a folha

Boletim da língua portuguesa nas instituições europeias

http://ec.europa.eu/translation/portuguese/magazine

English version of an article published in No. 37 — Autumn 2011

MACHINE TRANSLATION: CASE STUDY – ENGLISH INTO PORTUGUESE — EVALUATION OF MOSES IN DGT PORTUGUESE LANGUAGE DEPARTMENT USING MOSES FOR MERE MORTALS

Maria José Machado and Hilário Leal Fontes
Portuguese Language Department — Directorate General for Translation
European Commission

[click here for Portuguese version]

Abstract
The Portuguese Language Department (PTLD) of DG Translation of the European Commission has been testing the English-Portuguese language pair using open source Moses - installed and run with the set of scripts Moses for Mere Mortals - to evaluate its usefulness for professional translators (publishing) but also for general users (gist).
Moses for Mere Mortals (MMM) is an open-source application which builds a prototype of a real world translation chain enabling a fairly easy use of Moses, therefore making Moses available to a wider number of users.
In this paper we present the results obtained with a 12.4 million segment corpus and 2 Testing Sets with a total of 136 documents (about 1 million words). Four engines (both un-tuned and tuned) were trained and the output evaluated by automatic (BLEU and NIST) metrics and human evaluation.
For automatic evaluation, 10 parameter variants were selected for evaluation of the 8 engines. For human evaluation, 5 variants of Moses output were selected and 16,500 individual judgements for translations purposes (scores 1 to 5) and 16,500 for gist purposes (Yes/No) were collected from 11 evaluators.
The results are very promising even with these baseline engines, and their output has been used (as a CAT tool) for about a year and a half in the workflow of the PTLD.

1. Introduction

The European Commission supports and encourages the collaborative development and re-use of publicly-financed free and open source software (F/OSS) applications in European public administrations through its Open Source Observatory and Repository for European Public Administrations[1] and now, with a broader scope, through its Digital Agenda[2]. The EU Research and Technological Development Framework Programmes have been supporting research in Machine

[1] http://www.osor.eu/
[2] http://ec.europa.eu/information_society/digital-agenda/index_en.htm

Translation and, namely, the EuroMatrix(Plus) Project[3] which has made Moses available under a LGPL licence[4]. In the last decade, research has progressed with the development of Statistical Machine Translation, an approach that relies on corpora for the training of the system.

The EU 50 years' policy of multilingualism has produced a large body of high quality multilingual corpora and the European Commission has invested in, and been a user of, Machine Translation for the past 35 years. MT as a computer-assisted translation tool (CAT) can greatly assist the Commission in the fulfilment of its mission of treating all languages equally in its two-way communication with European citizens and companies. EURAMIS is the Translation Memory system where the alignments of texts translated by the EU institutions are stored. These corpora with millions of segments per language can therefore be easily used to train SMT language pairs.

2. Context

The authors have worked for more than 20 years in the field of translation and used and contributed to the improvement of the rule-based ECMT system for about 10 years. This evaluation was therefore carried out in a pragmatic way and from a translator's perspective as the authors have no formal NLP background – the authors have a language and translating background and the developer of the MMM scripts is a former translator and a trained physician.

3. Installation and running of Moses with the Moses for Mere Mortals scripts

Moses was installed using the set of scripts Moses for Mere Mortals (MMM) developed by João Rosas, with the collaboration of the authors as testers, and published under a GPL licence[5]. The MMM version used was the one published in Moses Sourceforge website which installs the Moses version released on August 14, 2010. For this Case Study we used a 4-processor PC with 8GB RAM.

MMM builds a translation chain prototype with Moses + IRSTLM + RandLM and MGIZA. These scripts do not cover factored training. The MMM scripts run on Linux (Ubuntu distribution) and automate the installation, the creation of a representative set of test files, the training, the translation and the scoring tasks.

MMM's main aims are: 1) To help build a prototype of a translation chain for the real world; 2) To guide the first steps of users who are just beginning to use Moses by providing an easily understandable Help-Tutorial, as well as a Quick-Start-Guide and a Demo; 3) To train large corpora; 4) To translate documents (in batch mode); 5) To enable a simple and quick evaluation of Moses with the automatic metrics BLEU and NIST (in batch mode), both for the whole document and line-by-line (very useful for quick human evaluation of best/worst BLEU/NIST scored segments); 6) To integrate machine translation and translation memories.

MMM consists of 6 scripts: *create* (in order to compile Moses and the packages it uses with a single command), *make-test-files, train, translate, score* and *transfer-training-to-another-location*. Included in MMM are two "Windows add-ins" - Extract_TMX_Corpus (ETC) and Moses2TMX - to complete the full chain from original Word documents to TMX files for import into the translation memory tool used by the translator.

---

[3] Euromatrix (2006-2009): http://www.euromatrix.net/; EuromatrixPlus (2009-2012): http://www.euromatrixplus.net/

[4] http://www.statmt.org/moses/

[5] https://github.com/moses-smt/mosesdecoder/tree/master/contrib/moses-for-mere-mortals; http://code.google.com/p/moses-for-mere-mortals/; http://en.wikipedia.org/wiki/Moses_for_Mere_Mortals

MMM also includes a "Nonbreaking_prefix file for the Portuguese language", a list of the main abbreviations for Portuguese. The MMM scripts enable the tailoring of the main Moses parameters (about 80) to particular language pairs and corpora/documents.
In this paper we present a Case Study which shows how mere mortals can use Moses.

4. Training and tuning data

For training purposes, a 12.4 million corpus (12M corpus), extracted from DGT's Euramis database by DGT Informatics Unit, was used. The corpus contains all DGT's translations of Commission documents and all the legislation and case-law aligned and stored up to November 2009.

This English-Portuguese corpus contains 468.9 million words (EN+PT) and 12,468,232 bilingual segments and it was cleaned of control characters, segments where source and target were identical and sentence pairs with larger than 4:1 token ratios. No merge of identical segments was performed.

The PT side of the corpus was used for language model training using the IRSTLM and RANDLM language models. The 800 and 2,000 segment tuning corpora used were composed of extractions of segments from documents of a large variety of domains/Directorates-General selected for their quality and not contained in the training corpus.

5. Engines

Four engines were trained with different training parameters and subsequently tuned using the 800 or 2,000 segment corpora (Table 1) with default settings, except when otherwise indicated[6].

Engine ID	LM	n-gram	Other non-default settings
E1	IRSTLM	7-gram	Tuning: no
E1t	RSTLM	7-gram	Tuning: 800
E2	IRSTLM	7-gram	Smoothing: improved Kneser Ney; Tuning: no
E2t	IRSTLM	7-gram	Smoothing: Improved Kneser Ney; Tuning: 800
E3	RANDLM	7-gram	Tuning: no
E3t	RANDLM	7-gram	Tuning: Corpus 2,000
E4	RANDLM	9-gram	MaxLen=80; Tuning: no
E4t	RANDLM	9-gram	MaxLen=80; Tuning: 2,000

Table 1. Engines trained with MMM default parameters, except when otherwise indicated.

6. Test sets

We tested 2 sets of documents (Table 2). Set 1 contained 88 documents which were evaluated individually by 32 translators of the PTLD as to their general usefulness for translation work. This evaluation was carried out during a 3-month period during which all translators could request a Moses translation to be used instead of our then available rule-based system (ECMT).

Therefore, these documents were not chosen according to specific criteria. They cover a wide range of Directorates-General/domains (20) and the Moses output used for these translations was from a previous engine trained with a 6.6 million segment corpus containing documents translated in DGT in a shorter period of time and trained with IRSTLM (Witten-Bell smoothing). Set 2 contained 48 documents selected from documents translated by colleagues who were not MT users at the time (11 translators).

---

[6] See MMM defaults parameters in Help-Tutorial doc at https://github.com/moses-smt/mosesdecoder/tree/master/contrib/moses-for-mere-mortals/docs).

The only other criterion was to cover a wide range of DGs/domains (19). Both sets contained documents representative of our work (both legislative and non-legislative documents), namely regulations, decisions, recommendations, communications, reports, the Commission General Report, opinions, staff working documents, memoranda, programmes, etc..

Document sets	No. pages (internal)	No. words	No. segments	Average no. words/segm	100% match with the TC (segm)	100% match percentage (segm)
Test Set 1	2,594	675,313	34,179	19,8	8,249	24,1%
Test Set 2	1,250	349,026	17,234	20,2	3,635	21,1%
Total	3,844	1,024,339	51,413	19,9	11,884	23,1%

Table 2. Sets of documents used for testing

7. Testing with different translation parameters

MMM *translate* script allows an easy definition of 17 translation parameters which may have a significant impact on the quality of the output. Various combinations of these parameters were tested with a small sample to determine if there was an improvement in Moses performance. After preliminary tests with different parameter combinations, 10 parameters were selected for further testing: weight_t, weight_l, weight_d, weight_w, mbr, searchalgorithm, cubepruningpoplimit, stack, maxphraselength and distortionlimit (Table 3).

Variants	MMM DEFAULTS. EXCEPT:									
	wp	*wd*	*wl*	*wt*	*searchalg*	*cubeprun*	*stack*	*mpl*	*mbr*	*distortion limit*
Var. A	-1.3	—	—	—	—	—	—	—	—	—
Var. 1	-1	0.5	0.9	1.5	1	2000	2000	30	1	7
Var. 1A	-1.3	0.5	0.9	1.5	1	2000	2000	30	1	7
Var. 1B	-1.6	0.5	0.9	1.5	1	2000	2000	30	1	7
Var. 1C	-2	0.5	0.9	1.5	1	2000	2000	30	1	7
Var. 1D	-2.5	0.5	0.9	1.5	1	2000	2000	30	1	7
Var. 2	-1	0.5	0.9	1.5	1	2000	2000	30	1	9
Var. 2A	-1.3	0.5	0.9	1.5	1	2000	2000	30	1	9
Var. 2B	-1.6	0.5	0.9	1.5	1	2000	2000	30	1	9

Table 3. Translation parameter combinations tested with Test Set 2.

8. Automatic Evaluation

The 2 Test Sets were translated by their translators and scored without the elimination of segments having a high match rate with segments in the training corpus.

The MMM *score* script was used to obtain BLEU and NIST scores for those documents globally and individually as it is important for us to have an idea of Moses' performance with very different types of documents. Therefore Moses translations of the individual documents were made and scored and afterwards the files were merged and scored again in order to obtain global scores.

The best BLEU score was obtained with E2-Var.1 in both Test Sets. The best NIST score was obtained with E4-Var.1 and E2-Def. The scores show maximum differences of up to 9.44 BLEU points, depending on the training and translation parameters used. Default parameters have

consistently yielded lower BLEU scores than some of the variants tested, of the order of -2.25 BLEU points to -9.44 BLEU points compared to the best performing variant (E2-Var.1).
Simply changing the word penalty from 0 (default) to values between -0.5 to -1.5 produced better BLEU scores, which seems logical as English is a more synthetic language than Portuguese. Changing the word penalty and some other parameters also produced somewhat better BLEU scores.

Human evaluation confirmed the lower performance of all the tuned engines, as well as of the RANDLM trained engines, not only in this evaluation as well as in other non-structured evaluations carried out with individual documents.

ENGINE/ VARIANT	LANGUAGE MODEL	Test Set 1						Test Set 2					
		BLEU	NIST	BLEU Dif. ***	NIST Dif ***	BLEU rank	NIST rank	BLEU	NIST	BLEU Dif ***	NIST Dif ***	BLEU rank	NIST rank
E1-Def **	IRSTLM-WB	49,5	11,689	-3,1	-0,006	11	7	46,28	10,806	-2,26	0,1062	17	6
E1-Var.1 *		52,1	11,698	-0,5	0,0028	3	5	48,5	10,727	-0,04	0,0271	2	9
E1-Var.A			-					48,06	10,616	-0,48	-0,083	5	12
E1t-Def	IRSTLM-WB-t800	47,1	11,133	-5,6	-0,562	14	14	44,16	10,314	-4,38	-0,386	24	21
E1t-Var.1								41,44	9,3611	-7,1	-1,339	28	29
E2-Def **	IRSTLM-IKN	50,4	11,793	-2,3	0,0981	8	2	47,07	10,888	-1,47	0,1887	14	1
E2-Var.1 *		52,6	11,695			1	6	48,54	10,7			1	10
E2-Var.A *		52,4	11,615	-0,2	-0,08	2	8	48,01	10,584	-0,53	-0,116	7	15
E2-Var.1B								46,56	10,317	-1,98	-0,383	16	20
E2t-Def *	IRSTLM-IKN-t800	48,7	11,32	-3,9	-0,375	12	12	45,54	10,432	-3	-0,268	21	18
E2-Var.1								40,62	9,1721	-7,92	-1,528	29	30
E3-Def **	RANDLM-7g	48,4	11,436	-4,2	-0,26	13	11	45	10,544	-3,54	-0,156	22	16
E3-Var.A			-					46,94	10,396	-1,73	-0,309	15	19
E3-Var.1		51,4	11,58	-1,2	-0,115	6	9	47,74	10,59	-0,93	-0,115	10	13
E3-Var.1A								47,17	10,47	-1,37	-0,23	13	17
E3t-Def	RANDLM-7g-t2000	46,4	10,963	-6,2	-0,733	15	16	43,4	10,135	-5,14	-0,565	26	25
E3t-Var1			-					43,78	9,8432	-4,76	-0,856	25	27
E4-Def **	RANDLM-9g	46,2	10,976	-6,5	-0,72	16	15	43,01	10,146	-5,53	-0,554	27	24
E4-Var.1		51,4	11,826	-1,2	0,1309	7	1	47,73	10,867	-0,81	0,1669	11	3
E4-Var.A		49,7	11,554	-3	-0,141	9	10	46,23	10,648	-2,31	-0,051	18	11
E4-Var.1A								48,03	10,827	-0,51	0,1274	6	5
E4-Var.1B		52	11,728	-0,6	0,033	4	4	48,25	10,751	-0,29	0,0514	3	8
E4-Var.1C								47,87	10,585	-0,67	-0,114	9	14
E4-Var.1D								46,03	10,233	-2,51	-0,467	19	23
E4-Var.2								47,67	10,874	-0,87	0,1741	12	2
E4-Var.2A								47,97	10,843	-0,57	0,1433	8	4
E4-Var.2B		51,9	11,739	-0,7	0,0434	5	3	48,24	10,774	-0,3	0,0747	4	7
E4t-Def **	RANDLM-9g-t800	43,2	10,289	-9,4	-1,406	17	17	40,18	9,4847	-8,36	-1,215	30	28
E4t-Var.A								44,59	9,9333	-3,95	-0,766	23	26
E4t-Var.1		49,7	11,206	-3	-0,49	10	13	46,03	10,277	-2,51	-0,423	20	22

* Variant with human evaluation of 300 segments by 11 evaluators;
** Variant with preliminary human evaluation of 125 segments in 13 variants by one evaluator;
*** BLEU/NIST score difference to best scoring engine (E2-Var.1)

Table 4. Global BLEU and NIST scores for Test Sets 1 and 2

The results obtained with the BLEU and NIST metrics differ significantly as can be seen by the results (with ranking) presented in Table 4. The human evaluation of the 300 segment sample (and other non-structured evaluations) corroborated BLEU scores globally and by document. We also used the *score-line-by-line* script to evaluate Moses output in selected documents and we observed that, at segment level, BLEU scores are not so "reliable", but even so they help us to detect problems more quickly and easily.

9. Human evaluation in real-life conditions

An evaluation in real-life conditions of Test Set 1 was carried out in which the 32 translators who participated were asked to give feedback concerning the usefulness of Moses for their work regarding each document. A free text comment and a Yes/No evaluation was requested. Only 2 non-MT users (at the time) considered Moses output not useful for specific documents (BLEU with E2-Var.1: 37.81 and 35.39), although some other translators considered Moses output useful even with similarly low score documents.

As was to be expected, the documents with a higher number of 100% match sentences had better scores, which confirms that Moses really "learns" from the data it is fed with. However, some documents with a very low 100% match also showed reasonably good scores (BLEU: ≥ 45). Detailed results by document are not presented in this paper. However, in Table 5 some figures are shown concerning scores by document obtained with the best performing engine (E2-Var.1) (which was not the Moses engine whose output was used by the translators for the translation of those documents).

BLEU	**Test Set 1**	**Test Set 2**
Global score	52.63	48.54
Highest score	74.79	80.22
Lowest score	30.83	28.98
Scores ≥ 50.00	56 docs (1,577 pp)	16 docs (328 pp)
Scores 40.00-49.99	27 docs (836 pp)	17 docs (506 pp)
Scores < 40.00	5 docs (181 pp)	15 docs (416 pp)

Table 5. Range of BLEU scores for Test Sets 1 and 2, by document.

10. Human evaluation of a 300 segment sample

A structured evaluation was carried out with a 300 segment sample extracted from both sets of documents using MMM *make-test-files* script. Before extracting the segments we ran a script prepared by Michael Jellinghaus (*filtersentences.perl*) to eliminate from the Test Sets the segments with a 100% match in the training corpus. We then ran the *make-test-files* script which divided Test Set 1 file in 80 sectors and Test Set 2 in 40 sectors to extract pseudo-randomly 3 segments/sector. An extra number of segments was extracted in order to eliminate those with no translation content (Official Journal references, numbers and short titles), thus creating a test sample of 300 segments (200 from Set 1 and 100 from Set 2).

One of the authors carried out a preliminary evaluation of 13 variants (among the best and the worst BLEU scoring (identified with (*) and (**) in Table 4) with 125 segments of this sample in order to select the most useful for evaluation by a larger number of fellow translators. This preliminary evaluation corroborated the BLEU scores.

Considering that BLEU is the most widely used metrics and that our previous human evaluations had definitely established that word penalty values between -0.5 and -1.5 produced better results, we decided to “trust” BLEU. Therefore, 5 variants were chosen for evaluation by 11 translators: 4 from

the best scoring variants of non-tuned engines and the best scoring of the tuned engines with the remaining translation default settings. The results are presented in Table 6.

In the evaluation of the 300 segments by 11 translators, our objective was to evaluate segments as they appear in our daily work, i.e., without any selection/limitation by sentence length, complexity, technicality or any other criteria (average of 23 words/segment). The design of this evaluation was mainly based on evaluations performed by the EuroMatrix(Plus) Project with adaptations to our context, purpose and resources. We know from experience that it is difficult to evaluate long segments with two or more clauses with different translations/mistakes in different Moses outputs, but those are the kind of segments we have to deal with in our daily work.

As the documents of these two Test Sets covered very different domains and many were very technical, besides the original and the 5 Moses outputs, a reference translation was included in the evaluation table provided to the evaluators since they were not specialised in all these domains. As these segments were randomly taken from the 2 Test Sets with about 70,000 segments, there is no context. This is a limitation that could not be avoided in this case.

The segments were evaluated in terms of their acceptability for translation and gist purposes. The evaluation for gist purposes is only indicative, considering it was not performed in lab conditions, as the evaluators had the original and reference translation (and some had even translated some of those documents) and this can have an impact on their evaluation. On the whole, 16,500 individual judgements for translations purposes (scores 1 to 5) and 16,500 for gist purposes (Yes/No) were collected for the selected 5 Moses variants.

10.1 Evaluation criteria (provided to the evaluators)

A. Scoring of segments according to their acceptability for translation work

This evaluation aims to evaluate the acceptability of Moses output (a possible translation), even if it is not the choice a particular translator would have made for her/his own translation. Several variants may have the same score. The objective is to classify those variants in a scale of 1 to 5, first of all as to the quality level for translation purposes. Score 5 should be reserved for translations that could be used without any change, from a linguistic and content point of view.

1 – **Bad:** Many changes for an acceptable translation; no time saved.

2 – **So so:** Quite a number of changes, but some time saved.

3 – **Good:** Few changes; time saved.

4 – **Very good:** Only minor changes, a lot of time saved.

5 – **Fully correct:** Could be used without any change, even if I would still change it if it was my own translation.

B. Verdict (Yes-1; No-0) as to the acceptability of each sentence for gist (assimilation) purposes

Evaluate if the full meaning of the segment can be understood, even if the sentence is not correct/fluent from a linguistic point of view. If a translation contains one or more words in the original language which should be translated, the verdict should be "No" (0) as, for assimilation purposes, we have to consider that the user may not understand a single word of the source language (although this is not the case in the present evaluation with EN-PT).

10.2 Results

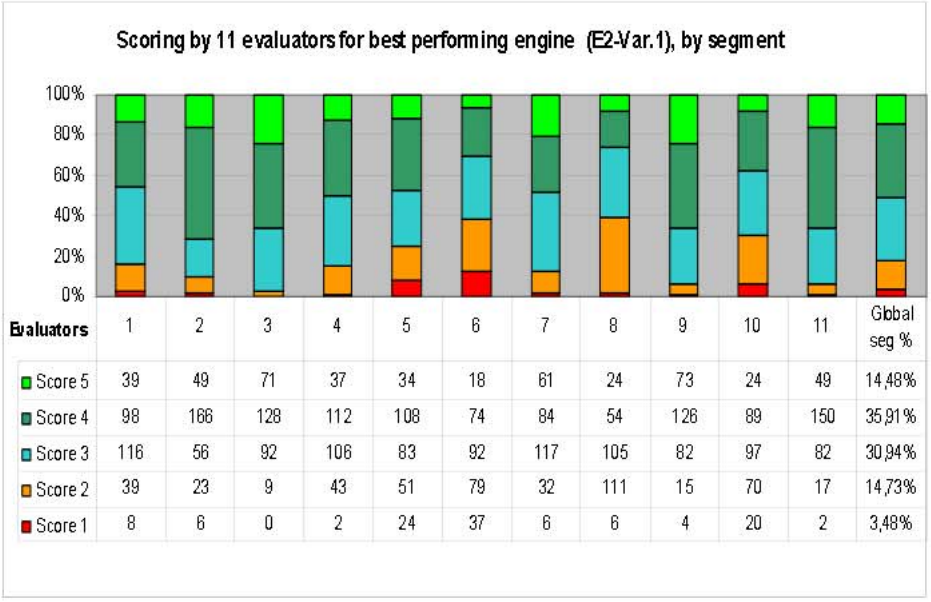
The global results are presented in Table 6 for the 5 engines/variants evaluated, for translation and gist purposes, with a global percentage per evaluator and variant.

	E2-Var.1		E2-Var.A		E2t-Def.		E1-Var.1		E4-Var.1B	
BLEU score – Test Set 1	52,63		52,43		48,73		52,09		52,00	
BLEU score – Test Set 2	48,54		48,01		45,54		48,50		48,25	
BLEU score-300 segments	46,79		45,63		41,98		46,68		44,99	
	T* (%)	**G* (%)**	**T (%)**	**G (%)**	**T (%)**	**G (%)**	**T (%)**	**G (%)**	**T (%)**	**G (%)**
Human evaluation average (% of points) – Global	68.69	68.45	68.19	66.94	63.62	59.48	67.30	66.58	65.65	62.73
Evaluator 1	68.07	69.67	67.47	69.67	62.93	58.33	67.00	67.67	65.53	63.33
Evaluator 2	75.27	69.33	78.80	66.00	70.93	58.33	79.20	65.67	72.93	60.00
Evaluator 3	76.20	74.67	75.80	72.67	72.33	66.33	75.07	71.00	73.13	66.33
Evaluator 4	69.27	52.00	68.60	49.67	65.80	45.33	67.60	49.00	67.27	48.00
Evaluator 5	65.13	73.33	61.67	71.00	54.80	61.67	59.33	70.67	62.53	69.33
Evaluator 6	57.13	48.00	55.53	44.67	52.60	37.00	55.33	45.67	53.87	42.67
Evaluator 7	70.80	68.67	70.13	65.00	63.47	51.67	70.13	67.33	67.53	62.00
Evaluator 8	58.60	87.00	61.20	89.33	57.00	87.33	58.67	90.33	58.87	89.33
Evaluator 9	78.20	76.33	75.40	75.33	70.20	67.00	74.33	79.67	70.93	69.00
Evaluator 10	61.80	68.33	59.93	66.00	56.00	59.67	57.53	62.67	55.20	58.67
Evaluator 11	75.13	61.00	75.60	67.00	73.73	61.67	76.13	62.67	74.40	61.33

* T — Translation; G — Gist

Table 6. Results of the human evaluation for translation and gist purposes of a 300 segment sample extracted from Test Sets 1 and 2.

In Table 7 are presented the results, by score, of the best performing engine/variant (E2-Var.1), which had a global percentage of 68.69 and 68.45 for translation and gist purposes, respectively.



Evaluators	1	2	3	4	5	6	7	8	9	10	11	Global seg %
Score 5	39	49	71	37	34	18	61	24	73	24	49	14,48%
Score 4	98	166	128	112	108	74	84	54	126	89	150	35,91%
Score 3	116	56	92	106	83	92	117	105	82	97	82	30,94%
Score 2	39	23	9	43	51	79	32	111	15	70	17	14,73%
Score 1	8	6	0	2	24	37	6	6	4	20	2	3,48%

Table 7. Results, by translator and by score, of the evaluation for translation purposes for the best scoring engine/variant (E2-Var.1).

11. Conclusions

Moses open-source MT system – as installed with the Moses for Mere Mortals scripts – proved to be reliable and robust, once the MMM scripts were duly tested, translating 81,000 pages (about 12 million words – 1 million segments) for this Case Study without any problems. Our main concern is quality, not speed, but we have also taken that factor into account and this explains our interest in the RANDLM Language Model. KenLM was not tested in this Case Study.

We were surprised with the consistently worst results obtained with the tuned engines, as shown by both automatic and human evaluation. As our approach is pragmatic and we obtained satisfying results with Variants 1 and A of the translation parameters, we introduced Variant 1 in our workflow, since it was evaluated as slightly better than Variant A in terms of quality and is better in terms of speed in the translation process.

Concerning human evaluation, and bearing in mind we were only testing baseline engines, the level of fluency and terminology accuracy was surprising high and mentioned by the vast majority of translators/evaluators. In practical terms, the use of MT (which was already high as in 2009 85% of the Portuguese translators reported using ECTM, at least for certain jobs) has increased with Moses and, in general, there is a high level of satisfaction among translators as to its usefulness, as confirmed by the global results obtained concerning acceptability for translation (above 60% with all the 5 engines/variants evaluated).

Inter-evaluator consistency was high in terms of global ranking (but not of level), as the engines/variants considered as best and worst performing were generally consistent (9 evaluators agreeing for each of them). The distribution by evaluator presented in Table 7 shows a high degree of variation between scores, reflecting different individual perceptions of usefulness. Although the evaluation for gist purposes is only indicative, it correlates well with the evaluation for translation purposes.

We have not carried out a thorough statistical analysis of the data presented in this Case Study, nor a segment level analysis/statistics, as our main interest is the usefulness of Moses output as a CAT tool for interactive translation using translations memories (TM) combined with MT. There seems to be a good correlation between automatic and human evaluation in the sense that the best and worst performing engines/variants in terms of BLEU scores have been corroborated by human evaluation in general.

12. Acknowledgments

We would like to thank our colleague João Rosas who designed and prepared the scripts which enabled us to carry out this study. We would also like to thank the 32 translators from the three Portuguese Units who actively participated in these evaluations and who, together with our other colleagues, are continuously giving us feedback on Moses performance. We would also like to thank Michael Jellinghaus, of DG TRAD of the European Parliament, for the *filtersentences.perl* script, which allowed us to eliminate 100% matches when preparing the 300 segment sample for human evaluation.

And, above all, we thank the Moses developers around the world who have contributed to the development of open-source Moses.

13. References

— Aziz, Wilker F.; Pardo, Thiago A. S.; Paraboni, Ivandré, "Fine-tuning in Portuguese-English Statistical Machine Translation". *Proceedings of the 2009 7th Brazilian Symposium in Information and Human Language Technology — STIL 2009*, São Carlos, 2009, http://portalsbc.sbc.org.br/download.php?paper=2816.

— Boyer, Vivian, "Human Evaluation of Machine Translation Quality: a Linguistic Oriented Approach", 2010.

— Callison-Burch, Chris; Koehn, Philipp; Monz, Christof; Peterson, Kay; Przybocki, Mark; Zaidan, Omar F., "Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation". *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, Uppsala, 2010, http://www.statmt.org/wmt10/pdf/wmt10-overview.pdf.

— Callison-Burch, Chris; Koehn, Philipp; Monz, Christof; Schroeder, Josh, "Findings of the 2009 Workshop on Statistical Machine Translation". *Proceedings of the 4th EACL Workshop on Statistical Machine Translation*, European Chapter of the Association for Computational Linguistics, Athens, 2009, pp. 1-28, http://homepages.inf.ed.ac.uk/pkoehn/publications/wmt09-overview.pdf.

— Callison-Burch, Chris; Fordyce, Cameron; Koehn, Philipp; Monz, Christof; Schroeder, Josh, "Further Meta-Evaluation of Machine Translation". *Proceedings of the Third Workshop on Statistical Machine Translation*, Association for Computational Linguistics, Columbus, 2008, pp. 70-106, http://aclweb.org/anthology-new/W/W08/W08-0309.pdf.

— Callison-Burch, Chris; Fordyce, Cameron; Koehn, Philipp; Monz, Christof, Schroeder, Josh, "(Meta-) Evaluation of Machine Translation". *Proceedings of the Second Workshop on Statistical Machine Translation*, Association for Computational Linguistics, Prague, 2007, pp. 136-158, http://www.statmt.org/wmt07/pdf/WMT18.pdf.

— Callison-Burch, Chris; Osborne, Miles; Koehn, Philipp, "Re-evaluating the Role of BLEU in Machine Translation Research". *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, 2006, pp. 249-256, http://www.aclweb.org/anthology/E/E06/E06-1032.pdf.

— Caseli, Helena de Medeiros; Nunes, Israel Aono, "Tradução Automática Estatística baseada em Frases e Fatorada: Experimentos com os idiomas Português do Brasil e Inglês usando o toolkit Moses (NILC-TR-09-07)". *Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional*, São Carlos, 2009, http://www2.dc.ufscar.br/~helenacaseli/pdf/2009/NILCTR-09-07.pdf.

— Eisele, Andreas; Federmann, Christian; Hodson, James, "Towards an effective toolkit for translators". *Proceedings of the 31rd Translating and the Computer Conference*, Association for Information Management, London, 2009, http://www.dfki.de/web/forschung/iwi/publikationen/renameFileForDownload?filename=TatC31.pdf&file_id=uploads_595.

— Jellinghaus, Michael; Poulis, Alexandros; Kolovratník, David, "Exodus — Exploring SMT for EU Institutions". *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, Uppsala, 2010, pp.116-120, http://www.statmt.org/wmt10/pdf/WMT15.pdf.

— Koehn, Philip — *Statistical Machine Translation*. Cambridge University Press, January 2010. ISBN-10: 0521874157.

— Koehn, Philipp; Birch, Alexandra; Steinberger,Ralf, "462 Machine Translation Systems for Europe". *Proceedings of the Machine Translation Summit XII*, International Association for Machine Translation (IAMT), Association for Machine Translation in the Americas (AMTA), Ontario, 2009, http://www.mt-archive.info/MTS-2009-Koehn-1.pdf.

— Koehn, Philipp; Schroeder, Josh; Osborne, Miles, "Edinburgh University System Description for the 2008 NIST Machine Translation Evaluation". *NIST Open Machine Translation 2009 Evaluation (MT09)* (collocated event with Machine Translation Summit XII), Ontario, 2009, http://homepages.inf.ed.ac.uk/pkoehn/publications/mteval08-report.pdf.

— Koehn, Philipp; Monz, Christof, "Manual and Automatic Evaluation of Machine Translation between European Languages". *Proceedings of the Workshop on Statistical Machine Translation* (Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting), Association for Computational Linguistics, New York, 2006, pp. 102-121, http://www.aclweb.org/anthology-new/W/W06/W06-3114.pdf?CFID=68022703&CFTOKEN=25854326.

— Kos, Kamil; Bojar, Ondřej, "Evaluation of Machine Translation Metrics for Czech as the Target Language". *The Prague Bulletin of Mathematical Linguistics*, no. 92, December 2009, pp. 135-147, http://ufal.mff.cuni.cz/pbml/92/art-pbml92-kos-bojar.pdf.

— Nunes, Israel Aono; Caseli, Helena de Medeiros, "Primeiros Experimentos na Investigação e Avaliação da Tradução Automática Estatística Inglês-Português". *Anais do I TILic - Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana* (collocated event with STIL 2009), 2009, São Carlos, pp. 1-3.

— Specia, Lucia; Raj, Dhwaj; Turchi, Marcho, "Machine Translation Evaluation versus Quality Estimation". *Machine Translation*, SpringerLink, vol. 24, no. 1, 2010, pp. 39-50.