

**A Comparative Evaluation
of
Adequacy
in
Six Machine Translation Systems**

Final Report

**Human and Automated Evaluations
of 6 MT Systems**

**Bogdan Babych, Debbie Elliott, Tony Hartley
Centre for Translation Studies, University of Leeds
www.leeds.ac.uk/cts**

Version: Final January 2004

Contents

Contents	i
Tables and figures	ii
1 Scope of the report.....	1
1.1 Objectives.....	1
1.2 Coverage.....	1
2 Summary of the results	4
2.1 Overall acceptability of MT output	4
2.2 Verdict on individual systems	4
2.2.1 System rankings by performance.....	4
2.2.2 System rankings by target language.....	6
2.2.3 Gaps in coverage.....	7
2.3 Calibration of automated scores by human scores	7
3 Set-up of experiment.....	9
3.1 Design of the MT systems.....	9
3.1.1 Statistics-based systems.....	9
3.1.2 Linguistic knowledge-based systems.....	9
3.2 Source and reference texts.....	10
3.2.1 Linguistic features of the whitepaper.....	11
3.2.2 Linguistic features of the emails.....	11
3.3 Dictionary adaptation.....	11
4 Human evaluations	12
4.1 Evaluation procedure	12
4.1.1 Evaluators	12
4.1.2 Evaluations and materials.....	12
4.1.3 Adequacy evaluation task.....	12
4.2 Results.....	14
4.2.1 Acceptability threshold.....	14
4.2.2 By system	15
4.2.3 By target language.....	17
5 Automated evaluations	29
5.1 Commentary on the automated metrics	29
5.1.1 BLEU.....	29
5.1.2 LTV	29
5.2 System rankings resulting from BLEU and LTV	30
5.3 BLEU scores.....	37
5.4 LTV scores	44
6 Calibration of automated scores.....	50
6.1 Purpose.....	50
6.2 Derivation of the coefficients	50
6.3 Application of the coefficients	51

Tables and figures

Table 1: Identifier codes for MT systems.....	2
Table 2: Language pairs subject too human evaluation	3
Table 3: Scale of acceptability weightings	14
Table 4: Rankings and acceptability according to human evaluations	16
Table 5: TL English – system rankings and acceptability.....	18
Table 6: TL French – system rankings and acceptability.....	20
Table 7: TL German – system rankings and acceptability	22
Table 8: TL Spanish – system rankings and acceptability	25
Table 9: TL Italian – system rankings and acceptability.....	27
Table 10: TL Portuguese – system rankings and acceptability	28
Table 11: TL English – system rankings whitepaper.....	31
Table 12: TL English – system rankings emails.....	32
Table 13: TL French – system rankings whitepaper and emails	33
Table 14: TL German – system rankings whitepaper and emails.....	34
Table 15: TL Spanish – system rankings whitepaper and emails.....	35
Table 16: TL Italian – system rankings whitepaper and emails	36
Table 17: TL Portuguese – system rankings whitepaper and emails	36
Table 18: Calibration table for automated scores	54
Figure 1: TL English – human evaluation scores for whitepaper and emails	17
Figure 2: TL French – human evaluation scores for whitepaper and emails	19
Figure 3: TL German – human evaluation scores for whitepaper and emails.....	21
Figure 4: TL Spanish – human evaluation scores for whitepaper and emails	24
Figure 5: TL Italian – human evaluation scores for whitepaper and emails	26
Figure 6: TL Portuguese – human evaluation scores for whitepaper and emails	28
Figure 7: TL English – BLEU scores for whitepaper and emails.....	38
Figure 8: TL French – BLEU scores for whitepaper and emails.....	39
Figure 9: TL German – BLEU scores for whitepaper and emails	40
Figure 10: TL Spanish – BLEU scores for whitepaper and emails	41
Figure 11: TL Italian – BLEU scores for whitepaper and emails.....	42
Figure 12: TL Portuguese – BLEU scores for whitepaper and emails.....	43
Figure 13: TL English – LTV scores for whitepaper and emails	44
Figure 14: TL French – LTV scores for whitepaper and emails.....	45
Figure 15: TL German – LTV scores for whitepaper and emails.....	46
Figure 16: TL Spanish – LTV scores for whitepaper and emails	47
Figure 17: TL Italian – LTV scores for whitepaper and emails	48
Figure 18: TL Portuguese – LTV scores for whitepaper and emails	49

1 Scope of the report

1.1 Objectives

The objectives of the present evaluation exercise are threefold:

- assess the relative quality of the output of a number of machine translation (MT) systems identified by Translution as potential ‘engines’ for its proposed translation service, taking emails and an EU whitepaper as representative inputs
- establish an acceptability threshold below which the quality of translation is deemed inadequate for the purposes of Translution’s future clients
- provide a benchmark for evaluating quickly and at low cost the performance of later versions of the MT systems evaluated here or of other systems which Translution might choose to consider in the future.

1.2 Coverage

Like the interim report delivered in August 2004, the present report focuses on a single attribute of translation quality:

- adequacy – the extent to which the information content of the original, source text is judged to be preserved in the translation produced by the MT system.

It was agreed not to evaluate the fluency, or naturalness, of the translations (which did figure in the comparative evaluation of 2 MT systems conducted in July 2004), this for two reasons.

- The notion of ‘fluency’ is subject to greater individual interpretation for the relatively new *genre* of emails than it is for the established *genre* of semi-technical reports represented by the EU whitepaper. Emails are not necessarily expected to be fluent in the same sense as reports.
- Previous MT research (in US, Japan and Europe) has consistently shown a high correlation between judgments of fluency and judgments of adequacy, which means a ranking of MT systems based on one attribute tends to be identical to a ranking based on the other.

This report extends the August 2004 interim report in three ways.

- It includes four additional MT systems, and thus additional language directions.
- The outputs are evaluated not only with automated measures but also by human evaluators.
- The automated evaluation uses not only the BLEU metric, but also the LTV metric.

For the language directions English>French and French>English, the dictionaries were adapted according to the guidelines followed in previous dictionary adaptation exercises (see section 3.3).

The systems evaluated are listed in Table 1, which also provides the identifier codes use in all subsequent tables and figures.

ID	System	AutoEval	HumanEval
s01	s01 Rules Based MT system – without dictionary adaptation. Previously evaluated in July 2004 for French <> English	√	
s02	s02 Rules Based MT system – without dictionary adaptation. Previously evaluated in July 2004 for French <> English	√	
s03	s03 Rules Based System – without dictionary adaptation	√	√
s04	s04 Statistical Based System – without dictionary adaptation	√	√
s05	s05 Rules Based System – without dictionary adaptation	√	√
s06	s06 Rules Based System – without dictionary adaptation. Same supplier as s01 but different version	√	√
u01	s01 – with dictionary adaptation	√	
u03	s03 – with dictionary adaptation	√	√
u05	s05– with dictionary adaptation	√	√

Table 1: Identifier codes for MT systems

All systems were included in the automated evaluations, in order to provide the maximum amount of data to inform the calibration objective. However, in order to minimise costs, the human evaluations were confined to those systems which Translution considered in November 2004 to be serious commercial contenders.

Table 2 shows the language directions that figured in the human evaluations. Filled cells indicate dictionary adaptation took place.

Languages	s06	s03	s04	s05
English > French	√	√	√	√
English > German	√			√
English > Spanish	√	√	√	√
English > Italian	√	√		
English > Portuguese	√			√
French > English	√	√	√	√

Languages	s06	s03	s04	s05
French > German	√			
French > Spanish	√			
French > Italian	√	√		
French > Portuguese	√			
German > English	√			√
German > French	√			
German > Spanish				
German > Italian		√		
Italian > English	√	√		
Italian > French	√	√		
Italian > German		√		
Italian > Spanish		√		
Portuguese > English	√			
Portuguese > French	√			
Spanish > English	√	√	√	√
Spanish > French	√			
Spanish > Italian		√		

Table 2: Language pairs subject to human evaluation

2 Summary of the results

This section summarises the most salient observations that emerge from the raw human scores presented in the figures and tables of sections 4.2.2 and 4.2.3.

2.1 Overall acceptability of MT output

The human evaluations show that, for almost all the language directions investigated, MT can provide an acceptable level of service in preserving the accuracy of content in translation for users primarily interested in conveying facts or extracting information from texts.

2.2 Verdict on individual systems

2.2.1 System rankings by performance

For these rankings, scores achieved after dictionary adaptation (u03 and u05) are set aside in the interests of a flat baseline comparison. For the detailed numerical data that underpins these conclusions, consult Table 4.

- **s03**

Where s03 is in competition with other systems, it comes last in all cases except English>Spanish for emails (3/4). Elsewhere, its rare competitive performances are Italian>English and Italian>French for emails. However, when performance on the whitepaper is also taken into account for these same two language directions, s01/s06 is better..

With respect to the acceptability threshold, it generally falls below standard (9/12 cases for emails, 10/12 cases for whitepaper).

Dictionary adaptation allows s03 to redeem a sub-standard performance and achieve the threshold in both cases tested (French>English and English>French). The improvement is, indeed, spectacular – up to 56%. But this should be seen as a sign of the serious deficiency of the s03 dictionaries in their current state. Moreover, our experience of dictionary adaptation showed that s03 required about double the effort compared to s02, s05 and s01/s06, all of which were approximately equal.

- **s04**

Human evaluators rank s04 in last or next-to-last position on both emails and whitepaper, across all language directions. This is in contrast to the automated scores, which always ranked it first on the whitepaper, whether into or out of English. We have previously highlighted this tendency of the automated metrics to overestimate the quality of statistics-based MT systems; this evidence underlines the need for calibration against human scores.

With respect to the acceptability threshold, it generally falls below standard (3/4 cases for both emails and whitepaper).

Moreover, it is always poorer for emails. This performance reflects the constituency of the corpus on which the system was trained. This sensitivity is both a strength and a weakness of statistical machine translation – good in domains for which they are tuned, less good in general purpose scenarios across a range of domains. While it is in principle possible to re-train s04 on data that includes email translations (assuming the availability of such a corpus), it is not clear what negative effect this might have on performance on other text types.

- **s05**

Overall, s05 is the best-performing system, especially on emails, where it is ranked first in all cases but one (6/7). While its performance on the whitepaper is better – in keeping with the general trend – on this text type it places first in fewer than half the cases (3/7), being headed by s01/s06 in all other instances.

With respect to the acceptability threshold, s05 fails narrowly in three cases. Its weakest directions are English>German/Italian/Spanish.

However, the evidence is that all of these failures would be redeemed by dictionary adaptation. Adapting the dictionaries for French>English and English>French yielded a clear improvement, even though the performance using just the default dictionaries had already exceeded the threshold.

- **s06**

s06 offers the widest range of language directions. Where it is in competition with other systems, it usually places second to s05 on emails and first on the whitepaper. Its weakest placing is third, in English>Spanish on both text types.

With respect to the acceptability threshold, s06 fails narrowly in 50% of the cases on emails (9/18) and 17% of the cases on the whitepaper ((3/18). Its weakest languages are Portuguese and Italian, with the notable exception of French>Italian. Its strongest directions are German>English, English<>French and (for the whitepaper) French<>Portuguese.

However, the evidence from our experience with s01 is that all of these failures would be redeemed by dictionary adaptation.

According to the automated scores, there was a very marginal regression of s06 compared with s01 for the following language directions: English>Spanish (emails), Spanish>English, English>Portuguese and Italian>English (whitepaper). Moreover, s06 failed on one French>Portuguese and one German>French email segment, which were successfully translated by s01.

2.2.2 System rankings by target language

Once again, baseline scores (without dictionary adaptation) are used. For the detailed numerical data that underpins these conclusions, consult the tables and figures in section 4.2.3.

- **TL English**

There is generally acceptable – even highly acceptable – translation into English from all source languages and for both text types, and by more than system

•
SLs French, German and Spanish are highly satisfactory with s05 and s01/s06

•
For SL Italian, on emails s03 outperforms s01/s06 (which is just below the threshold), but over both text types s01/s06 appears the safer option.

For SL Portuguese, only s01/s06 is available. While it fails on emails, it is satisfactory on the whitepaper.

- **TL French**

For SLs German and English the acceptability is good. s01/s06 offers the best cover.

For SL Italian, s03 is equal to s01/s06 on emails (where both fail very marginally) but inferior on the whitepaper. s01/s06 offers the better cover.

For SLs Spanish and Portuguese, s01/s06 is the only available system. While it fails marginally on emails from both SLs, it is highly satisfactory on the whitepaper, again from both SLs.

- **TL German**

For SL French, s01/s06 performs satisfactorily across both text types.

For SL English, s05 and s01/s06 are very close. Both perform less well on the whitepaper, where they fail marginally.

For SL Italian, s03 is the only available system. It fails on both text types, doing worst on the whitepaper. The quality is probably redeemable through dictionary adaptation, but the effort will be considerable.

- **TL Spanish**

This proves to be a very challenging TL.

For SL English, the only immediately satisfactory performance is by s05 on the whitepaper. s03 is very slightly ahead of s01/s06, but s01/s06 nevertheless holds out the better prospect of the two, since dictionary adaptation is likely to

be less onerous.

For SL French, s01/s06 is the only available system and performs well – outstandingly so on emails.

For SL Italian, s03 is the only contender and fails marginally on both text types.

- **TL Italian**

For SLs French and Spanish, the quality is good. s01/s06 is exceptional from French on the whitepaper, while from Spanish s03 (the only contender) is satisfactory on both text types.

For SL English, s01/s06 offers the better quality, although it fails marginally on emails.

For SL German, s03 is the only contender. Its performance on both text types is very poor and probably irredeemable through dictionary adaptation alone.

- **TL Portuguese**

s01/s06 is good in Portuguese >Italian on the whitepaper.

2.2.3 Gaps in coverage

No satisfactory system is currently available for the following language directions offered only by s03 and where the quality of translation is currently seriously deficient.

- Italian>German
- Italian>Spanish
- German>Italian

2.3 Calibration of automated scores by human scores

We have produced a first version of a tool for benchmarking quickly and at low cost the performance of later versions of the MT systems evaluated here or of other systems which Translution might choose to consider in the future.

This tool takes the form of a table of coefficients that vary according to the parameters of text type (email or whitepaper), source language and target language. The coefficients are derived from observed correlations between the human evaluation scores and the automated evaluation scores.

An updated or new system can be evaluated relative to the systems discussed in this report and to the quality threshold in four simple steps:

- generate the translation of the email and/or whitepaper
- calculate the BLEU and/or LTV scores

- apply the coefficients to the automated scores
- check the result against the threshold value.

The optimisation of the coefficients presented in section 6 is the focus of ongoing research.

3 Set-up of experiment

This section describes the principal resources engaged in the experiment – the MT systems themselves, the texts for translation, and the dictionaries.

3.1 Design of the MT systems

3.1.1 Statistics-based systems

s04 is a statistics-based MT system. Such systems are built on the basis of large bilingual collections (ideally several million words) of texts and their translations, aligned sentence by sentence.

From this data, the systems are ‘trained’ by machine-learning techniques to acquire two models: the translation model, which gives the likelihood of some segment in the source language being translated by some segment in the target language; and the language model, which gives the likelihood of some segment in the target language being followed by some other segment of the target language.

These two statistical models are used jointly to generate translations of ‘new’ texts which were not part of the training data. The more closely the constituents of the new texts resemble the training data, the better the translation; conversely, words and expressions in the new text which are not present in the training data will disrupt the translation process.

It is important to understand that statistics-based systems do not have a distinct dictionary module.

3.1.2 Linguistic knowledge-based systems

All three other systems are ‘linguistic knowledge-based systems’, relying on rules which attempt to represent the grammars of the source and target languages.

In addition to their grammar rules, the systems are supplied with dictionaries of varying sizes and offer the user the possibility of creating new, customised dictionaries to fill gaps or to override existing translations of words or expressions with a preferred variant.

While s01/s06 and s05 both have very large dictionaries, the lexical resources of s03 are either smaller or less well tuned to the texts translated.

Architecturally, s01/s06 and s05 are known to be similar. Both are based on the so-called ‘transfer’ model, which decomposes the translation process into three main stages:

- analysis – This phase is concerned with establishing an unambiguous ‘understanding’ of the source text; it is independent of any considerations of the target language.
- transfer – This phase is concerned with mapping the more or less abstract representation of the source text into a corresponding representation of the target text; it is inherently bilingual and contrastive.

- synthesis – This phase is concerned with transforming the abstract representation of the target text into a string of grammatically and semantically correct words and sentences; it is independent of any considerations of the source language.

This modular architecture is intended to ensure the re-usability of the analysis and synthesis modules for a particular language, whatever other language it is paired with.

It is unclear to what extent s03 implements the transfer model or whether it uses a simpler and less flexible direct approach where translation decisions are closely tied to a particular source language-target language pair.

Creating a new dictionary entry is, then, a matter of associating a word or expression in the analysis dictionary (or ‘source lexicon’) with a corresponding word or expression in the synthesis dictionary (or ‘target lexicon’). This association is stored in the ‘transfer lexicon’, which is thus specific to translation from a particular source language into a particular target language.

3.2 Source and reference texts

Two types of document were used to evaluate the performance of the MT systems.

- The first 3,000 words (approximately) of a European Commission whitepaper on safe Internet access. For the purposes of human evaluation, this was divided into 150 text segments. Each segment comprised a complete sentence or heading, with the exception of very long sentences, which were split as appropriate.
- A set of 36 emails (24 business-related and 12 personal). The emails varied in length between 31 and 210 words, the average being 107 words. They were divided into 228 segments.

These texts were supplied by Translution as being representative of the anticipated use of MT by the target business community.

The whitepaper existed in official EU versions in all the languages under consideration. All language versions were checked manually to remove from any one version those few segments that did not have a direct counterpart in all the other versions. Thus a strict parallelism was enforced across all language versions.

Translations of the emails from English into all language versions were provided by Translution. These were checked for parallelism in the same way as the whitepaper.

Thus each text served two functions:

- source text for translation into all available target languages
- reference text against which to check all translations into that target language.

3.2.1 Linguistic features of the whitepaper

This document presented a number of challenges to the MT systems.

- It contained many strings of nouns and names of organisations, policies and legislation, which were not likely to be found in the system dictionaries, e.g. *Organisation Mondiale de la Propriété Intellectuelle* – World Intellectual Property Organisation.
- Some sentences were very long and complex.

3.2.2 Linguistic features of the emails

The emails were expected to pose a different set of problems for the MT systems:

- highly general support verbs: e.g. *get*
- phrasal verbs with multiple interpretations: e.g. *put up*
- abbreviated words: e.g. *ordi - ordinateur* (computer), *t'as - tu as* ('you have')
- onomatopoeic words: e.g. *ouf*
- colloquial or new words: e.g. *beep, yak*
- colloquial phrases akin to speech, which would probably not be found in system dictionaries
- occasional long sentences, and instances of email writers simply forgetting to use full stops
- acronyms used in business, unknown by system dictionaries
- names of places, companies, etc.

3.3 Dictionary adaptation

With s03 and s05, in the directions French>English and English>French, we created new dictionaries to account for missing and mis-translated words.

We first generated translations of all source texts using the dictionaries supplied with the MT systems. We then created a new user dictionary for each system, using the appropriate human-produced reference version of the text as a gold standard for the target language translations.

The user-defined dictionaries were then set to take precedence over the default system dictionaries and a second batch of translation was generated.

4 Human evaluations

We take the human judgments as our gold standard quality benchmark. This section describes the procedure followed in obtaining the scores and presents all the results in figures and tables. These substantiate the conclusions drawn in section 2. The human scores are also used to calibrate the automated scores presented in section 0 by means of the calibration table presented in section 6.1.

4.1 Evaluation procedure

4.1.1 Evaluators

With the collaboration of research partners in Europe, we engaged one evaluation coordinator per target language and a total of 135 evaluators to judge the emails and the whitepaper for adequacy. There were six sets of judges, the majority of whom were postgraduate students and non-linguists.

- English: 45 native speakers of English at the Leeds University, England
- French: 33 native speakers of French in Switzerland
- Italian: 18 native speakers of Italian in Italy
- Spanish: 18 native speakers of Spanish in Spain
- German: 12 native speakers of German in Germany
- Portuguese: 9 native speakers of Portuguese in Portugal

4.1.2 Evaluations and materials

All machine translation output was first collated according to the target language. Segments of output text from different systems and different source languages were then combined automatically to create one file per evaluator, containing all emails and the entire whitepaper document. The number of files created was designed to provide three scores per segment for each language pair and system.

The resulting evaluator packs were sent electronically to the coordinators in the six countries. Coordinators were given precise instructions on how to conduct the evaluations. They were asked to explain the evaluator instructions (shown below) in the target language, and to tell students to work at their own pace and take a break whenever they needed to.

4.1.3 Adequacy evaluation task

All of the machine translations were evaluated for adequacy. The emails were divided into 228 segments (often sentences or headings) and the whitepaper into 150 segments. Each segment was paired with the ‘gold standard’ human translation, referred to as a ‘reference text’.

Each evaluator judged all 378 segments in order, unaware that the candidate texts were translations or that they came from different sources. In this way, each judge would see (and intuitively compare) segments of varying quality instead of output

from one system alone. There were exactly three copies of each pack so that not more than three evaluators saw the same combination of segments.

Each pack was stored as a separate Word document. The segments were presented in the form of a table containing 378 rows with a scoring box adjacent to each segment. Judges worked in a computer cluster and entered their scores electronically, which enabled us to automate their collation and avoid all possibilities of transcription errors. The time taken for students to complete the evaluation varied between 1.5 and 3 hours.

This is the adequacy task set to the evaluators:

For each numbered segment, read the reference text on the left very carefully. Then decide how much of the same information you can find in the candidate text on the right. You should NOT be concerned with grammatical errors or differences in the choice of words.

For each segment, enter your score in the appropriate box in the right hand column. Please DO NOT go back to a segment once you have made a judgement. Save your file regularly during the evaluation. When you have finished, check that every segment in the file has been given a score.

Give each segment of text a score of 5, 4, 3, 2 or 1 where:

5 = All of the content is present

1 = None of the content is present (OR the text completely contradicts the information given on the left hand side).

NB Please bear in mind that this is a running piece of text and that it has been segmented in this way only for the purposes of this experiment.

4.2 Results

4.2.1 Acceptability threshold

We set the threshold of acceptability at the human score of **3.5**. This value was established experimentally.

First, human scores given by individual judges for individual segments were mapped into the scale of weightings given in Table 3.

Human score		Acceptability weighting
5	→	+ 2
4	→	+ 1
3	→	- 1
2	→	- 2
1	→	- 4

Table 3: Scale of acceptability weightings

This scale is weighted against segments that receive bad, poor or average scores. It penalises segments that preserve none, little or some of the content of the source text, while rewarding – but more modestly – segments that preserve most or all of the information. It is a severe rather than a lenient scale.

The resulting score was multiplied by the number of words in each evaluated segment, e.g., if a segment with 15 words received the score -2, the product is -30.

Then we summed all the products for each evaluated system in each translation direction. The intuition is that the acceptability threshold corresponds to a zero sum.

Thus, if the sum is greater than 0, the level of MT quality is ‘acceptable’ (since the majority of segments receive positive marks), otherwise the quality is ‘not acceptable’. Such an ‘acceptability score’ lies in the range +41040 to -82080 for the emails and +27000 to -54000 for the whitepaper. It is given in the following tables in column ‘hPass’.

Acceptability level 0 corresponds to an average score of 3.5 in terms of human evaluation. Therefore, those systems which scored higher than 3.5 can be considered to offer an acceptable quality of output.

In the following tables, any systems not evaluated by human judges are ranked by their LTV adequacy scores. Discrepancies in ranking are colour-coded: **red** figures show that the MT systems are ranked differently for emails and the whitepaper, **blue** figures – that automated scores and human scores ranked systems differently.

Systems that achieve the acceptability threshold (see section 4.2.1) are shown in **green**, while those that fall below the threshold are shown in **blue**.

Experience with the benefits of dictionary adaptation in the present evaluation exercise and the evaluation of s01 and s02 reported in July 2004 allows us to predict that systems achieving in the range 2.5—3.4 without dictionary adaptation will achieve 3.5 or better with dictionary adaptation.

4.2.2 By system

Sys	SL	TL	TxtT	r=	hAve	hPass	TxtT	r=	hAve	hPass
s03	it	de	em		3.184	-10500	wp		2.707	-16770
s03	es	en	em	4	3.294	-6690	wp	4	3.147	-7380
u03	fr	en	em		3.845	9420	wp		4.338	15810
s03	fr	en	em	4	3.423	-2460	wp	4	3.131	-7950
s03	it	en	em	1	3.746	6870	wp	2	2.907	-12240
s03	en	es	em	2	3.149	-11520	wp	2	3.498	-540
s03	it	es	em		3.339	-5520	wp		3.458	-1470
u03	en	fr	em		3.811	8460	wp		4.118	11190
s03	en	fr	em	4	2.854	-20730	wp	4	2.647	-17730
s03	it	fr	em	1	3.45	-3000	wp	2	3.436	-1950
s03	de	it	em		2.551	-30150	wp		2.287	-25470
s03	en	it	em	2	3.282	-7860	wp	2	2.964	-11310
s03	es	it	em		3.705	4590	wp		3.667	2580
s03	fr	it	em	2	3.598	1650	wp	2	3.902	6870
Sys	SL	TL	TxtT	r=	hAve	hPass	TxtT	r=	hAve	hPass
s04	es	en	em	3	3.447	-2190	wp	3	3.927	7830
s04	fr	en	em	3	3.689	5010	wp	3	4.224	13620
s04	en	es	em	4	2.49	-31920	wp	4	3.171	-7110
s04	en	fr	em	3	3.351	-5760	wp	3	3.62	1740
Sys	SL	TL	TxtT	r=	hAve	hPass	TxtT	r=	hAve	hPass
s05	en	de	em	1	3.602	2130	wp	2	3.342	-3780
s05	de	en	em	1	4.383	24900	wp	2	4.071	10890
s05	es	en	em	1	4.151	17610	wp	1	4.242	14040
u05	fr	en	em		4.247	21960	wp		4.589	20550
s05	fr	en	em	1	4.151	17250	wp	2	4.273	14160
s05	en	es	em	1	3.379	-4410	wp	1	3.696	3240
u05	en	fr	em		3.846	9150	wp		4.298	14700
s05	en	fr	em	2	3.649	3840	wp	2	3.902	7170
s05	en	pt	em	1	3.409	-3930	wp	1	3.771	4710

Sys	SL	TL	TxtT	r=	hAve	hPass	TxtT	r=	hAve	hPass
s06	en	de	em	2	3.503	-1020	wp	1	3.469	-900
s06	fr	de	em		3.665	3690	wp		3.818	5670
s06	de	en	em	2	4.194	19500	wp	1	4.153	11790
s06	es	en	em	2	3.902	11310	wp	2	4.018	9480
s06	fr	en	em	2	4.08	16710	wp	1	4.347	16170
s06	it	en	em	2	3.25	-8010	wp	1	3.971	9000
s06	pt	en	em		3.124	-11730	wp		3.711	3900
s06	en	es	em	3	3.126	-12090	wp	3	3.46	-1260
s06	fr	es	em		3.618	3060	wp		4.456	18150
s06	de	fr	em		3.654	3480	wp		3.882	6660
s06	en	fr	em	1	3.974	13620	wp	1	3.924	8010
s06	es	fr	em		3.377	-4800	wp		4.562	19920
s06	it	fr	em	2	3.446	-2940	wp	1	3.88	6900
s06	pt	fr	em		3.303	-6690	wp		4.204	12900
s06	en	it	em	1	3.333	-6270	wp	1	3.611	1410
s06	fr	it	em	1	3.743	6240	wp	1	4.5	18300
s06	en	pt	em	2	3.114	-13080	wp	2	3.196	-6450
s06	fr	pt	em		3.377	-4620	wp		4.262	14460

Table 4: Rankings and acceptability according to human evaluations

4.2.3 By target language

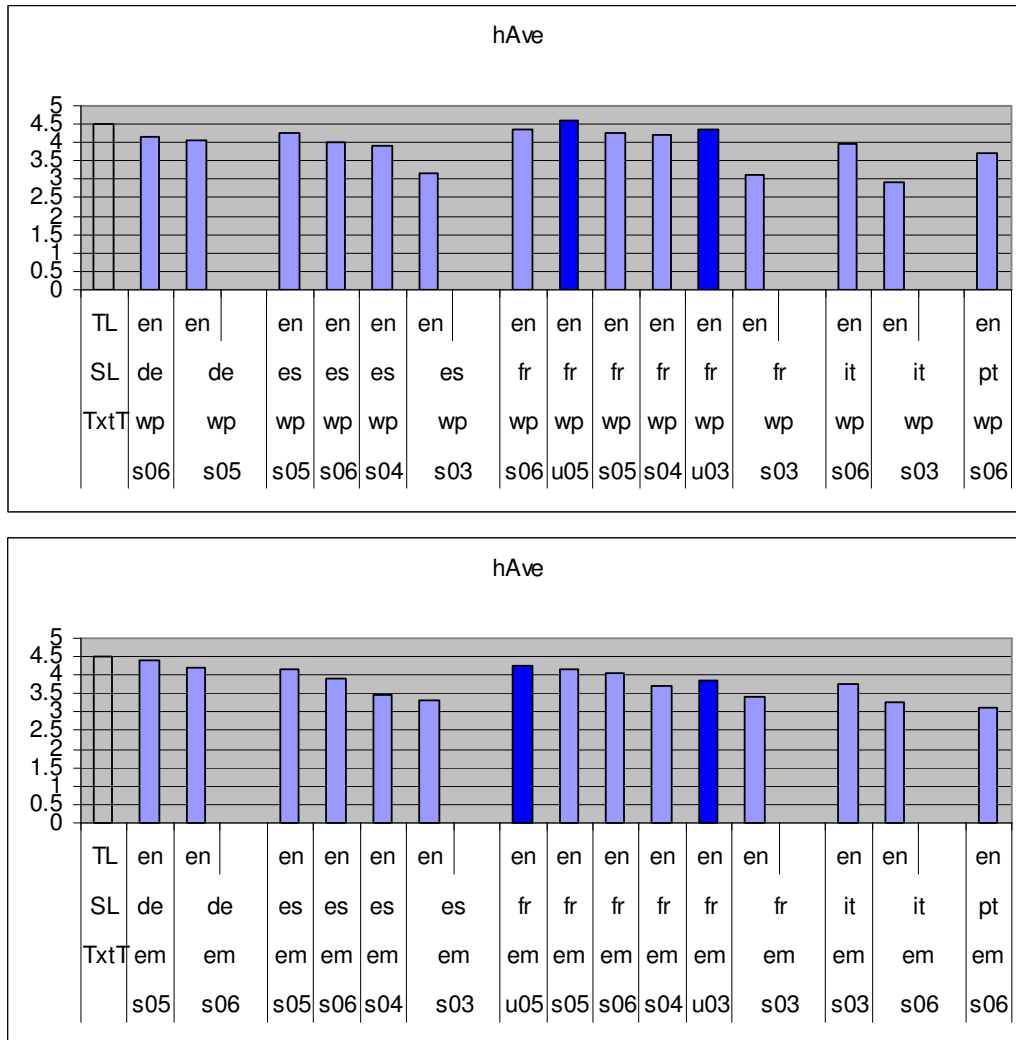


Figure 1: TL English – human evaluation scores for whitepaper and emails

Sys	TxtT	SL	TL	r=	hAve	hPass	r=	Bleu	cBlue	r=	LTVade	cLTV
s06	wp	de	en	1	4.153	11790	1	0.1374		1	0.232	
s05	wp	de	en	2	4.071	10890	2	0.1258		2	0.2055	
s05	wp	es	en	1	4.242	14040	2	0.1702		2	0.2721	
s06	wp	es	en	2	4.018	9480	3	0.1585		3	0.2528	
s04	wp	es	en	3	3.927	7830	1	0.2237		1	0.3308	
s03	wp	es	en	4	3.147	-7380	4	0.1278		4	0.2231	
corr									0.5591			0.54755
s06	wp	fr	en	1	4.347	16170	2	0.2242		2	0.3091	
u05	wp	fr	en		4.589	20550		0.3354			0.433	
s05	wp	fr	en	2	4.273	14160	3	0.2002		3	0.2565	
s04	wp	fr	en	3	4.224	13620	1	0.2746		1	0.3674	
u03	wp	fr	en		4.338	15810		0.237			0.3636	
s03	wp	fr	en	4	3.131	-7950	4	0.1253		4	0.2026	
corr									0.8393			0.7957
s06	wp	it	en	1	3.971	9000	1	0.1296		1	0.2339	
s03	wp	it	en	2	2.907	-12240	2	0.0974		2	0.2024	
s06	wp	pt	en		3.711	3900		0.1216			0.2256	
corrA									0.7086			0.67425
Sys	TxtT	SL	TL	r=	hAve	hPass	r=	Bleu	cBlue	r=	LTVade	cLTV
s05	em	de	en	1	4.383	24900	1	0.3207		1	0.4213	
s06	em	de	en	2	4.194	19500	2	0.2951		2	0.4005	
s05	em	es	en	1	4.151	17610	1	0.218		1	0.3473	
s06	em	es	en	2	3.902	11310	2	0.1959		2	0.3196	
s04	em	es	en	3	3.447	-2190	3	0.1712		3	0.2612	
s03	em	es	en	4	3.294	-6690	4	0.1432		4	0.2518	
corr									0.97826			0.997148
u05	em	fr	en		4.247	21960		0.3339			0.4446	
s05	em	fr	en	1	4.151	17250	2	0.2475		2	0.3513	
s06	em	fr	en	2	4.08	16710	1	0.2862		1	0.392	
s04	em	fr	en	3	3.689	5010	3	0.2259		3	0.3294	
u03	em	fr	en		3.845	9420		0.2659			0.388	
s03	em	fr	en	4	3.423	-2460	4	0.174		4	0.2982	
corr									0.88387			0.824693
s03	em	it	en	1	3.746	6870	2	0.132		2	0.2716	
s06	em	it	en	2	3.25	-8010	1	0.1746		1	0.2856	
s06	em	pt	en		3.124	-11730		0.2051			0.3075	
corrA									0.76986			0.821518

Table 5: TL English – system rankings and acceptability

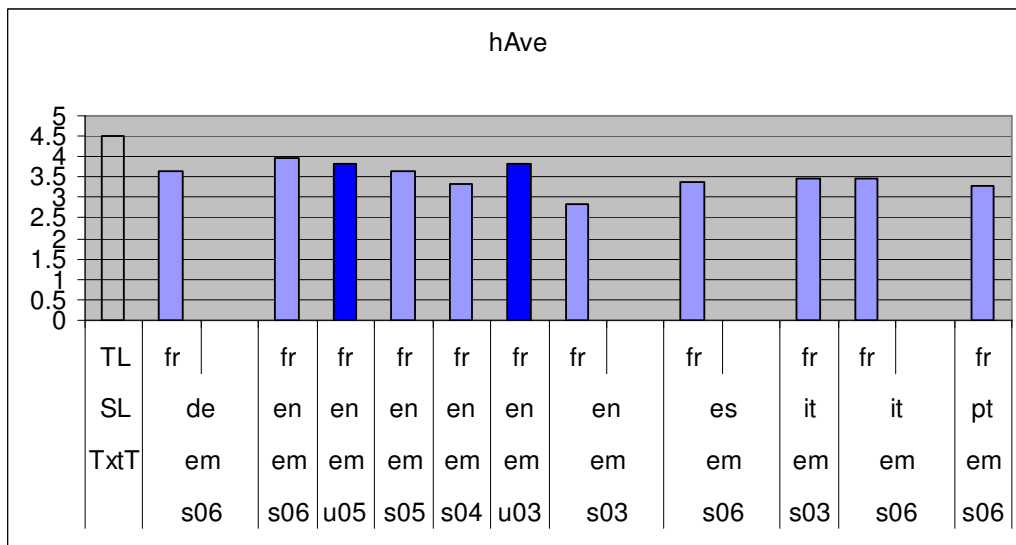
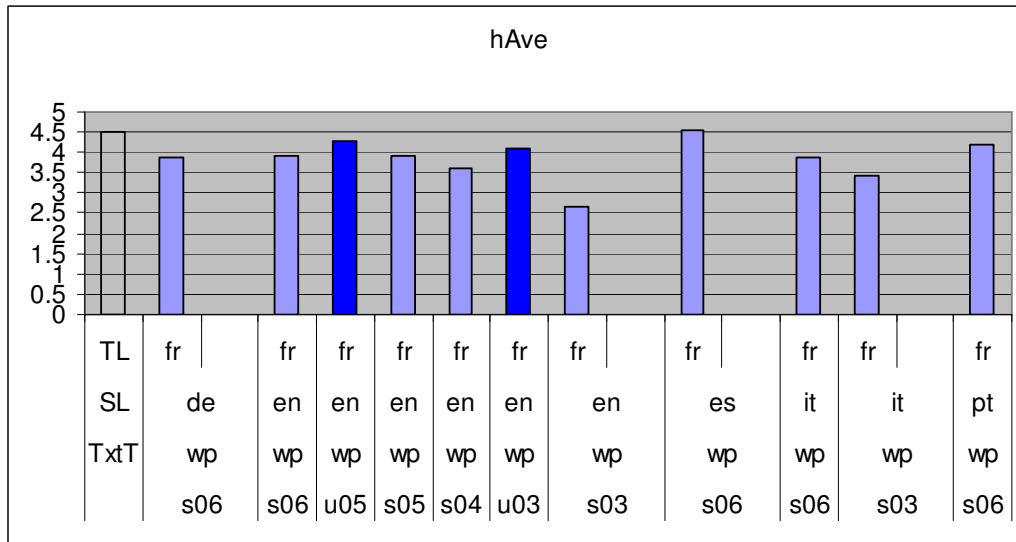


Figure 2: TL French – human evaluation scores for whitepaper and emails

Sys	TxtT	SL	TL	r=	hAve	hPass	r=	Bleu	cBlue	r=	LTVade	cLTV
s06	wp	de	fr		3.882	6660		0.2328			0.3294	
s06	wp	en	fr	1	3.924	8010	2	0.2551		2	0.3285	
u05	wp	en	fr		4.298	14700		0.3321			0.4276	
s05	wp	en	fr	2	3.902	7170	3	0.246		3	0.3247	
s04	wp	en	fr	3	3.62	1740	1	0.2998		1	0.399	
u03	wp	en	fr		4.118	11190		0.2617			0.3814	
s03	wp	en	fr	4	2.647	-17730	4	0.1341		4	0.2215	
corr									0.8643			0.84379
s06	wp	es	fr		4.562	19920		0.4748			0.536	
s06	wp	it	fr	1	3.88	6900	1	0.3721		1	0.4351	
s03	wp	it	fr	2	3.436	-1950	2	0.3381		2	0.3851	
s06	wp	pt	fr		4.204	12900		0.4236			0.4991	
corrA									0.7182			0.78831

Sys	TxtT	SL	TL	r=	hAve	hPass	r=	Bleu	cBlue	r=	LTVade	cLTV
s06	em	de	fr		3.654	3480		0.1656			0.2573	
s06	em	en	fr	1	3.974	13620	1	0.2414		1	0.3079	
u05	em	en	fr		3.846	9150		0.2498			0.3024	
s05	em	en	fr	2	3.649	3840	2	0.237		2	0.291	
s04	em	en	fr	3	3.351	-5760	3	0.201		3	0.2523	
u03	em	en	fr		3.811	8460		0.2228			0.2909	
s03	em	en	fr	4	2.854	-20730	4	0.139		4	0.2204	
corr									0.95382			0.986863
s06	em	es	fr		3.377	-4800		0.1613			0.2474	
s03	em	it	fr	1	3.45	-3000	2	0.1398		2	0.2075	
s06	em	it	fr	2	3.446	-2940	1	0.1695		1	0.2513	
s06	em	pt	fr		3.303	-6690		0.178			0.2508	
corrA									0.7732			0.820173

Table 6: TL French – system rankings and acceptability

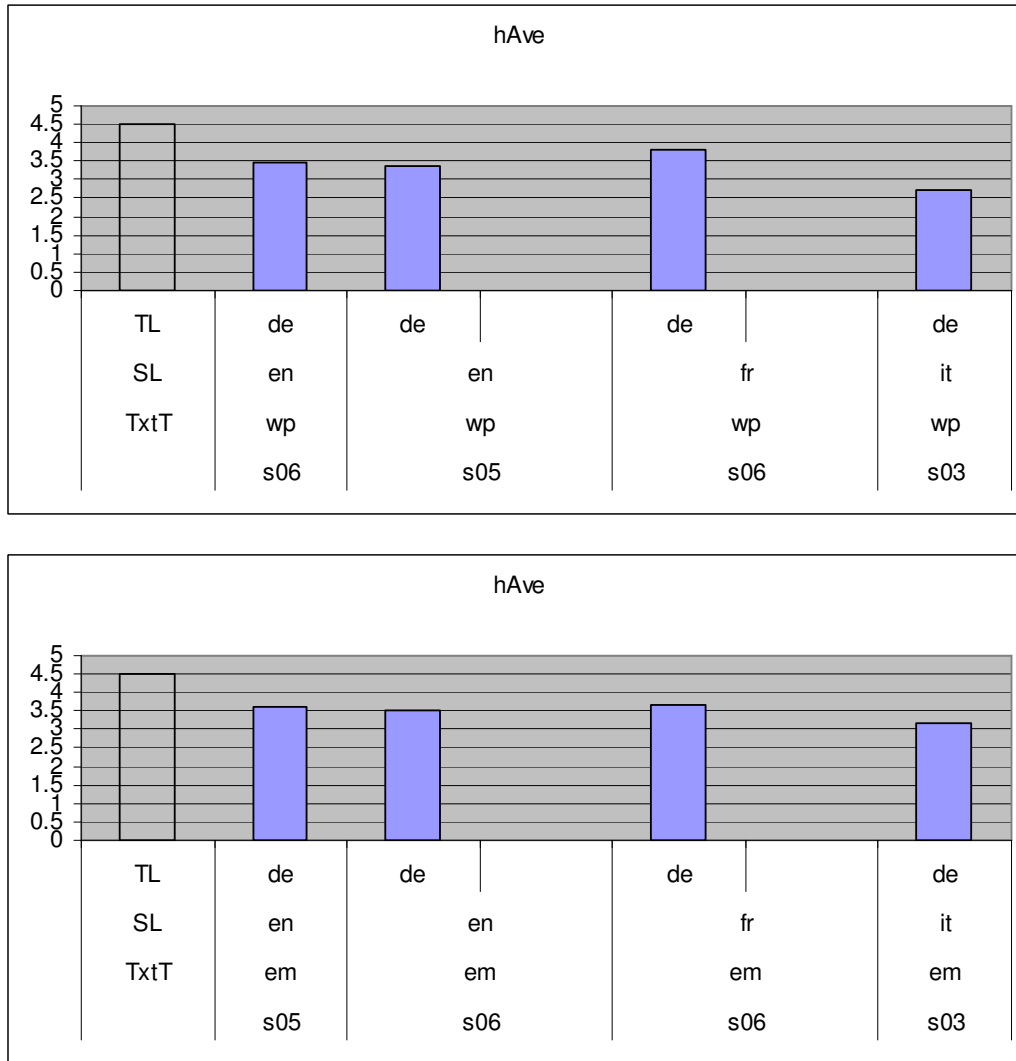


Figure 3: TL German – human evaluation scores for whitepaper and emails

Sys	TxtT	SL	TL	r=	hAve	hPass	r=	Bleu	cBlue	r=	LTVade	cLTV
s06	wp	en	de	1	3.469	-900	2	0.0581		1	0.1441	
s05	wp	en	de	2	3.342	-3780	1	0.0762		2	0.1386	
s06	wp	fr	de		3.818	5670		0.1314			0.2061	
s03	wp	it	de		2.707	-16770		0.0224			0.0977	
corrA									0.9168	0.94869		
Sys	TxtT	SL	TL	r=	hAve	hPass	r=	Bleu	cBlue	r=	LTVade	cLTV
s05	em	en	de	1	3.602	2130	1	0.236		1	0.3029	
s06	em	en	de	2	3.503	-1020	2	0.1969		2	0.2759	
s06	em	fr	de		3.665	3690		0.1496			0.2653	
s03	em	it	de		3.184	-10500		0.0644			0.1901	
corrA									0.7694	0.882381		

Table 7: TL German – system rankings and acceptability

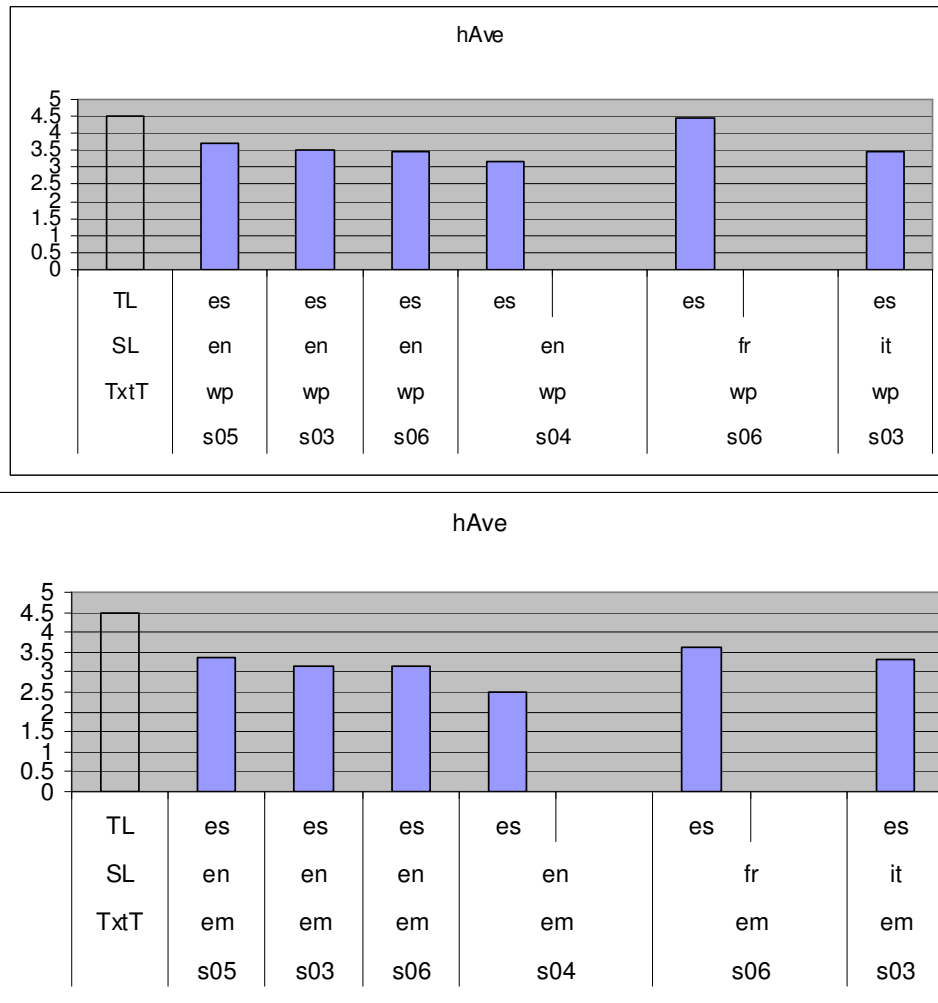


Figure 4: TL Spanish – human evaluation scores for whitepaper and emails

Sys	TxtT	SL	TL	r=	hAve	hPass	r=	Bleu	cBlue	r=	LTVade	cLTV
s05	wp	en	es	1	3.696	3240	2	0.2158		2	0.2785	
s03	wp	en	es	2	3.498	-540	3	0.2058		3	0.2704	
s06	wp	en	es	3	3.46	-1260	4	0.1987		4	0.2539	
s04	wp	en	es	4	3.171	-7110	1	0.2414		1	0.3269	
corr									-0.642			-0.6988
s06	wp	fr	es		4.456	18150		0.557			0.5771	
s03	wp	it	es		3.458	-1470		0.2802			0.3251	
corrA									0.8777			0.84309
Sys	TxtT	SL	TL	r=	hAve	hPass	r=	Bleu	cBlue	r=	LTVade	cLTV
s05	em	en	es	1	3.379	-4410	1	0.1988		1	0.26	
s03	em	en	es	2	3.149	-11520	4	0.1412		4	0.2136	
s06	em	en	es	3	3.126	-12090	2	0.1925		2	0.241	
s04	em	en	es	4	2.49	-31920	3	0.1592		3	0.225	
corr									0.46433			0.526954
s06	em	fr	es		3.618	3060		0.172			0.246	
s03	em	it	es		3.339	-5520		0.1379			0.2323	
corrA									0.17749			0.528202

Table 8: TL Spanish – system rankings and acceptability

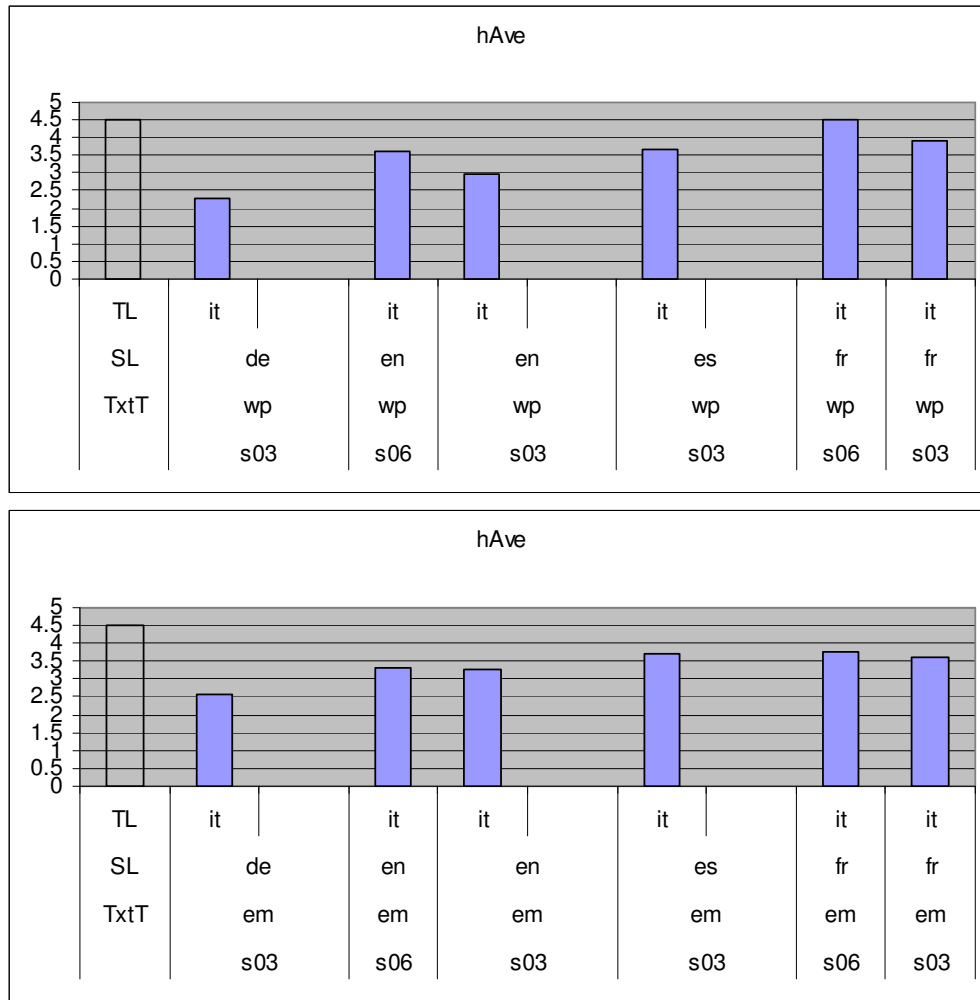


Figure 5: TL Italian – human evaluation scores for whitepaper and emails

Sys	TxtT	SL	TL	r=	hAve	hPass	r=	Bleu	cBlue	r=	LTVade	cLTV
s03	wp	de	it		2.287	-25470		0.0648			0.1837	
s06	wp	en	it	1	3.611	1410	1	0.1397		1	0.2362	
s03	wp	en	it	2	2.964	-11310	2	0.0898		2	0.18	
s03	wp	es	it		3.667	2580		0.2028			0.2761	
s06	wp	fr	it	1	4.5	18300	1	0.4348		1	0.4848	
s03	wp	fr	it	2	3.902	6870	2	0.3172		2	0.3738	
corrA									0.9064	0.88729		
Sys	TxtT	SL	TL	r=	hAve	hPass	r=	Bleu	cBlue	r=	LTVade	cLTV
s03	em	de	it		2.551	-30150		0.0858			0.1869	
s06	em	en	it	1	3.333	-6270	1	0.1544		1	0.2206	
s03	em	en	it	2	3.282	-7860	2	0.117		2	0.2018	
s03	em	es	it		3.705	4590		0.1166			0.2049	
s06	em	fr	it	1	3.743	6240	1	0.1799		1	0.253	
s03	em	fr	it	2	3.598	1650	2	0.1536		2	0.232	
corrA									0.74004	0.734464		

Table 9: TL Italian – system rankings and acceptability

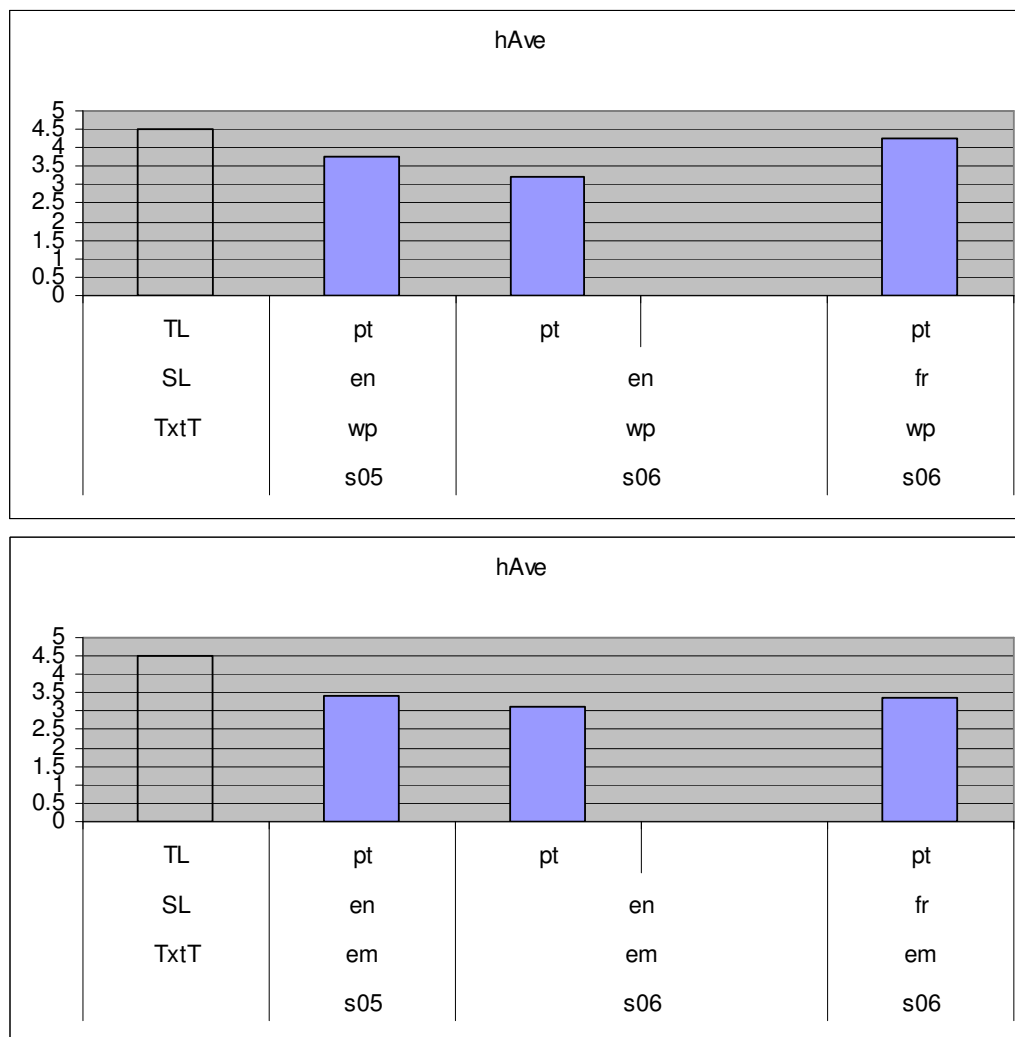


Figure 6: TL Portuguese – human evaluation scores for whitepaper and emails

Sys	TxtT	SL	TL	r=	hAve	hPass	r=	Bleu	cBlue	r=	LTVade	cLTV
s05	wp	en	pt	1	3.771	4710	1	0.1341		1	0.201	
s06	wp	en	pt	2	3.196	-6450	2	0.0901		2	0.1497	
s06	wp	fr	pt		4.262	14460		0.4214			0.4512	
corrA									0.902			
corrA									0.902	0.91737		
Sys	TxtT	SL	TL	r=	hAve	hPass	r=	Bleu	cBlue	r=	LTVade	cLTV
s05	em	en	pt	1	3.409	-3930	1	0.2076		1	0.2703	
s06	em	en	pt	2	3.114	-13080	2	0.1435		2	0.2232	
s06	em	fr	pt		3.377	-4620		0.1616			0.2353	
corrA									0.78333			
corrA									0.78333	0.765963		

Table 10: TL Portuguese – system rankings and acceptability

5 Automated evaluations

We take the automated scores as an attempt to replicate or reliably approximate the human judgments. This section describes the procedure followed in obtaining the automated scores and presents all the results in figures and tables. The human scores presented in section 4.2 are used to calibrate the automated scores by means of the calibration table presented in section 6.1.

5.1 Commentary on the automated metrics

5.1.1 BLEU

Absolute performance figures for BLEU¹ may vary substantially depending on the **target language (TL)**, the **source language (SL)** and the **text type** – emails or the whitepaper.

Variation of BLEU scores by the **TL** is due to reasons external to translation – the **TL's** linguistic structure. BLEU computes the distance between an MT text and a human reference in terms of the number of matched word tokens. But the absolute number of matches depends on whether the TL allows a greater or smaller degree of variation for different forms of the same word and for the word order. For English, this is a relatively minor issue, but it can become a major issue for heavily inflected Romance languages (such as French, Italian and Portuguese) or languages with a 'free' word order (such as German).

Therefore, absolute scores are not comparable across different TLs and need to be calibrated separately for each language, using human evaluation scores.

However, variation of BLEU scores by the **SL** and **text type** is due to translation-internal reasons, namely the difficulty of the translation task:

- similarity between the SL and TL (translation between closely related languages is easier)
- amenability of language used in a particular text type for a given MT system (some systems may be tuned to for specific text types or subject domains).

Although variation of BLEU scores by these parameters reflects the difficulty of the translation task, the scores are comparable and represent a level of MT quality which is achievable given such difficulty.

5.1.2 LTV

The LTV² (Legitimate Translation Variation) method takes into account the relative salience of matched items. LTV computes two scores:

¹ Paineni K, Roukos S, Ward, T, Zhu W-J 2001 *Bleu: a method for automatic evaluation of machine translation* IBM Research Report RC2176 (W0109-022) September 17, 2001.

² Babych B, Hartley A 2004a *Extending the BLEU MT Evaluation Method with Frequency Weightings* Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, July, 2004. pp. 622-629.

Babych B, Hartley A 2004b *Modelling legitimate translation variation for automatic evaluation of MT*

- weighted recall, which usually correlates with human judgements about adequacy
- weighted F-score (a combination of precision and recall), which usually correlates with human judgements about fluency.

For the systems evaluated, both parameters correlate with each other very closely ($r = 0.9947$), so both scores typically yield the same ranking of the evaluated systems.

5.2 System rankings resulting from BLEU and LTV

The tables in this section show, for each set of systems grouped by the same language direction, the rankings according to the LTV score for adequacy (ADE), the LTC score for fluency (FLU) and the BLEU score.

- LTV weighted recall (adequacy) scores for the evaluated systems typically confirm the ranking produced by BLEU scores ($r = 0.9686$).
- BLEU scores usually better correlate with fluency, and naturally they show greater agreement with the weighted F-score ($r = 0.97978$) than with weighted recall.

Wherever there is difference in ranking produced by LTV (ADE) and the other two scores we indicate this by highlighting the discrepancy in **blue**.

sys	txtT	SL	TL	rLtv-	ADE	rLtv-	FLU	r-	BLEU
s06	wp	de	en	1	0.232	1	0.2903	1	0.1374
s01	wp	de	en	2	0.2314	2	0.2898	2	0.1349
s02	wp	de	en	3	0.2084	4	0.2636	4	0.1193
s05	wp	de	en	4	0.2055	3	0.2642	3	0.1258
s04	wp	es	en	1	0.3308	1	0.3862	1	0.2237
s05	wp	es	en	2	0.2721	2	0.3327	3	0.1702
s02	wp	es	en	3	0.2642	3	0.3236	2	0.1707
s01	wp	es	en	4	0.2532	4	0.3096	4	0.1595
s06	wp	es	en	5	0.2528	5	0.3091	5	0.1585
s03	wp	es	en	6	0.2231	6	0.2836	6	0.1278
s04	wp	fr	en	1	0.3674	1	0.418	1	0.2746
s06	wp	fr	en	2	0.3091	2	0.3591	3	0.2242
s01	wp	fr	en	3	0.3068	3	0.3568	2	0.2247
s02	wp	fr	en	4	0.2694	4	0.3272	5	0.1935
u05	wp	fr	en		0.433		0.4757		0.3354
s05	wp	fr	en	5	0.2565	5	0.3121	4	0.2002
u03	wp	fr	en		0.3636		0.4065		0.237
s03	wp	fr	en	6	0.2026	6	0.2629	6	0.1253
s01	wp	it	en	1	0.2342	1	0.293	1	0.1317
s06	wp	it	en	2	0.2339	2	0.2921	2	0.1296
s03	wp	it	en	3	0.2024	3	0.2596	3	0.0974
s06	wp	pt	en	1	0.2256	1	0.2834	1	0.1216
s01	wp	pt	en	2	0.2244	2	0.2828	2	0.1216

Table 11: TL English – system rankings whitepaper

sys	txtT	SL	TL	rLtv-	ADE	rLtv	FLU	r-	BLEU
s05	em	de	en	1	0.4213	1	0.462	1	0.3207
s06	em	de	en	2	0.4005	2	0.4478	2	0.2951
s02	em	de	en	3	0.3895	3	0.4407	4	0.2841
s01	em	de	en	4	0.3876	4	0.4395	3	0.2933
s05	em	es	en	1	0.3473	1	0.3886	2	0.218
s02	em	es	en	2	0.3419	2	0.3842	1	0.222
s06	em	es	en	3	0.3196	3	0.3696	3	0.1959
s01	em	es	en	4	0.3122	4	0.3622	4	0.1932
s04	em	es	en	5	0.2612	5	0.3272	5	0.1712
s03	em	es	en	6	0.2518	6	0.302	6	0.1432
s06	em	fr	en	1	0.392	1	0.4428	1	0.2862
s01	em	fr	en	2	0.3699	2	0.4229	2	0.2759
s02	em	fr	en	3	0.3666	3	0.4139	4	0.2473
u05	em	fr	en		0.4446		0.4751		0.3339
s05	em	fr	en	4	0.3513	4	0.3994	3	0.2475
s04	em	fr	en	5	0.3294	5	0.3896	5	0.2259
u03	em	fr	en		0.388		0.4257		0.2659
s03	em	fr	en	6	0.2982	6	0.3468	6	0.174
s06	em	it	en	1	0.2856	1	0.3403	1	0.1746
s03	em	it	en	2	0.2716	3	0.2932	3	0.132
s01	em	it	en	3	0.2705	2	0.3272	2	0.1676
s06	em	pt	en	1	0.3075	1	0.3676	1	0.2051
s01	em	pt	en	2	0.2846	2	0.346	2	0.192

Table 12: TL English – system rankings emails

sys	txtT	SL	TL	rLtv-	ADE	rLtv-	FLU	r-	BLEU
s06	wp	de	fr	1	0.3294	1	0.3948	1	0.2328
s02	wp	de	fr	2	0.2524	2	0.3168	2	0.1471
s04	wp	en	fr	1	0.399	1	0.4624	1	0.2998
s06	wp	en	fr	2	0.3285	2	0.3984	2	0.2551
u01	wp	en	fr		0.4359		0.4941		0.3377
s01	wp	en	fr	3	0.3256	3	0.3979	3	0.2537
u05	wp	en	fr		0.4276		0.4837		0.3321
s05	wp	en	fr	4	0.3247	4	0.3931	4	0.246
s02	wp	en	fr	5	0.2914	5	0.3582	5	0.2042
u03	wp	en	fr		0.3814		0.4344		0.2617
s03	wp	en	fr	6	0.2215	6	0.2874	6	0.1341
s06	wp	es	fr	1	0.536	1	0.5906	1	0.4748
s02	wp	es	fr	2	0.5086	2	0.5714	2	0.4433
s06	wp	it	fr	1	0.4351	1	0.5043	1	0.3721
s03	wp	it	fr	2	0.3851	2	0.4574	2	0.3381
s06	wp	pt	fr		0.4991		0.5543		0.4236

sys	txtT	SL	TL	rLtv-	ADE	rLtv-	FLU	r-	BLEU
s06	em	de	fr	1	0.2573	1	0.3144	2	0.1656
s02	em	de	fr	2	0.2507	2	0.3126	1	0.1712
s06	em	en	fr	1	0.3079	1	0.371	1	0.2414
u05	em	en	fr		0.3024		0.3638		0.2498
s05	em	en	fr	2	0.291	2	0.3546	2	0.237
u01	em	en	fr		0.3175		0.3818		0.265
s01	em	en	fr	3	0.2839	3	0.3494	3	0.2352
s02	em	en	fr	4	0.2745	4	0.3419	4	0.2156
s04	em	en	fr	5	0.2523	5	0.3181	5	0.201
u03	em	en	fr		0.2909		0.3513		0.2228
s03	em	en	fr	6	0.2204	6	0.2795	6	0.139
s06	em	es	fr	1	0.2474	1	0.3105	1	0.1613
s02	em	es	fr	2	0.2336	2	0.2904	2	0.1544
s06	em	it	fr	1	0.2513	1	0.3147	1	0.1695
s03	em	it	fr	2	0.2075	2	0.2652	2	0.1398
s06	em	pt	fr		0.2508		0.3175		0.178

Table 13: TL French – system rankings whitepaper and emails

sys	txtT	SL	TL	rLtv-	ADE	rLtv-	FLU	r-	BLEU
s06	wp	en	de	1	0.1441	1	0.1983	3	0.0581
s05	wp	en	de	2.5	0.1386	2	0.1958	1	0.0762
s01	wp	en	de	2.5	0.1386	3	0.1914	4	0.052
s02	wp	en	de	4	0.1229	4	0.1743	2	0.0583
s02	wp	es	de		0.1387		0.1941		0.0678
s06	wp	fr	de	1	0.2061	1	0.2672	1	0.1314
s02	wp	fr	de	2	0.1349	2	0.1868	2	0.0649
s03	wp	it	de		0.0977		0.1422		0.0224

sys	txtT	SL	TL	rLtv-	ADE	rLtv-	FLU	r-	BLEU
s05	em	en	de	1	0.3029	1	0.3704	1	0.236
s02	em	en	de	2	0.2966	2	0.3639	2	0.2244
s06	em	en	de	3	0.2759	3	0.3398	4	0.1969
s01	em	en	de	4	0.2741	4	0.3367	3	0.197
s02	em	es	de		0.2294		0.2745		0.1088
s06	em	fr	de	1	0.2653	1	0.3176	1	0.1496
s02	em	fr	de	2	0.2434	2	0.301	2	0.1245
s03	em	it	de		0.1901		0.2184		0.0644

Table 14: TL German – system rankings whitepaper and emails

sys	txtT	SL	TL	rLtv-	ADE	rLtv-	FLU	r-	BLEU
s02	wp	de	es		0.2269		0.2931		0.1522
s04	wp	en	es	1	0.3269	1	0.3998	1	0.2414
s05	wp	en	es	2	0.2785	2	0.3514	2	0.2158
s03	wp	en	es	3	0.2704	3	0.3446	3	0.2058
s02	wp	en	es	4	0.266	4	0.3358	5	0.2031
s01	wp	en	es	5	0.2565	5	0.329	4	0.2034
s06	wp	en	es	6	0.2539	6	0.3265	6	0.1987
s06	wp	fr	es	1	0.5771	1	0.6361	1	0.557
s02	wp	fr	es	2	0.495	2	0.5652	2	0.4579
s03	wp	it	es		0.3251		0.3984		0.2802
sys	txtT	SL	TL	rLtv-	ADE	rLtv-	FLU	r-	BLEU
s02	em	de	es		0.2423		0.3042		0.1534
s05	em	en	es	1	0.26	1	0.3164	1	0.1988
s02	em	en	es	2	0.2467	2	0.3091	3	0.1897
s06	em	en	es	3	0.241	3	0.2953	2	0.1925
s01	em	en	es	4	0.2301	5	0.285	4	0.1833
s04	em	en	es	5	0.225	4	0.2901	5	0.1592
s03	em	en	es	6	0.2136	6	0.2626	6	0.1412
s06	em	fr	es	1	0.246	1	0.3143	1	0.172
s02	em	fr	es	2	0.2405	2	0.3078	2	0.1707
s03	em	it	es		0.2323		0.2898		0.1379

Table 15: TL Spanish – system rankings whitepaper and emails

sys	txtT	SL	TL	rLtv-	ADE	rLtv-	FLU	r-	BLEU
s03	wp	de	it		0.1837		0.2394		0.0648
s06	wp	en	it	1	0.2362	1	0.2966	1	0.1397
s01	wp	en	it	2	0.2313	2	0.2905	2	0.1342
s02	wp	en	it	3	0.1919	3	0.2541	3	0.1156
s03	wp	en	it	4	0.18	4	0.2349	4	0.0898
s03	wp	es	it		0.2761		0.3389		0.2028
s06	wp	fr	it	1	0.4848	1	0.5414	1	0.4348
s03	wp	fr	it	2	0.3738	2	0.4418	2	0.3172

sys	txtT	SL	TL	rLtv-	ADE	rLtv-	FLU	r-	BLEU
s03	em	de	it		0.1869		0.2347		0.0858
s06	em	en	it	1	0.2206	1	0.2849	2	0.1544
s02	em	en	it	2	0.219	2	0.28	1	0.1567
s01	em	en	it	3	0.2077	3	0.27	3	0.1475
s03	em	en	it	4	0.2018	4	0.2535	4	0.117
s03	em	es	it		0.2049		0.2623		0.1166
s06	em	fr	it	1	0.253	1	0.3272	1	0.1799
s03	em	fr	it	2	0.232	2	0.2914	2	0.1536

Table 16: TL Italian – system rankings whitepaper and emails

sys	txtT	SL	TL	rLtv-	ADE	rLtv-	FLU	r-	BLEU
s05	wp	en	pt	1	0.201	1	0.2679	1	0.1341
s01	wp	en	pt	2	0.1578	2	0.2199	2	0.096
s06	wp	en	pt	3	0.1497	3	0.2103	3	0.0901
s06	wp	fr	pt		0.4512		0.5233		0.4214
s05	em	en	pt	1	0.2703	1	0.3427	1	0.2076
s06	em	en	pt	2	0.2232	2	0.2766	2	0.1435
s01	em	en	pt	3	0.2093	3	0.262	3	0.1352
s06	em	fr	pt		0.2353		0.3004		0.1616

Table 17: TL Portuguese – system rankings whitepaper and emails

5.3 BLEU scores

We group the evaluation results by target language. The results for the whitepaper and the emails are given in separate charts on the same page, so that they may be compared directly. In each chart the scores for the same language direction are grouped together. Systems are sorted by quality rank within their group, better systems coming first.

Versions of the system with dictionary adaptation are shown next to the baseline systems; in dark blue.

Remember from section 5.1.1 that, for reasons inherent in the structure of each language, absolute scores are not comparable across different target languages and that the scores need to be calibrated separately for each target language, using human evaluation scores.

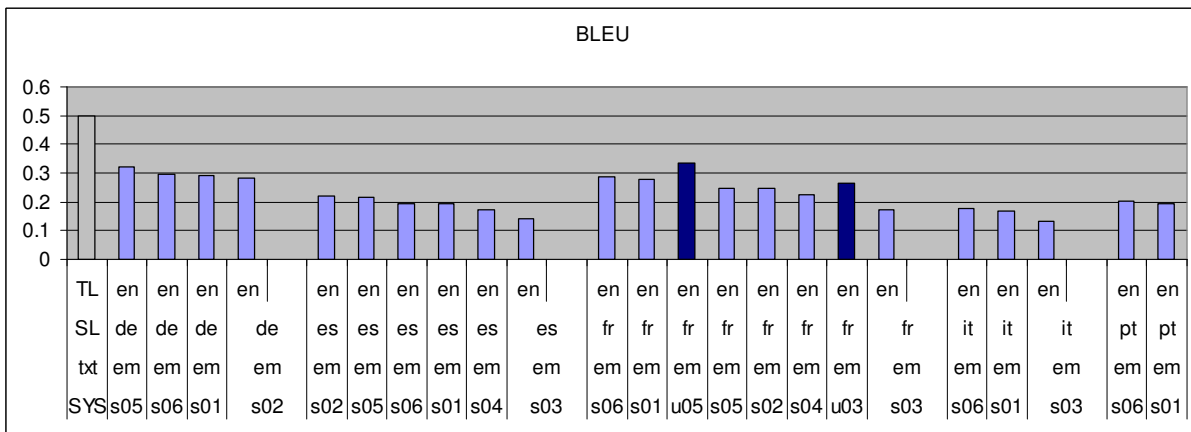
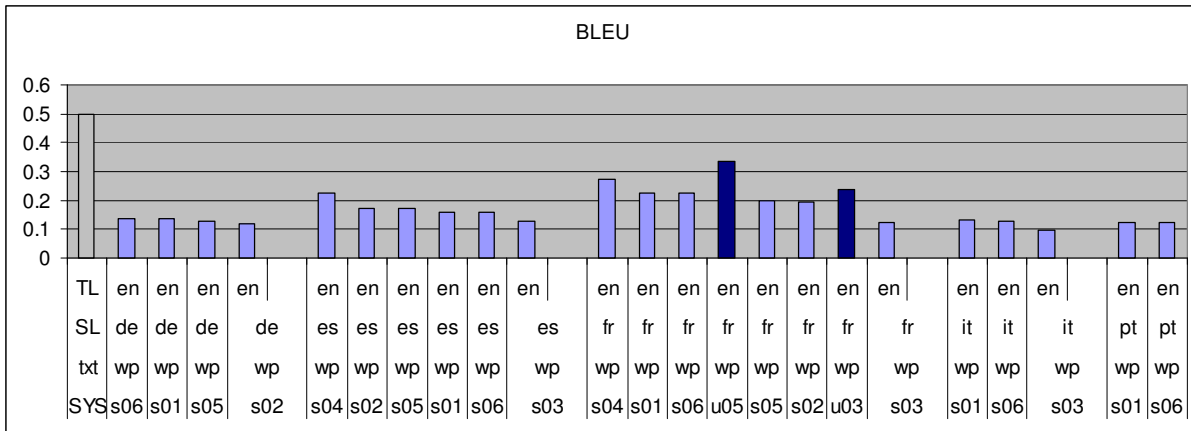


Figure 7: TL English – BLEU scores for whitepaper and emails

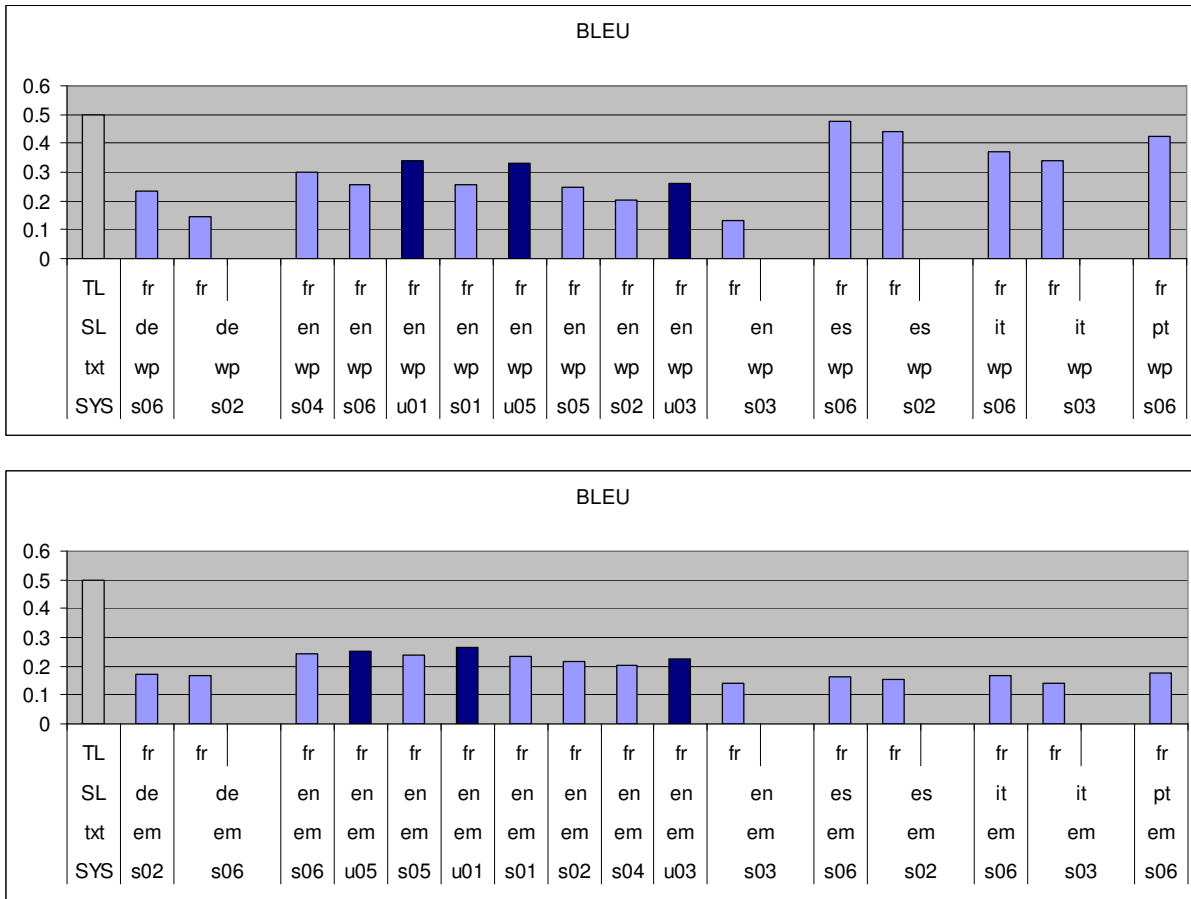


Figure 8: TL French – BLEU scores for whitepaper and emails

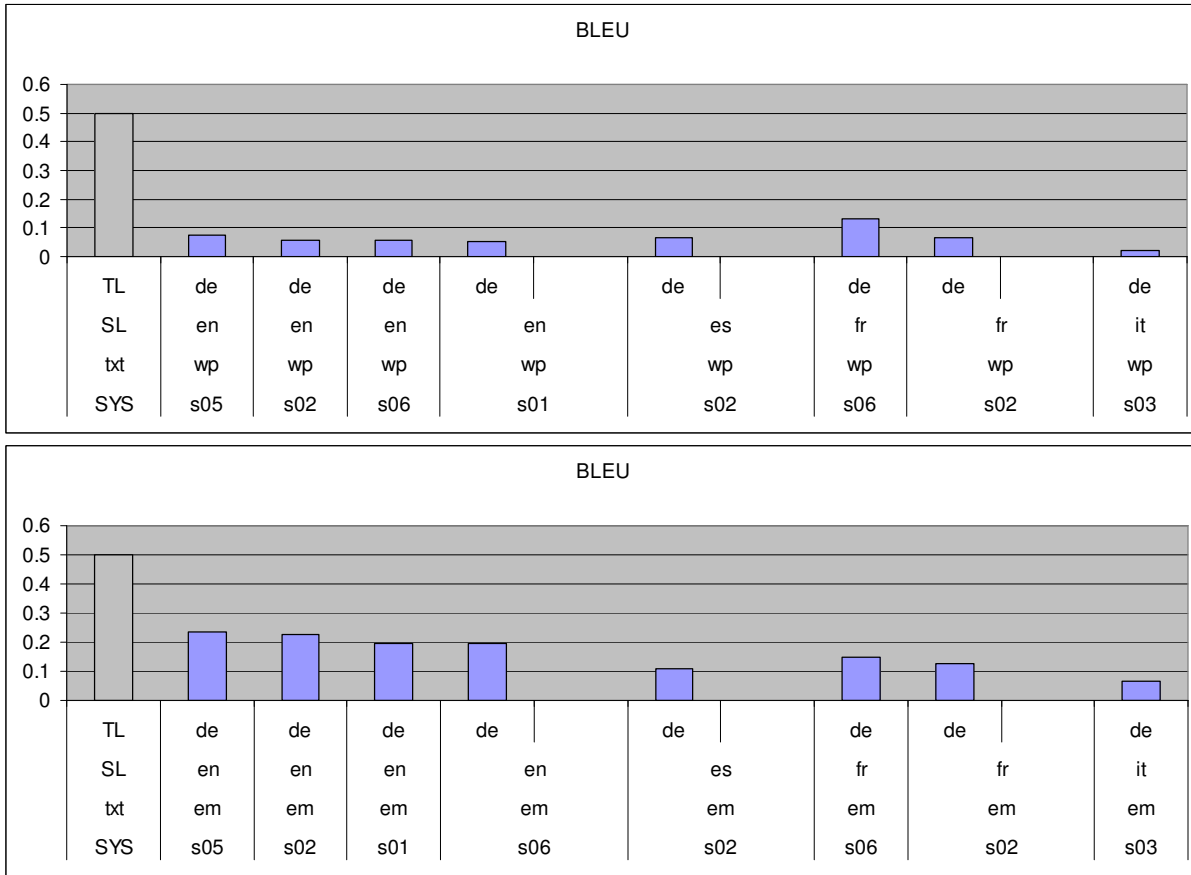


Figure 9: TL German – BLEU scores for whitepaper and emails

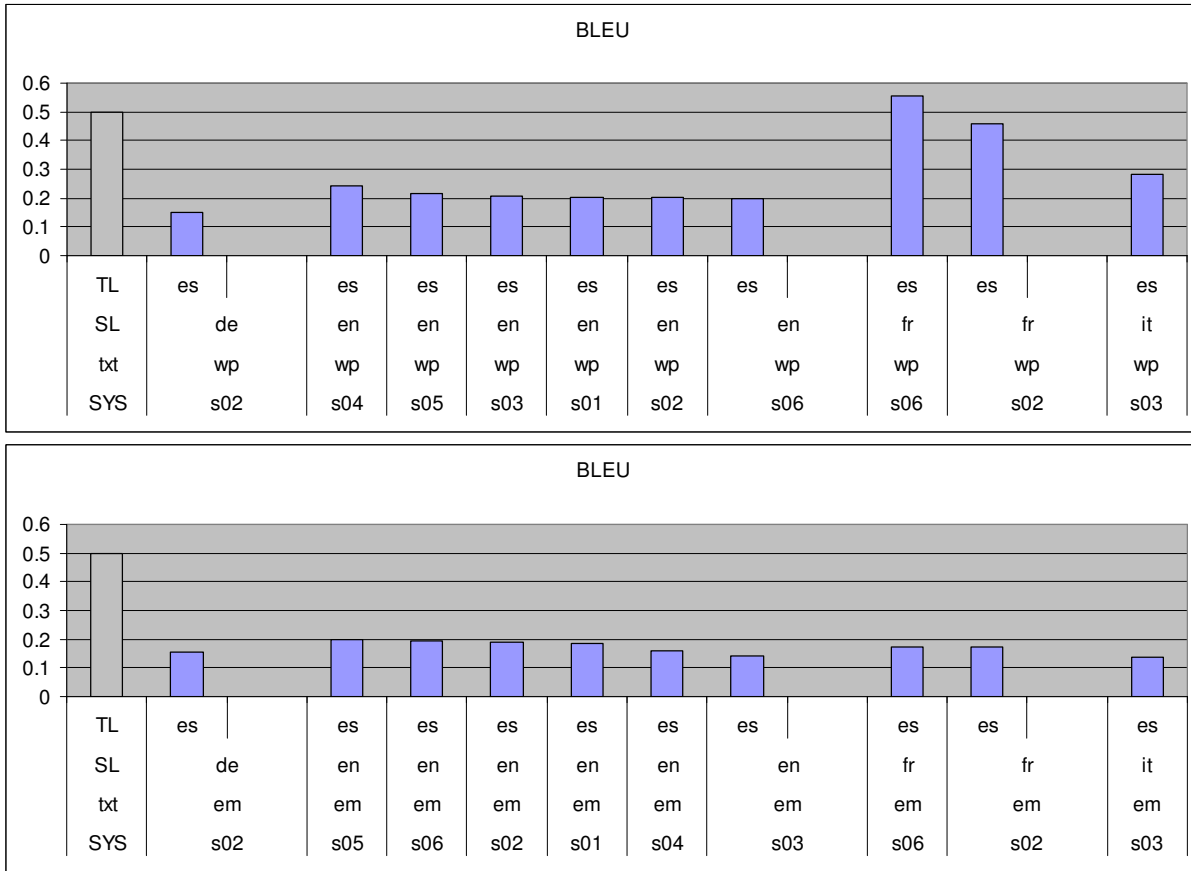


Figure 10: TL Spanish – BLEU scores for whitepaper and emails

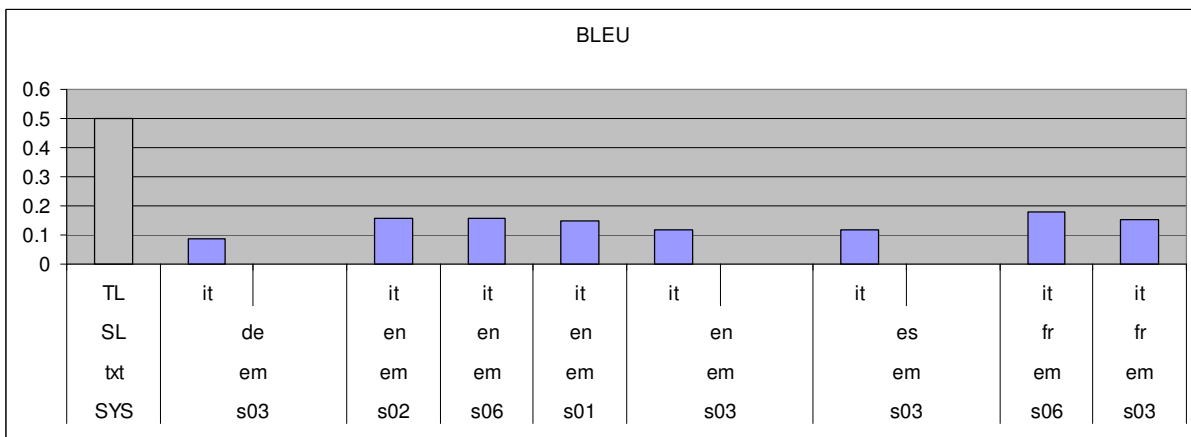
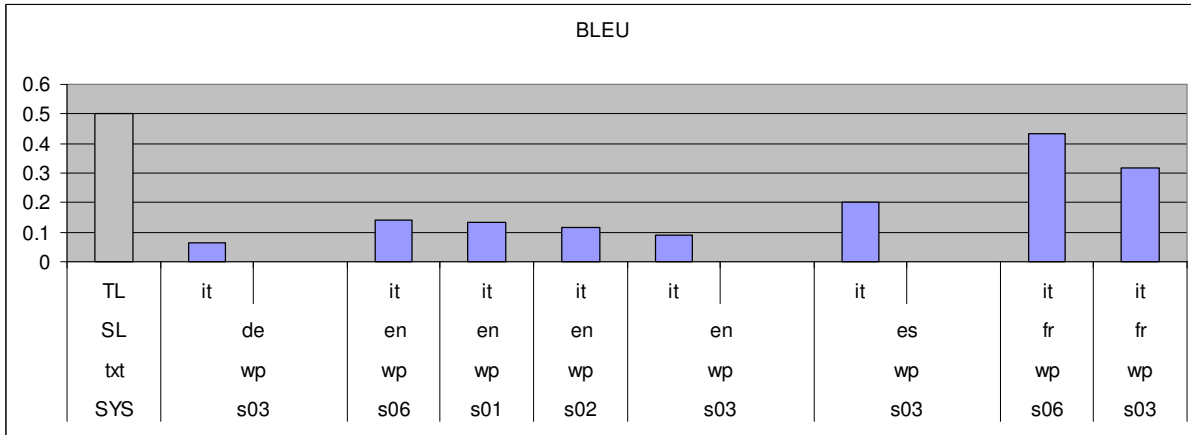


Figure 11: TL Italian – BLEU scores for whitepaper and emails

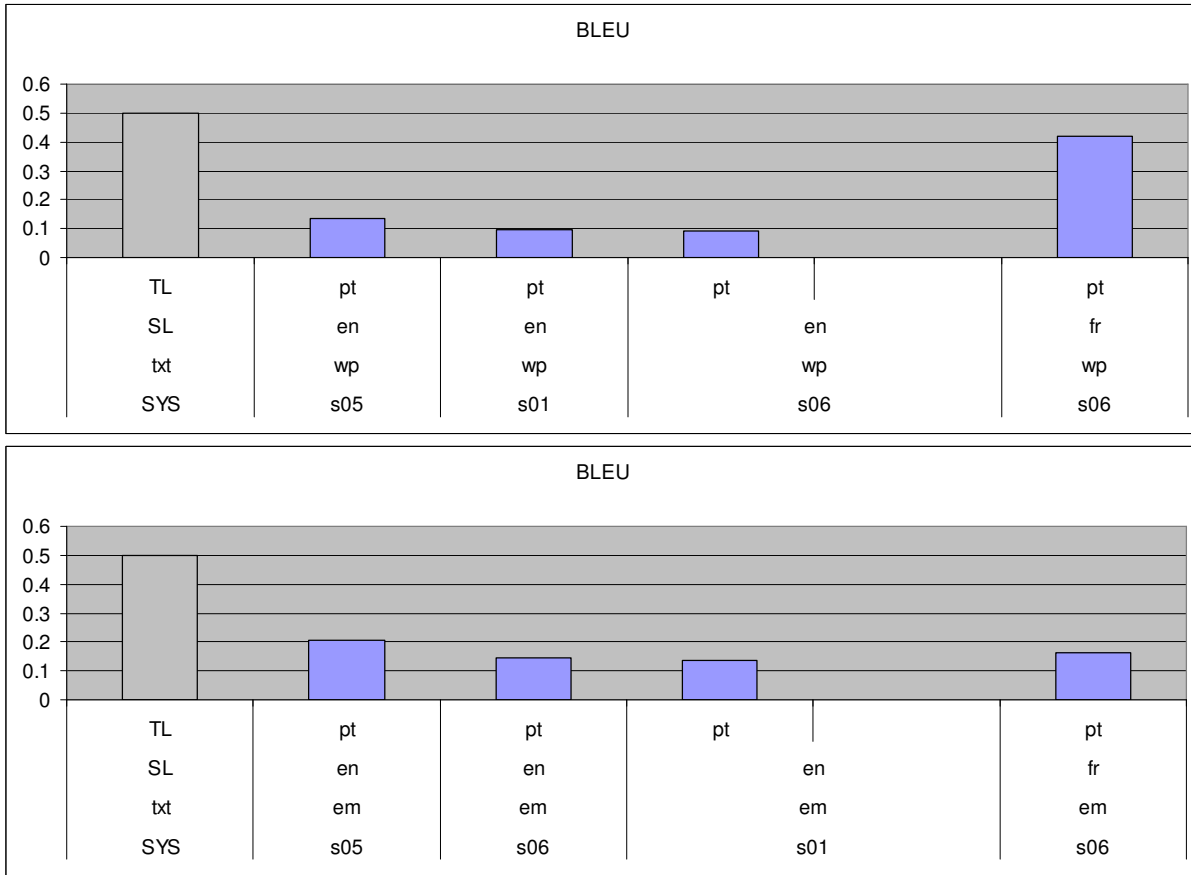


Figure 12: TL Portuguese – BLEU scores for whitepaper and emails

5.4 LTV scores

Since weighted recall (adequacy) and the weighted F-score (fluency) correlate with each other very closely ($r = 0.9947$), the charts show only the weighted recall (ADE) figures.

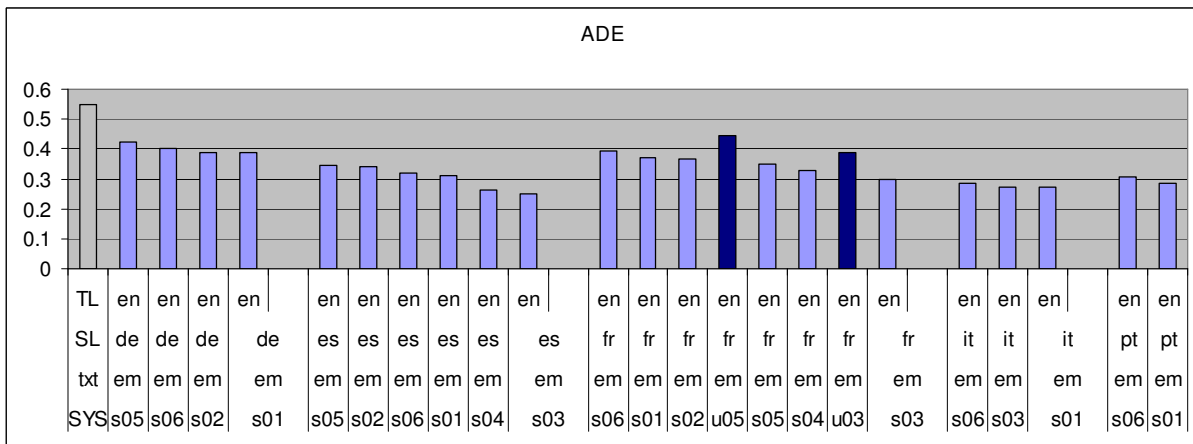
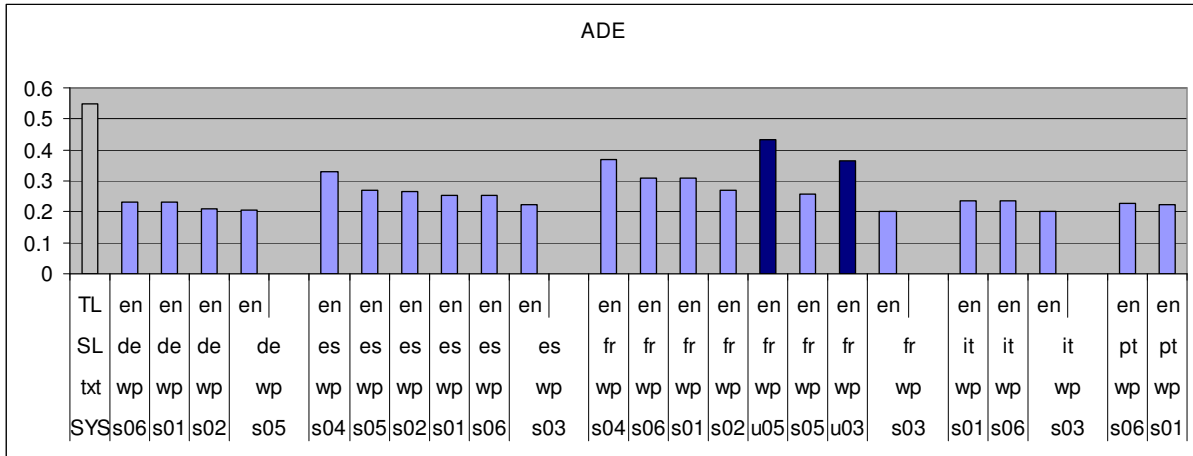


Figure 13: TL English – LTV scores for whitepaper and emails

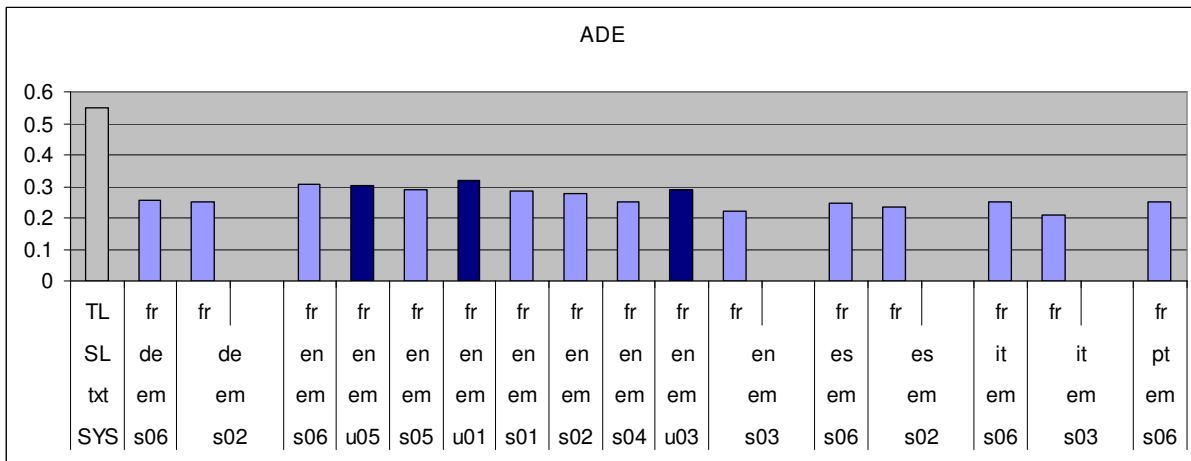
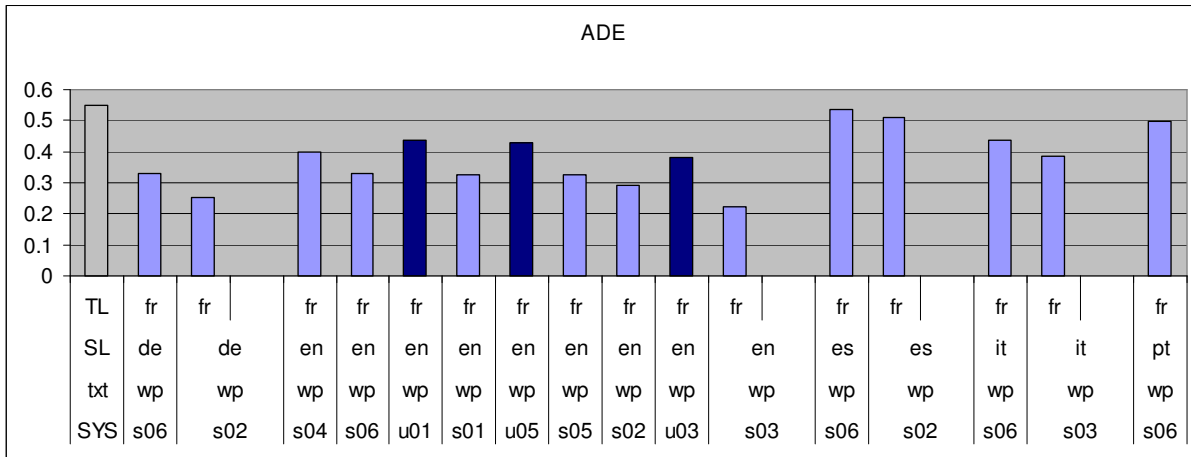


Figure 14: TL French – LTV scores for whitepaper and emails

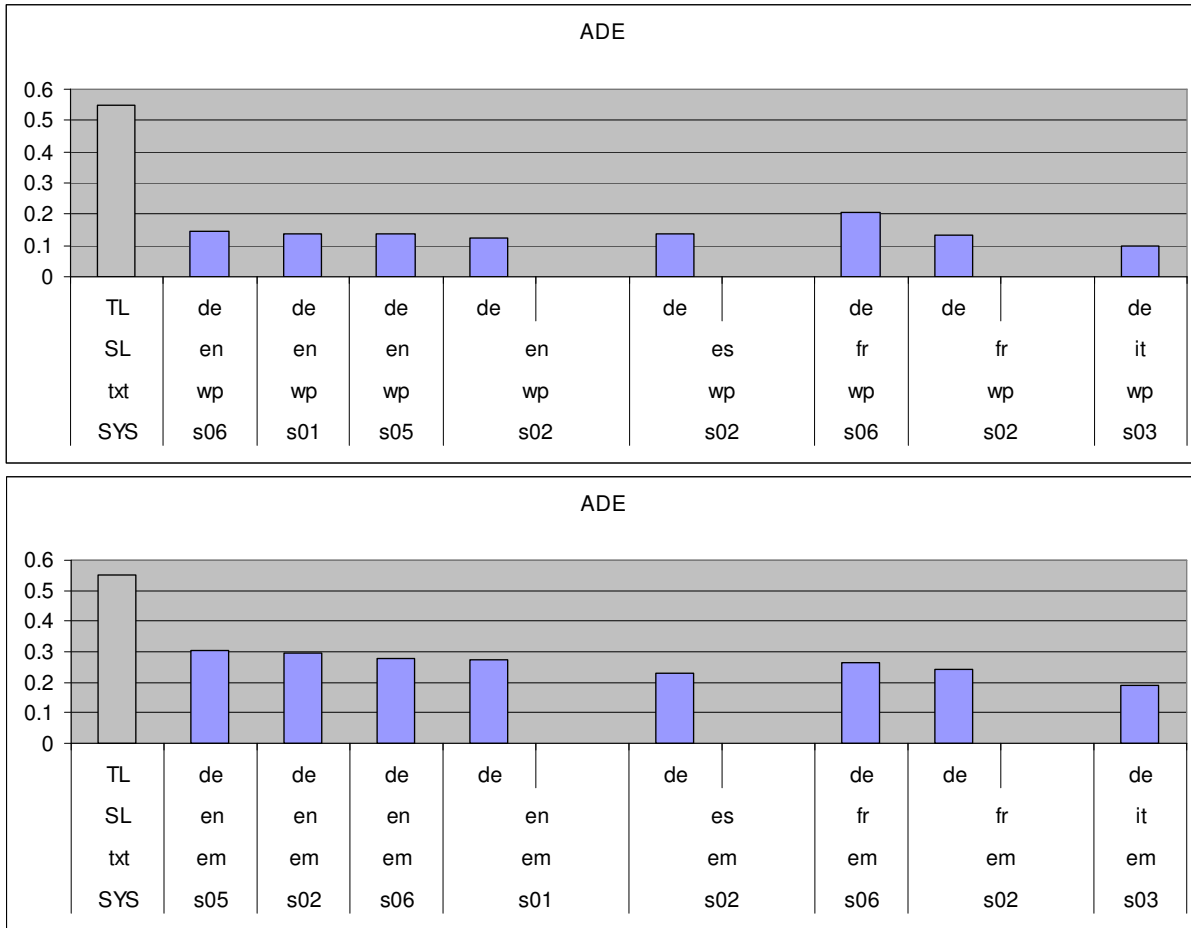


Figure 15: TL German – LTV scores for whitepaper and emails

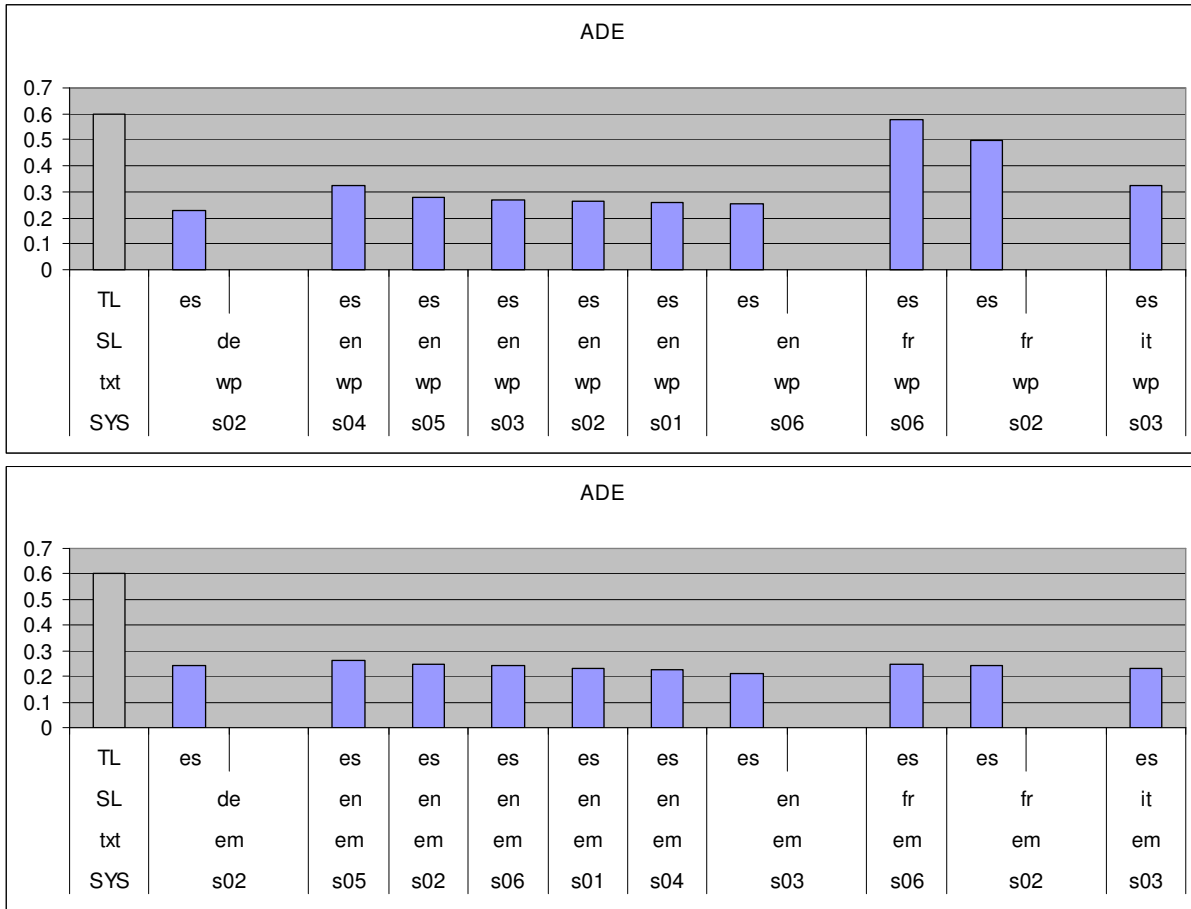


Figure 16: TL Spanish – LTV scores for whitepaper and emails

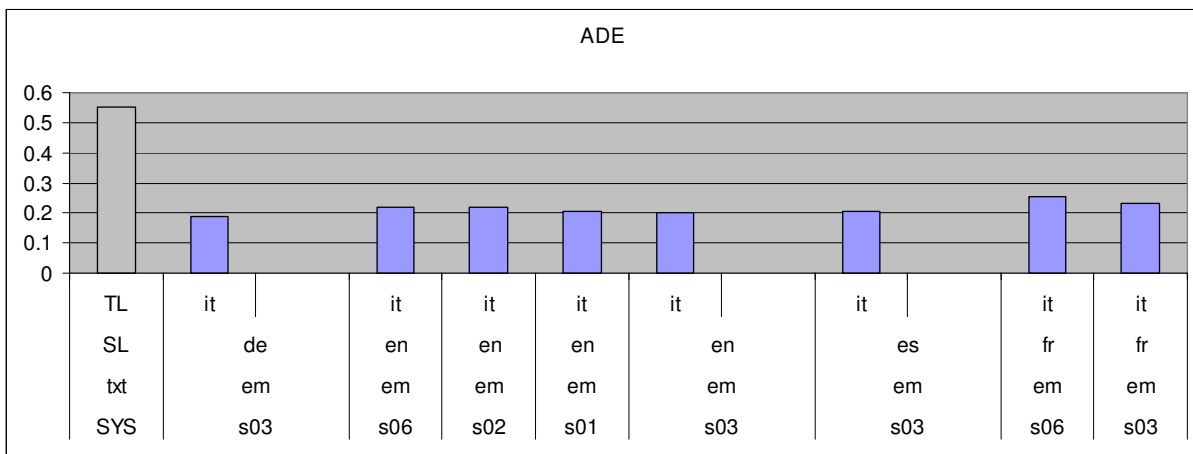
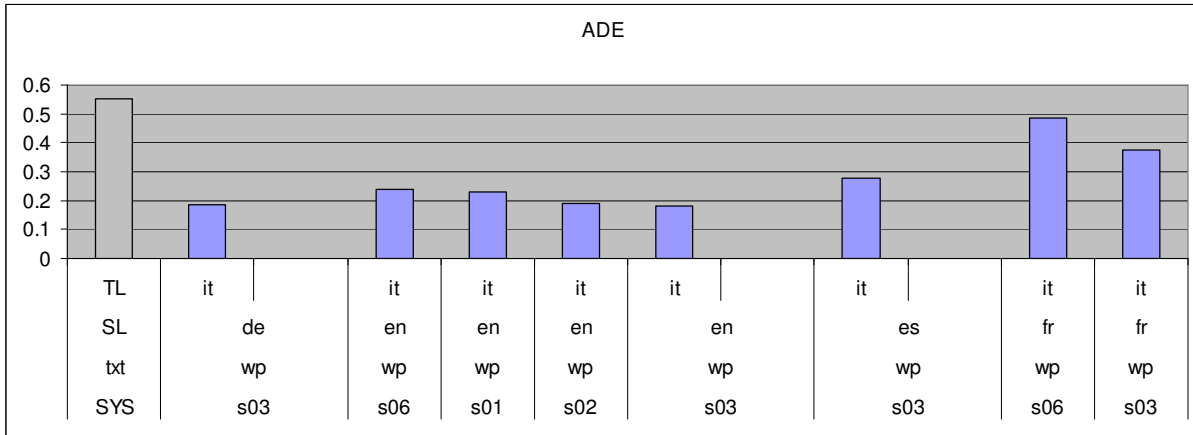


Figure 17: TL Italian – LTV scores for whitepaper and emails

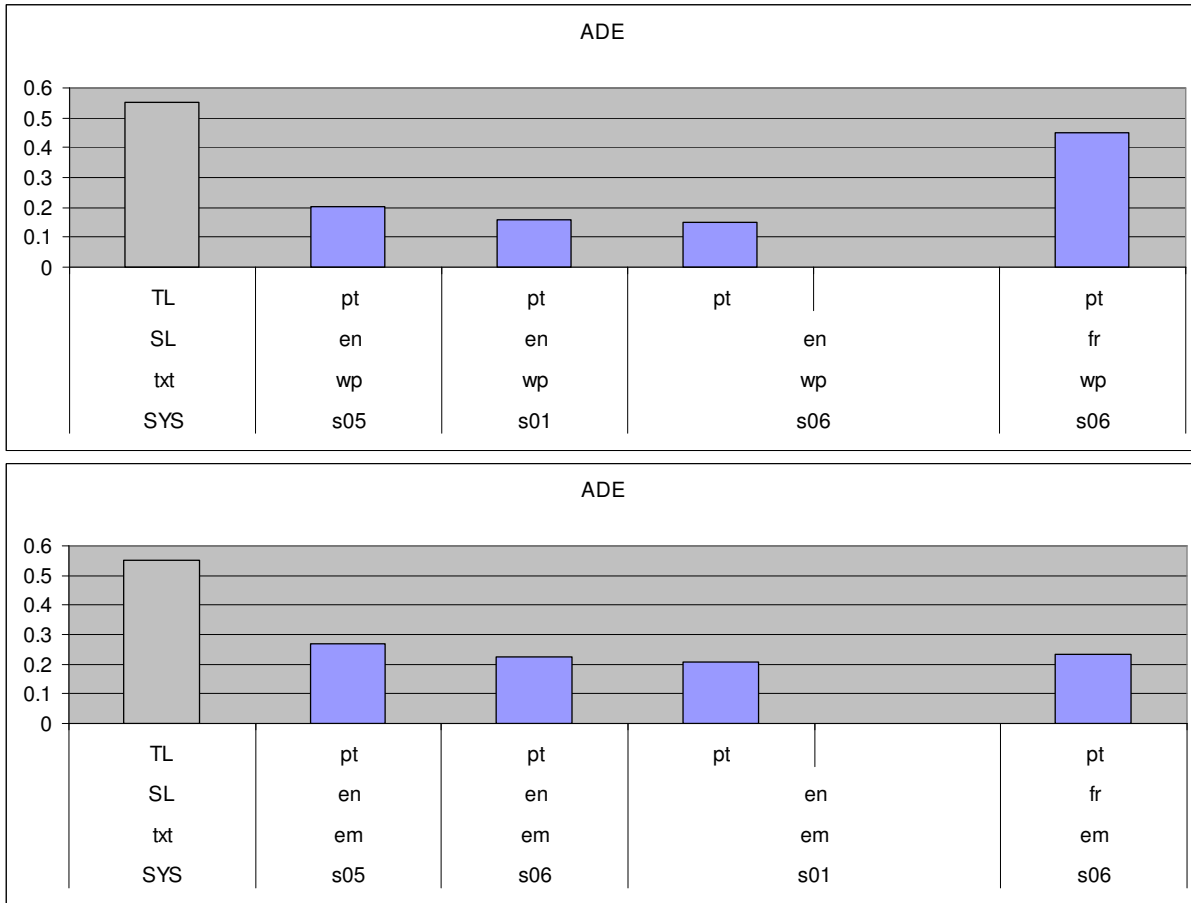


Figure 18: TL Portuguese – LTV scores for whitepaper and emails

6 Calibration of automated scores

6.1 Purpose

The purpose of the calibration table given in section **Error! Reference source not found.** is to benchmark quickly and at low cost the performance of later versions of the MT systems evaluated here or of other systems which Translution might choose to consider in the future.

This tool takes the form of a table of coefficients that vary according to the parameters of text type (email or whitepaper), source language and target language. The coefficients are derived from observed correlations between the human evaluation scores and the automated evaluation scores.

An updated or new system can be evaluated relative to the systems discussed in this report and to the quality threshold in four simple steps:

- generate the translation of the email and/or whitepaper
- calculate the BLEU and/or LTV scores
- apply the coefficients to the automated scores
- check the result against the threshold value.

6.2 Derivation of the coefficients

We experimented with a number of approaches in order to calibrate the automated scores. Currently the best results are achieved with a method which takes into account information about *text type* and *target language*. Given the automated scores, the method allows us to predict human scores with reasonable accuracy.

The coefficients were obtained using the following method:

For each target language and text type we took the two least controversial ‘highest scoring’ and ‘lowest scoring’ systems. By ‘least controversial’ we mean the system closest to the top (or to the bottom) of the rankings according to the human evaluation results and both automated evaluation scores – BLEU and LTV. Human scores for these ‘least controversial’ systems are highlighted in **blue** in the following table. The values of E_1 and E_2 (automated evaluation scores for each system), X_1 and X_2 (human evaluation scores for each system) were taken from our data for these systems. Then coefficients a and b were computed by solving the system of 2 equations:

- $a * E_1 + b = X_1$
- $a * E_2 + b = X_2$

6.3 Application of the coefficients

Scores for all other systems were extrapolated using formula (1). It can be seen from Table 18 that most of the extrapolated scores (given in grey) are very close to real human scores. The higher the correlation between human and automated scores, the more accurate is the prediction.

In order to compute a human score X the following formula should be used:

$$(1) \quad X = a * E + b \quad \text{where}$$

- E is the automated score known to the evaluator
- a and b are coefficients which depend on the text type, target language and the type of automated score used – BLEU or LTV. These coefficients are given in Table 18.
- X is the projected human evaluation score which we are looking for.

For example, to compute an extrapolated human evaluation score for a translation of **email** text into German, which has an LTV adequacy score of 0.2759, we need to consult the first section of the table (which contains the data: **TL = de, LTVadeEm, BleuEm**), where we find the a and b coefficients for translations of Emails into German for LTV scores:

$$a = 3.71; b = 2.48 \text{ (approximately)}$$

Applying formula (1)

$$0.2759 * 3.71 + 2.48 = 3.501$$

we obtain a result which is very close to the observed human score of 3.503.

Sys	SL	TL		hAveEm	clbLTV	clbBleu	hPassEm	LTVadeEm	BleuEm		hAveWp	clbLTV	clbBleu	hPassWp	LTVadeWp	BleuWp	
1	s06	fr	de	1	3.665	3.462	3.359	3690	0.2653	0.1496	1	3.818	4.3887	3.9905	5670	0.2061	0.1314
2	s05	en	de	2	3.602	3.601	3.5664	2130	0.3029	0.236	3	3.342	3.3411	3.3392	-3780	0.1386	0.0762
3	s06	en	de	3	3.503	3.501	3.4726	-1020	0.2759	0.1969	2	3.469	3.4264	3.1256	-900	0.1441	0.0581
4	s03	it	de	4	3.184	3.183	3.1546	-10500	0.1901	0.0644	4	2.707	2.7063	2.7043	-16770	0.0977	0.0224
								correl=	0.8823805	0.7694					0.9486933	0.9168	
								a=	3.7056738	2.4359					15.525672	11.803	
								b=	2.4795514	3.0271					1.1901418	2.4426	
Sys	SL	TL		hAveEm	clbLTV	clbBleu	hPassEm	LTVadeEm	BleuEm		hAveWp	clbLTV	clbBleu	hPassWp	LTVadeWp	BleuWp	
1	s05	de	en	1	4.383	4.121	4.1835	24900	0.4213	0.3207	8	4.071	3.0218	3.1231	10890	0.2055	0.1258
2	u05	fr	en	2	4.247	4.236	4.2495	21960	0.4446	0.3339	1	4.589	4.5847	4.5777	20550	0.433	0.3354
3	s06	de	en	3	4.194	4.018	4.0555	19500	0.4005	0.2951	7	4.153	3.2038	3.2036	11790	0.232	0.1374
4	s05	fr	en	5	4.151	3.775	3.8175	17250	0.3513	0.2475	4	4.273	3.3722	3.6394	14160	0.2565	0.2002
5	s05	es	en	4	4.151	3.756	3.67	17610	0.3473	0.218	5	4.242	3.4793	3.4312	14040	0.2721	0.1702
6	s06	fr	en	6	4.08	3.976	4.011	16710	0.392	0.2862	2	4.347	3.7335	3.8059	16170	0.3091	0.2242
7	s06	es	en	7	3.902	3.619	3.5595	11310	0.3196	0.1959	9	4.018	3.3467	3.35	9480	0.2528	0.1585
8	u03	fr	en	8	3.845	3.957	3.9095	9420	0.388	0.2659	3	4.338	4.1079	3.8948	15810	0.3636	0.237
9	s03	it	en	9	3.746	3.382	3.24	6870	0.2716	0.132	15	2.907	3.0005	2.926	-12240	0.2024	0.0974
10	s04	fr	en	10	3.689	3.667	3.7095	5010	0.3294	0.2259	6	4.224	4.134	4.1557	13620	0.3674	0.2746
11	s04	es	en	11	3.447	3.33	3.436	-2190	0.2612	0.1712	11	3.927	3.8826	3.8025	7830	0.3308	0.2237
12	s03	fr	en	12	3.423	3.513	3.45	-2460	0.2982	0.174	14	3.131	3.0019	3.1196	-7950	0.2026	0.1253
13	s03	es	en	13	3.294	3.284	3.296	-6690	0.2518	0.1432	13	3.147	3.1427	3.1369	-7380	0.2231	0.1278
14	s06	it	en	14	3.25	3.451	3.453	-8010	0.2856	0.1746	10	3.971	3.2169	3.1494	9000	0.2339	0.1296
15	s06	pt	en	15	3.124	3.559	3.6055	-11730	0.3075	0.2051	12	3.711	3.1599	3.0939	3900	0.2256	0.1216
								correl=	0.8215176	0.7699					0.6742491	0.7086	
								a=	4.9429461	4.9974					6.8699381	6.9461	
								b=	2.0493662	2.5784					1.6143168	2.2593	

Sys	SL	TL		hAveEm	clbLTV	clbBleu	hPassEm	LTVadeEm	BleuEm		hAveWp	clbLTV	clbBleu	hPassWp	LTVadeWp	BleuWp	
1	s06	fr	es	1	3.618	3.308	3.2663	3060	0.246	0.172	1	4.456	10.971	10.449	18150	0.5771	0.557
2	s05	en	es	2	3.379	3.377	3.3732	-4410	0.26	0.1988	2	3.696	3.6854	3.6928	3240	0.2785	0.2158
3	s03	en	es	3	3.149	3.147	3.1434	-11520	0.2136	0.1412	3	3.498	3.4878	3.4948	-540	0.2704	0.2058
4	s06	en	es	4	3.126	3.283	3.3481	-12090	0.241	0.1925	4	3.46	3.0852	3.3543	-1260	0.2539	0.1987
5	s04	en	es	5	2.49	3.204	3.2152	-31920	0.225	0.1592	5	3.171	4.8664	4.1997	-7110	0.3269	0.2414
								correl=	0.5674152	0.3539						0.8486513	0.8909
								a=	4.9568966	3.9931						24.444444	19.8
								b=	2.0902069	2.5852						-3.111778	-0.5768
Sys	SL	TL		hAveEm	clbLTV	clbBleu	hPassEm	LTVadeEm	BleuEm		hAveWp	clbLTV	clbBleu	hPassWp	LTVadeWp	BleuWp	
1	s06	en	fr	1	3.974	3.913	3.7605	13620	0.3079	0.2414	5	3.924	3.5013	3.655	8010	0.3285	0.2551
2	u05	en	fr	2	3.846	3.846	3.8357	9150	0.3024	0.2498	2	4.298	4.2951	4.2964	14700	0.4276	0.3321
3	u03	en	fr	3	3.811	3.707	3.5941	8460	0.2909	0.2228	4	4.118	3.925	3.71	11190	0.3814	0.2617
4	s06	de	fr	4	3.654	3.301	3.0821	3480	0.2573	0.1656	7	3.882	3.5085	3.4692	6660	0.3294	0.2328
5	s05	en	fr	5	3.649	3.708	3.7212	3840	0.291	0.237	6	3.902	3.4708	3.5792	7170	0.3247	0.246
6	s03	it	fr	6	3.45	2.699	2.8512	-3000	0.2075	0.1398	10	3.436	3.9547	4.3464	-1950	0.3851	0.3381
7	s06	it	fr	7	3.446	3.228	3.117	-2940	0.2513	0.1695	8	3.88	4.3552	4.6296	6900	0.4351	0.3721
8	s06	es	fr	8	3.377	3.181	3.0436	-4800	0.2474	0.1613	1	4.562	5.1634	5.4851	19920	0.536	0.4748
9	s04	en	fr	9	3.351	3.24	3.399	-5760	0.2523	0.201	9	3.62	4.066	4.0273	1740	0.399	0.2998
10	s06	pt	fr	10	3.303	3.222	3.1931	-6690	0.2508	0.178	3	4.204	4.8678	5.0586	12900	0.4991	0.4236
11	s03	en	fr	11	2.854	2.855	2.8441	-20730	0.2204	0.139	11	2.647	2.6442	2.6471	-17730	0.2215	0.1341
								correl=	0.8201731	0.7732						0.7883126	0.7182
								a=	12.097561	8.9531						8.0106744	8.3384
								b=	0.1876976	1.6095						0.8726356	1.5288

Sys	SL	TL		hAveEm	clbLTV	clbBleu	hPassEm	LTVadeEm	BleuEm		hAveWp	clbLTV	clbBleu	hPassWp	LTVadeWp	BleuWp	
1	s06	fr	it	1	3.743	3.742	3.7375	6240	0.253	0.1799	1	4.5	4.4984	4.5001	18300	0.4848	0.4348
2	s03	es	it	2	3.705	2.874	2.9362	4590	0.2049	0.1166	3	3.667	2.9666	3.1127	2580	0.2761	0.2028
3	s03	fr	it	3	3.598	3.363	3.4046	1650	0.232	0.1536	2	3.902	3.6837	3.7969	6870	0.3738	0.3172
4	s06	en	it	4	3.333	3.157	3.4147	-6270	0.2206	0.1544	4	3.611	2.6737	2.7354	1410	0.2362	0.1397
5	s03	en	it	5	3.282	2.818	2.9412	-7860	0.2018	0.117	5	2.964	2.2612	2.437	-11310	0.18	0.0898
6	s03	de	it	6	2.551	2.55	2.5462	-30150	0.1869	0.0858	6	2.287	2.2884	2.2875	-25470	0.1837	0.0648
								correl=	0.7344635	0.74					0.8872925	0.9064	
								a=	18.033283	12.667					7.3497177	5.9811	
								b=	-0.819421	1.4641					0.9368569	1.8994	
Sys	SL	TL		hAveEm	clbLTV	clbBleu	hPassEm	LTVadeEm	BleuEm		hAveWp	clbLTV	clbBleu	hPassWp	LTVadeWp	BleuWp	
1	s05	en	pt	1	3.409	4.136	4.0464	-3930	0.2703	0.2076	2	3.771	3.3715	3.3305	4710	0.201	0.1341
2	s06	fr	pt	2	3.377	3.375	3.378	-4620	0.2353	0.1616	1	4.262	4.2572	4.2527	14460	0.4512	0.4214
3	s06	en	pt	3	3.114	3.112	3.1151	-13080	0.2232	0.1435	3	3.196	3.1899	3.1892	-6450	0.1497	0.0901
								correl=	0.7659626	0.7833					0.9173685	0.902	
								a=	21.735537	14.53					3.5356551	3.2176	
								b=	-1.737372	1.0289					2.6667124	2.9061	

Table 18: Calibration table for automated scores