DELiC4MT is an open-source toolkit for diagnostic MT evaluation (Toral et al., 2012).[3] 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 analyse when evaluating the quality of 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 granularity desired by the user, considering lexical, morphological, syntactic and/or semantic information.

Any of these layers of linguistic description can be combined to create 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 word sequences in the source text composed of 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 often a PoS tagger); clearly, some languages are better served than others in this respect. The data pre-processing steps that are required to use DELiC4MT are described in Section 3.1, while Section 3.3.2 discusses the way in which its output is presented to the user.

DELiC4MT produces a score, indicating how many of the relevant checkpoints detected on the source side were translated correctly by the MT system under investigation. This diagnostic feedback can then be incorporated into the further development, fine-tuning and customisation of the MT software to optimise its performance. One advantage of the toolkit over standard automatic MT evaluation metrics such as BLEU (Papineni et al., 2002) is that it supports more flexible, transparent and fine-grained evaluation: the scores of automatic MT evaluation metrics are often difficult to interpret and do not always help one to understand the actual linguistic strengths and weaknesses of an MT system.

Toral et al. (2012) describe the different modules that make up the DELiC4MT 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 tutorial is also available, showing how the toolkit works, applying it to a specific language pair, test set and linguistic checkpoint. [4] 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). The toolkit 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.

2.2 DELiC4MT-Based Analysis of Translation Quality Barriers

DELiC4MT has so far been used to evaluate the overall quality of MT systems with respect to their performance on user-defined source-side linguistic phenomena. The novelty of the work presented in this paper lies in the application of this toolkit to the investigation of translation quality barriers. These are investigated with DELiC4MT according to two main variables. Firstly, we consider different MT system types: this variable enables us to compare the performance of statistical (SMT), rule-based (RbMT) and hybrid (HMT) MT software on a selection of source-language linguistic checkpoints, which are explained in more detail in Section 2.3. We thus have a clear view of those quality barriers encountered by the various types of MT software for each translation direction, broken down according to a range of checkpoints as salient linguistically-motivated morphosyntactic units of evaluation.

Secondly, we look at human quality rankings of the MT output: this variable concerns the quality band assigned by human evaluators to the output of each MT system, whereby each sentence was rated as either good (rank 1), near-miss (rank 2) or poor (rank 3). We are thus able to evaluate the performance of the MT systems on each checkpoint separately for those sentences that fall into each of these rating bands. Both variables under consideration lend themselves to comparative evaluations, which are investigated in Section 3 with a view to shedding light on translation quality barriers.

2.3 From Linguistic Checkpoints to Translation Quality Barriers

On the basis of some preliminary tests, we decided to focus our analysis on linguistic checkpoints consisting of individual PoS classes (rather than PoS sequences), which were deemed sufficiently fine-grained to obtain interesting and useful information on translation quality barriers. This decision mitigated the data sparseness problems that we would have run into using more elaborate and specific linguistic checkpoints, given the limited amount of data available (cf. Section 3.1).

Following some explorations of the possibilities, we eventually selected 9 linguistic checkpoints for our analysis, consisting of the following individual PoS classes: adjectives (ADJ), adverbs (ADV), determiners (DET), common nouns (NOC), nouns (NOU, combining

[3] www.computing.dcu.ie/~atoral/delic4mt/

[4] http://github.com/antot/DELiC4MT/blob/master/doc/tutorial/delic4mt_tutorial.pdf

NOC and NOP), proper nouns (NOP), particles (PAR), pronouns (PRO) and verbs (VER). These are grouping abstractions over the possibly different PoS sets used for the three languages under investigation, and we thought that they represented a reasonable balance between granularity and high-level description. The scores provided in the analysis for any of these checkpoints 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 in question. Thus, the lower the score for a checkpoint, the worse the quality of the translations in the MT output for words corresponding to that linguistic phenomenon (in this study, PoS class) in the input, which reveals a potential translation quality barrier when referenced against the human evaluation of the output.

3. Evaluation

3.1 Data, Pre-Processing and Experimental Set-Up

We conducted this analysis of translation quality barriers focusing on news data, relying on the 2013 WMT data sets for which human reference translations were available. Table 1 shows the data used for the evaluation, detailing the number of sentences and the types of MT systems available for each translation direction.

Translation Direction	Number of Sentences	MT Systems
EN→ES	500	SMT, RbMT, HMT
ES→EN	203	SMT, RbMT
EN→DE	500	SMT, RbMT, HMT
DE→EN	500	SMT, RbMT

Table 1: Datasets used for the evaluation.[5]

The sentences used were translated by the winning MT systems from the 2013 WMT shared task. In this case, the SMT system is a phrase-based system from one of the leading European academic teams in MT research, while both the RbMT and HMT systems are leading systems on the market nowadays. Since these systems were used in the shared task, they had training/reference data consisting of news articles and the translations were all of novel sentences from news articles. It should be noted that WMT uses paid human translators to generate source sentences in all language pairs, so, for example, a segment authored in Spanish would be translated by a human into German and then translated from German by the MT systems into the various WMT target languages (including back into Spanish). To control for the issue of "pivot" or "relay" translation, our corpus used only "native" source segments, i.e., those segments authored in the source language of each language pair we considered.

Two Language Service Providers (LSPs) plus an in-house team at DFKI (for ES→EN only) carried out human assessment of the quality of the MT output for these sentences in the various language pairs, ranking them into three quality categories: rank 1 (perfect output, not requiring any editing to be published); rank 2 (near-misses, i.e. sentences with fewer than 3 errors, thereby deemed to be easily post-editable); and, finally, rank 3 (poor-quality output, with 3 or more errors, requiring time-consuming and resource-intensive post-editing).

In order to pre-process the data so that it could be used with DELiC4MT, we PoS-tagged the source and target sides of the references. Freeling [6] (Padró and Stanilovsky, 2012) was used for English and Spanish, with TreeTagger [7] (Schmid, 1995) used for German. Subsequently, the source and target sides of the reference were word-aligned with GIZA++ (Och and Ney, 2003). As the reference datasets were rather small for word alignment, in order to obtain alignments of higher quality, they were appended to a bigger corpus of the same domain as the WMT data (news commentary),[8] before performing word alignment. Once the alignment was ready, we extracted the subset that corresponded to the sentences of the reference set and discarded the rest.

Before proceeding further, we need to provide clarification regarding the data sets used. For each of the four translation directions, the diagnostic evaluation presented here concerns the very same input when comparisons of MT systems take place on the whole input data. In contrast, this is not the case for the identification of the quality barriers considering the MT output categorised according to the three quality rankings. This is because DELiC4MT was run separately on a subset of the input, depending on how that subset was classified by the human judges, resulting in three different data sets divided according to their quality. This means, for example, that the subset of rank 1 sentences translated with the SMT system for EN→ES is different from the subset of the same rank and translation direction for the RbMT system, so no direct comparison is possible in such cases.

3.2 Results

This section presents in turn the results obtained with DELiC4MT (Y axis in the figures below) on the 9 chosen linguistic checkpoints (X axis in the figures) for each of the four translation directions. This enables us to directly relate the translation quality barriers identified for each MT system type as well as across the three quality rankings to specific source-text properties. The figures in this section presenting the data (1-16) would be better represented by scatter plots. However, some of the data points for individual PoS-based linguistic checkpoints for the different MT system types are very close, which

[5] At the time of writing, the ES→EN data has only been partially rated, resulting in a smaller number of data points for this translation direction.

[6] http://nlp.lsi.upc.edu/freeling/

[7] www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[8] www.statmt.org/wmt13/translation-task.html#download

makes it difficult to differentiate them. As a result, in the interest of clarity, all the figures 1-16 include the trend lines connecting the data points for the various PoS-based linguistic checkpoints.

3.2.1. Results for EN→ES

One overall finding for the EN→ES language pair is that the SMT system is the best in general, followed by HMT and RbMT (in this order), even though SMT receives (virtually) the same scores as HMT for the PAR, PRO and VER linguistic checkpoints.

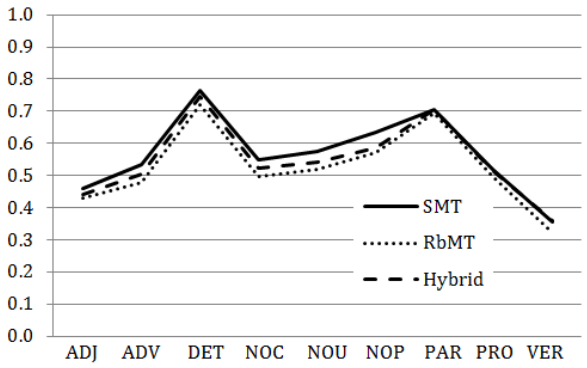

Figure 1: EN→ES results (overall).

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

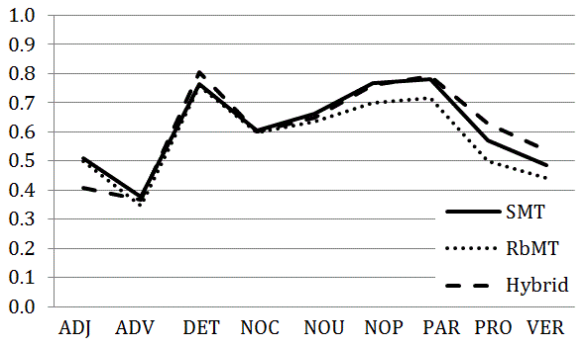

Figure 2: EN→ES results for rank 1.

As can be seen in Figure 3, rank 2 translations show similar results for all three systems, with RbMT lagging slightly behind, especially for ADV, NOC and VER. Equivalent trends can be observed in Figure 4 for rank 3 translations, with SMT obtaining an even bigger advantage on DET, NOU and NOP. For these three checkpoints HMT comes second, and RbMT last. However, we observe that the results by the three MT system types are very similar for the remaining checkpoints.

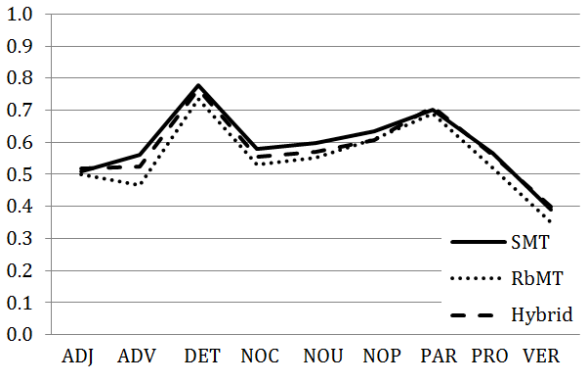

Figure 3: EN→ES results for rank 2.

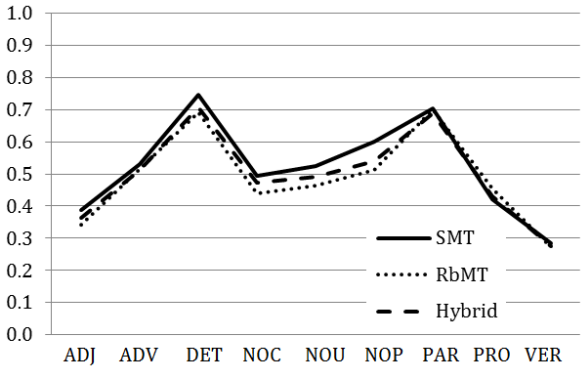

Figure 4: EN→ES results for rank 3.

3.2.2. Results for ES→EN

In overall terms, for the ES→EN translation direction the performance of SMT is consistently better than that of RbMT for all the 9 linguistic checkpoints, with approximately a 10% relative difference in the respective DELiC4MT scores. Particularly severe quality barriers for RbMT seem to be ADV, NOC, PRO and VER.

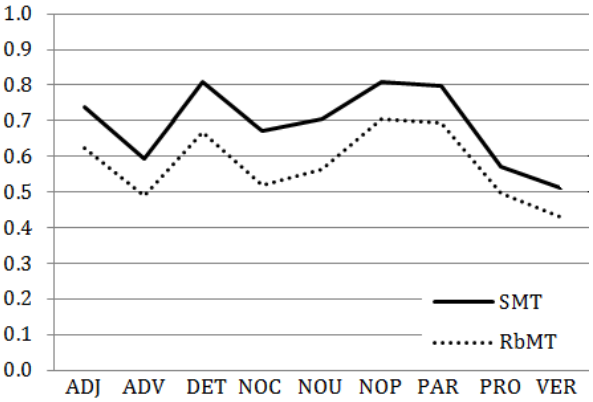

Figure 5: ES→EN results (overall).

More specifically, for rank 1 translations (bearing in mind the comparatively small numbers of checkpoint instances with respect to the other two quality bands, cf. Section 3.3.1 and in particular Tables 2 and 3), the performance of RbMT is particularly modest for ADJ, NOC, NOU, PAR, PRO and VER (Figure 6). On the other hand, SMT and RbMT have very similar performances for ADV and NOP in rank 1 translations, showing that these two categories are not specifically affected by quality differences depending on the two MT system types.

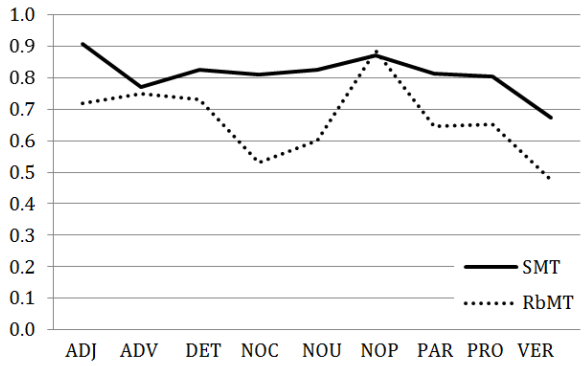

Figure 6: ES→EN results for rank 1.

The rank 2 translations show that SMT outperforms RbMT by a similar margin across all the linguistic checkpoints (Figure 7). As a result, in this case, the breakdown into the linguistic checkpoints does not allow us to gain particularly useful insights, showing that translation quality barriers are fairly consistent across the board for all the considered PoS-based linguistic checkpoints in near-miss translations, regardless of the type of MT system that generated them.

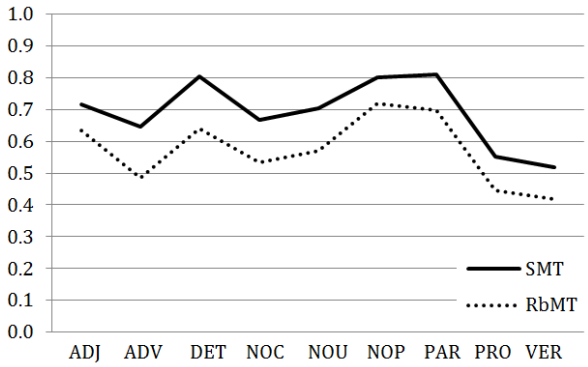

Figure 7: ES→EN results for rank 2.

The situation is more interesting for the rank 3 translations, where both SMT and RbMT show specific weaknesses in the translation of ADV, PRO and VER (Figure 8). Interestingly, although these three checkpoints show the lowest scores, they are also the ones where RbMT performs better than SMT, by a clear margin. For the remaining six checkpoints, the SMT output obtains higher scores, with a difference of approximately 10% in value at times.

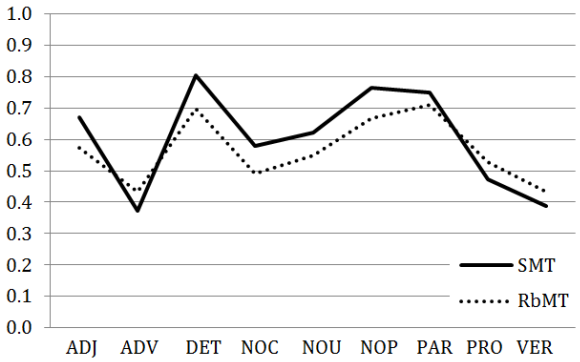

Figure 8: ES→EN results for rank 3.

3.2.3. Results for EN→DE

For the EN→DE translation direction, in overall terms, the performance of the three systems is very similar, with SMT giving slightly better scores than RbMT for all the checkpoints, while also beating HMT most of the time, except for PAR (where there is a tie), PRO and VER. As a result, it is difficult to identify prominent translation quality barriers from this analysis, except for a comparatively poor performance of RbMT, particularly for ADJ, NOC, NOU and PAR.

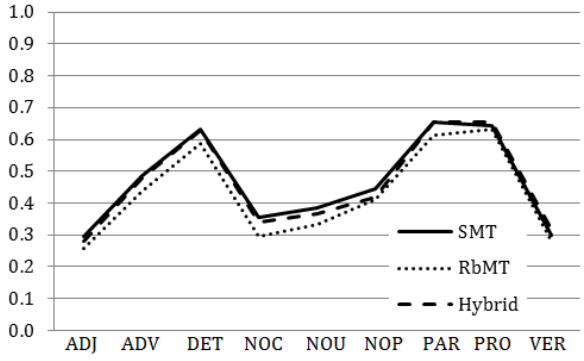

Figure 9: EN→DE results (overall).

Looking at the results by ranking, on the other hand, gives a more interesting picture. For rank 1 translations, SMT shows a particularly disappointing performance for NOC and NOU, while it is by far the top system for ADJ, NOP and PAR (Figure 10). RbMT receives the lowest score of the three systems for the ADJ checkpoint, where HMT also performs particularly badly. RbMT also showed the worst performance for VER, where HMT came out on top.

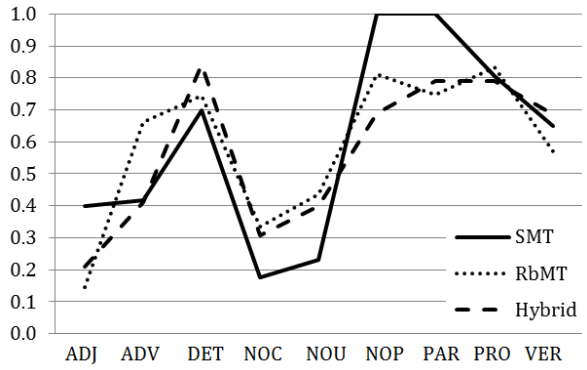

Figure 10: EN→DE results for rank 1.

The rank 2 translations (Figure 11) show a consistent trend, with SMT obtaining the best results for all the checkpoints (there is a tie with HMT for verbs), and RbMT lagging behind.

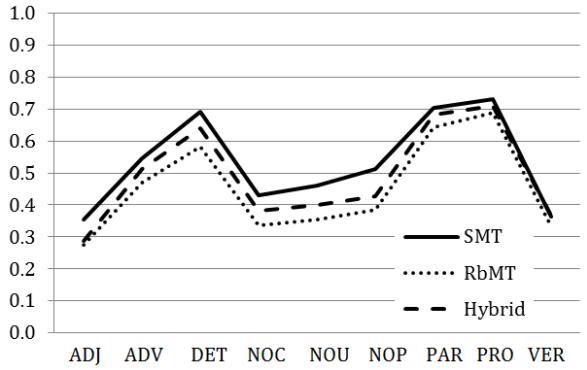

Figure 11: EN→DE results for rank 2.

Finally, looking at rank 3 translations (Figure 12), all three MT systems find ADJ and VER similarly problematic to translate (which was to be expected, due to a large extent to agreement problems), whereas RbMT runs into noticeably more difficulties with NOC. 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 this particular translation direction.

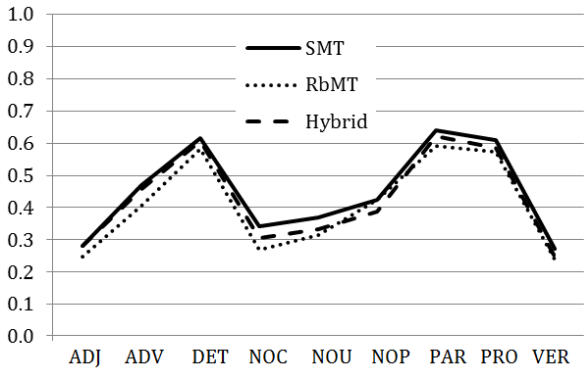

Figure 12: EN→DE results for rank 3.

3.2.4. Results for DE→EN

Finally, coming to the DE→EN translation direction, whose overall results are summarised in Figure 13, both SMT and RbMT encounter specific difficulties with the translation of ADJ, NOC and NOU checkpoints, with similarly low performance levels (the scores of RbMT are slightly lower in all these three cases). In contrast, DELiC4MT reveals that there are only relatively minor problems for the translation of DET, where both systems perform very well – determiners are much easier to translate from German into English, due to the much smaller set of non-inflected options available in the target. The other checkpoints show equivalent scores, but RbMT is comparatively weaker, especially for ADV and PRO.

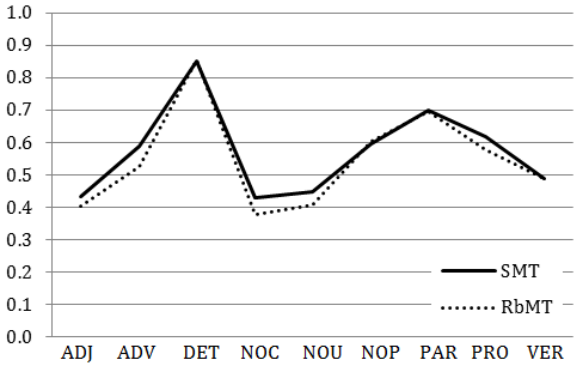

Figure 13: DE→EN results (overall).

With regard to the translation quality ranks, RbMT receives distinctly lower scores for ADV, NOP, PAR and PRO when considering the rank 1 translations (Figure 14). On the other hand, the performance of SMT is particularly bad for adjectives (20% lower than RbMT), thus pointing to a clear quality barrier within the better translations. RbMT also obtains a better evaluation than SMT for DET, where RbMT translates correctly 97.6% of them.

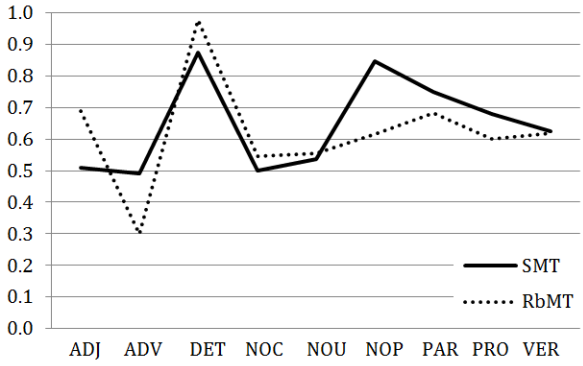

Figure 14: DE→EN results for rank 1.

As far as rank 2 translations are concerned (Figure 15), the performance of SMT and RbMT is very similar across all the checkpoints: some are handled slightly better by SMT (e.g. ADJ, NOU and PRO), while in particular for NOP the score of RbMT is higher.

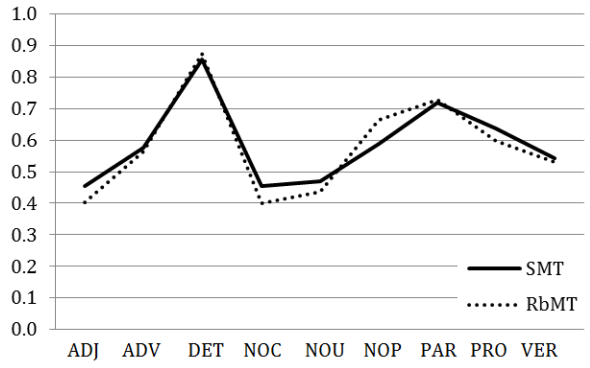

Figure 15: DE→EN results for rank 2.

Finally, for rank 3 translations (Figure 16) the performance tends to be equivalent again for most checkpoints, but RbMT struggles more with ADV, NOC and NOU. On the other hand, for these low-quality translations SMT seems to find more serious barriers in the translation of VER, for which RbMT receives a 5% higher score.

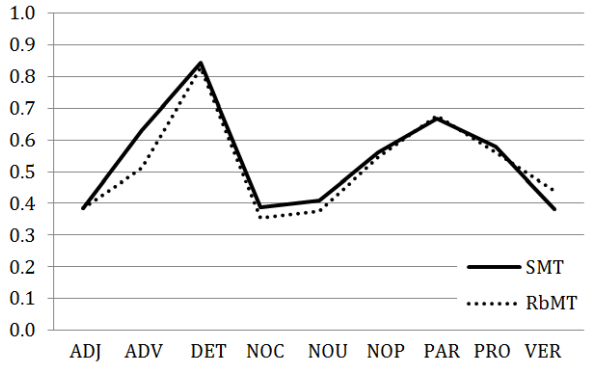

Figure 16: DE→EN results for rank 3.

3.3 Analysis

3.3.1. Correlations between Human Ratings and DELiC4MT

It should be noted that across all of the translation directions and MT system types, 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 those subsets of the data, thus making it particularly difficult to draw reliable generalisations in such circumstances, due to data sparseness problems. In Tables 2 and 3 we thus provide the number of instances of each checkpoint detected on the source/input side of the various data subsets (Table 2 shows the EN--ES language pair, and Table 3 presents the data for EN--DE), to help put in perspective the DELiC4MT scores and our findings in terms of translation quality barriers across the three ranks. For both language pairs there are much higher numbers of detected checkpoints for the rank 2 and rank 3 quality bands. When looking at SMT, RbMT and HMT alike, we can therefore be more confident in the analysis of our findings for near-miss and poor translations, whereas particular caution must be exercised when interpreting the results of rank 1 (i.e. good) translations.

		SMT		RbMT		HMT	
		ES>EN	EN>ES	ES>EN	EN>ES	ES>EN	EN>ES
RANK1	ADJ	53	59	27	44	-	71
	ADV	24	69	17	55	-	82
	DET	119	68	67	37	-	66
	NOC	193	197	100	97	-	183
	NOU	254	304	124	150	-	258
	NOP	61	107	24	53	-	75
	PAR	134	110	76	67	-	100
	PRO	41	56	29	48	-	70
	VER	177	193	157	102	-	198
RANK2	ADJ	196	492	196	496	-	448
	ADV	119	443	109	443	-	394
	DET	335	569	327	575	-	527
	NOC	639	1723	662	1772	-	1655
	NOU	853	2508	823	2499	-	2390
	NOP	214	785	161	727	-	735
	PAR	482	1156	459	1153	-	1073
	PRO	125	512	127	510	-	450
	VER	786	1449	687	1515	-	1373
RANK3	ADJ	70	380	93	412	-	433
	ADV	51	302	68	327	-	352
	DET	132	340	188	373	-	393
	NOC	286	1202	354	1275	-	1308
	NOU	371	1649	528	1840	-	1852
	NOP	85	447	174	565	-	544
	PAR	163	729	240	793	-	845
	PRO	72	274	81	289	-	329
	VER	273	1016	388	1069	-	1117

Table 2: Numbers of checkpoint instances detected on the source side for the EN--ES language pair.

		SMT		RbMT		HMT	
		DE>EN	EN>DE	DE>EN	EN>DE	DE>EN	EN>DE
RANK1	ADJ	173	5	45	20	-	38
	ADV	63	12	20	16	-	38
	DET	102	10	42	13	-	29
	NOC	356	40	152	44	-	140
	NOU	396	43	178	56	-	185
	NOP	39	3	26	12	-	45
	PAR	152	5	47	16	-	48
	PRO	87	16	35	18	-	30
	VER	179	40	89	42	-	75
RANK2	ADJ	591	180	587	360	-	419
	ADV	203	156	195	252	-	275
	DET	479	191	438	373	-	434
	NOC	2023	593	1655	1201	-	1312
	NOU	2294	905	1921	1878	-	1995
	NOP	270	312	264	677	-	683
	PAR	673	290	581	678	-	743
	PRO	298	167	278	317	-	336
	VER	680	493	640	896	-	1086
RANK3	ADJ	536	708	673	512	-	436
	ADV	203	488	254	388	-	343
	DET	403	749	504	560	-	487
	NOC	1737	2551	2299	1930	-	1732
	NOU	1995	3771	2574	2766	-	2539
	NOP	258	1220	275	836	-	807
	PAR	581	1435	778	1034	-	939
	PRO	251	509	323	355	-	326
	VER	578	1801	708	1396	-	1173

Table 3: Numbers of checkpoint instances detected on the source side for the EN--DE language pair.

To ascertain the correlation between the DELiC4MT scores and the human evaluations, we calculated Pearson's *r* values for the DELiC4MT scores and the human ratings. This correlation concerns individual PoS-based checkpoints and the human quality ranking of MT output of whole sentences for which the relevant checkpoint is detected by DELiC4MT on the source side. Noise introduced by the PoS tagger and the word aligner might have an impact on these results, however our previous work (Naskar et al., 2013) shows rather conclusively that the noise introduced by state-of-the-art PoS taggers and word aligners does not have a noticeable impact on DELiC4MT results. Normalising these results to positive values (since the rating is an inverse scale) gives the results shown in Figure 17.

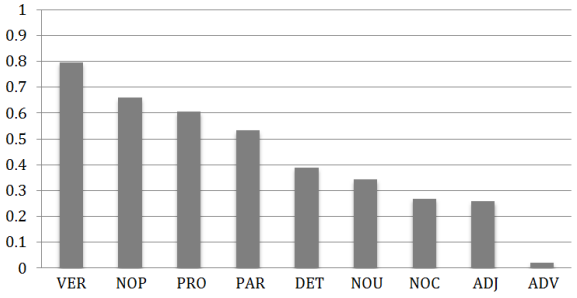

Figure 17: Pearson's *r* correlation for DELiC4MT scores and human quality ratings.

As seen in Figure 17, quality levels correlated with all of the PoS-based linguistic checkpoints, but many of the correlations are weak at best (and virtually nonexistent for ADV). Assuming significance over the approximately 4,400 translations examined, it seems that the VER, NOP, and PRO checkpoints stand out as the best predictors of human quality assessment. PAR and DET also exhibit reasonable correlation.

Figures 18 and 19 show the score clustering at each quality level along with the trend lines for each checkpoint. In an examination of human error annotation (Burchardt et al., 2014:39-4l) we found that PAR and DET were among the most problematic PoS classes for MT in general, and especially for RbMT. An examination of the respective scores and trend lines as shown in Figures 18 and 19 reveals that MT systems were generally quite reliable in producing high DELiC4MT scores for these items, with among the highest scores for these checkpoints across all quality bands. While they differentiate between the bands, due to the low standard deviation evident in each cluster, the differentiation is also quite small.

Furthermore, the use of particles and determiners is among the most variant of grammatical features across languages, and accurately transferring these items between languages is quite likely to be error prone. Accordingly, although the MT systems were consistently good in matching these two checkpoints, an examination of human error markup shows that a high DELiC4MT score for these two checkpoints is not necessarily a good predictor of overall quality (approximately 15% of the errors annotated in the corpus described in Burchardt et al. (2014) had to do with so-called "function words" such as particles and determiners), unlike VER, NOP, and PRO, where a high degree of correspondence between presence of these checkpoints in both source and target would generally be a good predictor of accuracy and quality. Thus a comparison of these results with the findings of the human annotation task described in Burchardt et al. (2014) shows that automatic analysis, such as this study carries out, can contribute to a better understanding of human annotation and vice-versa.

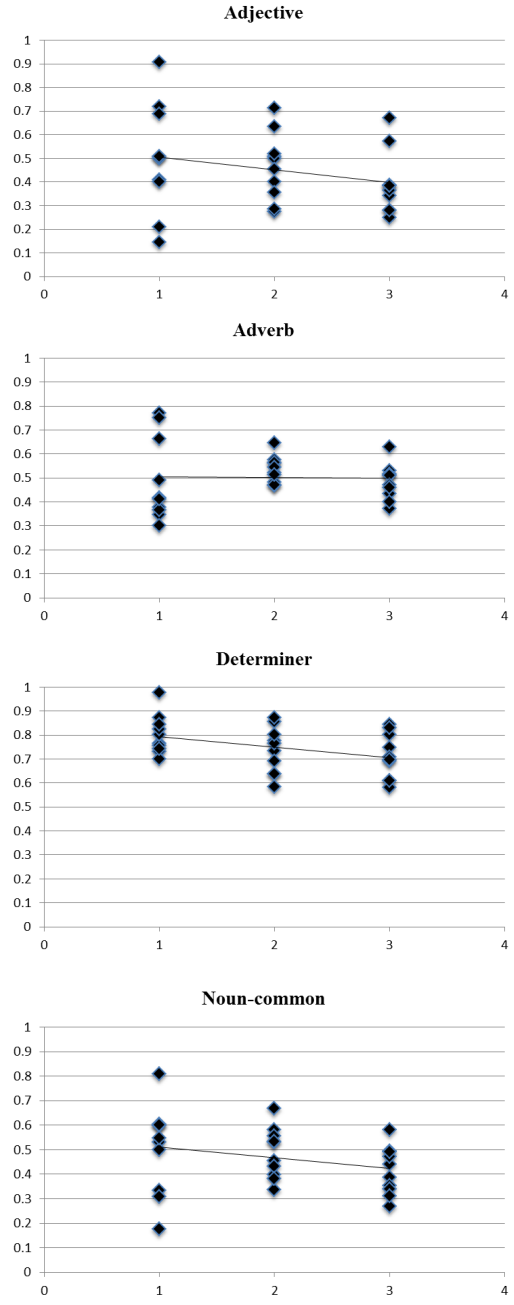

Figure 18: DELiC4MT scores by quality band with trend lines for ADJ, ADV, DET, NOC.

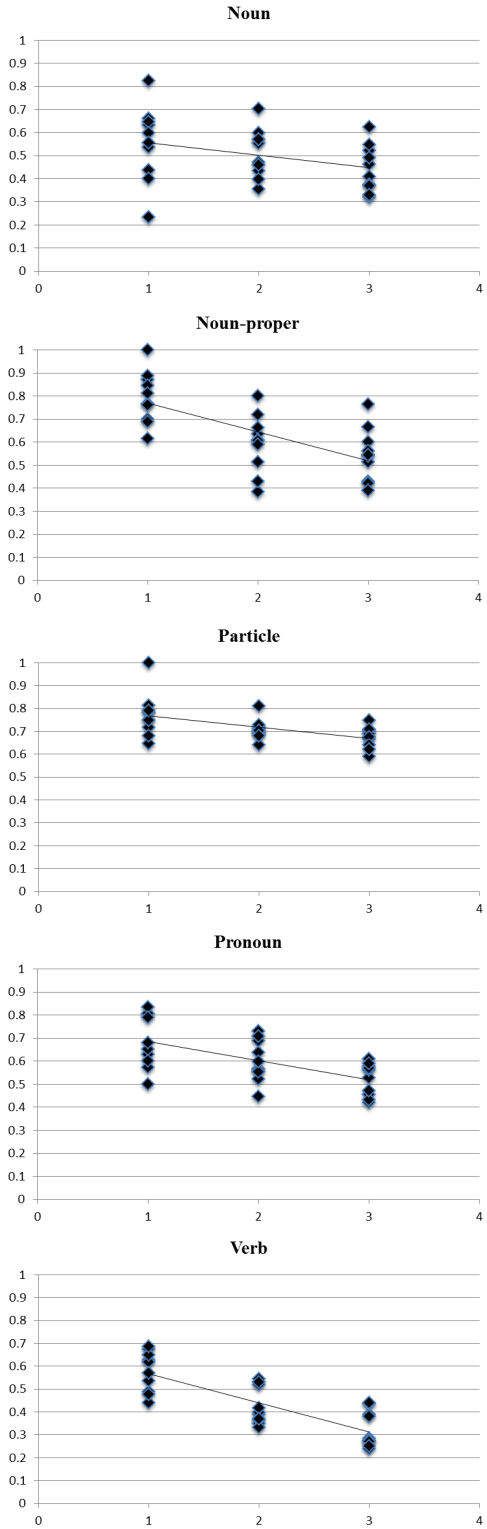

Figure 19: DELiC4MT scores by quality band with trend lines for NOU, NOP, PAR, PRO, VER.

3.3.2. DELiC4MT Sample Output

To clarify the mechanism of the analysis performed by DELiC4MT on the PoS-based linguistic checkpoints, here we show three sample outputs on the segment level (all of them for the language direction Spanish to English and for the VER checkpoint on output produced by the RbMT system). 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. As explained in more detail in Section 2.1, DELiC4MT detects the relevant PoS-based checkpoint on the source side, matches it with the aligned word in the MT output/hypothesis, and checks this against the corresponding word in the reference translation.

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 shows a correct translation, scored successfully by DELiC4MT. The Spanish form "es" (3rd person of the present tense of the verb "ser", i.e. 'to be' in Spanish) is correctly translated to its equivalent in English, "is", matching the aligned reference translation.

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 shows a verb translated literally (and incorrectly): the source "hizo" (3rd person of the past tense of the verb "hacer", i.e. 'to make/to do' in Spanish) would normally correspond to "made/did" in English; however, in the expression "hacer un regalo" it corresponds to "give a present". The diagnostic evaluation tool correctly identifies this as a mistake, i.e. it detects a specific case contributing to translation quality barriers.

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 shows a correct translation which DELiC4MT fails to assess positively: the verb "respira" (3rd person of the present tense of the verb "respirar", i.e. 'to breathe' in Spanish) is correctly translated as "breathing" in English; 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.

4. Conclusions and Future Work

This paper has explored the joint use of automatic diagnostic evaluation and human quality rankings to identify source-side linguistic phenomena that cause quality barriers in MT, looking at the two bidirectional language pairs EN↔ES and EN↔DE. We have evaluated output sentences produced by three types of MT systems

(statistical, rule-based and hybrid) belonging to different quality ranks (perfect, near-miss and poor translations), as classified by human annotators. The evaluation has been performed on a set of 9 PoS-based linguistic checkpoints with DELiC4MT, thus allowing us to draw conclusions on the quality barriers encountered by the different MT systems on a range of linguistic phenomena, for all three quality ranks across the four translation combinations.

On the basis of this evaluation, the paper has analysed the correlation between the scores obtained for each of these source-side linguistic phenomena and the human quality ratings, thus assessing the extent to which these phenomena can be used to predict human quality evaluation. Considering all the MT system types evaluated together, it turns out that the best predictors are VER (r=0.795), NOP (r=0.658) and PRO (r=0.604), while the worst one is by far ADV (r=0.02).

Regarding future work, taking into account the limitations of the current study (the small amount of data and somewhat limited translation combinations), we would like to confirm the findings reported here by performing experiments on larger data sets, including a more varied and extended set of language pairs for a wider collection of linguistic checkpoints. We are also planning to explore the use of diagnostic MT evaluation to analyse the errors identified by the Multidimensional Quality Metric (MQM) (Lommel and Uszkoreit, 2013). The MQM is a new paradigm for translation quality assessment, in which errors are categorised according to a hierarchy of issue types. By using DELiC4MT with a variety of suitable linguistic checkpoints to analyse translations annotated with the MQM, we intend to investigate which source-side linguistic phenomena cause the various MQM error types, as further indicators of translation quality barriers.

5. Acknowledgements

The work presented here has been conducted as part of the QTLaunchPad project, which has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 296347. Thanks are due to Aljoscha Burchardt, Maja Popović, Kim Harris and Lucia Specia for facilitating the annotation of the data used for this study and for interesting discussions that led to the work presented here, for which however the authors are solely responsible.

6. References

Balyan, R., Naskar, S.K., Toral, A. and Chatterjee, N. (2012). A Diagnostic Evaluation Approach Targeting MT Systems for Indian Languages. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), COLING 2012*. Mumbai, India, December 2012, pp. 61--72.

Balyan, R., Naskar, S.K., Toral, A. and Chatterjee, N. (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)*. Samos, Greece. 2013. LNCS 7817. Berlin: Springer, pp. 285--296.

Burchardt, A., Gaspari, F., Lommel, A., Popović, M., and Toral, A. (2014). *Barriers for High-Quality Machine Translation*. QTLaunchPad Deliverable 1.3.1. Available from www.qt21.eu/launchpad/system/files/deliverables/QTLP-Deliverable-1_3_1.pdf (accessed 10 February 2014).

Lommel, A. and Uszkoreit, H. (2013). Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment. Paper presented at *Localization World*, 12-14 June 2013, London, United Kingdom.

Naskar, S.K., Toral, A., Gaspari, F. and Way, A. (2011). A Framework for Diagnostic Evaluation of MT Based on Linguistic Checkpoints. In *Proceedings of Machine Translation Summit XIII*. Xiamen, China, 19-23 September 2011, pp. 529--536.

Naskar, S.K., Toral, A., Gaspari, F. and Groves, D. (2013). Meta-Evaluation of a Diagnostic Quality Metric for Machine Translation. In K. Sima'an, M.L. Forcada, D. Grasmick, H. Depraetere and A. Way (Eds.), *Proceedings of the XIV Machine Translation Summit*. Nice, France, 2-6 September 2013. Allschwil: The European Association for Machine Translation, pp. 135--142.

Och, F.J., and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics,* volume 29, number 1, pp. 19--51.

Padró, L. and Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference* . ELRA. Istanbul, Turkey. 21-27 May 2012, pp. 2473--2479.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Philadelphia, July 2002, pp. 311--318.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. P*roceedings of the ACL SIGDAT-Workshop.* Dublin, Ireland, pp. 47--50.

Toral, A., Naskar, S.K., Gaspari, F. and Groves, D. (2012). DELiC4MT: A Tool for Diagnostic MT Evaluation over User-defined Linguistic Phenomena. *The Prague Bulletin of Mathematical Linguistics*, 98(1), pp. 121--131.

Toral, A., Naskar, S.K., Vreeke, J., Gaspari, F. and Groves, D. (2013). A Web Application for the Diagnostic Evaluation of Machine Translation over Specific Linguistic Phenomena. In C. Dyer and D. Higgins (Eds.), *Proceedings of the 2013 NAACL HLTConference - Demonstration Session.* Atlanta, GA, USA. 10-12 June 2013. Stroudsburg, PA: Association for Computational Linguistics, pp. 20--23.