DELIVERABLE

Project Acronym: iTranslate4

Grant Agreement number: 250405

Project Title: Internet Translators for all European Languages

---

4.3 Evaluation campaign report

Revision: version 1.4

Authors:

Csaba Oravecz (Research Institute for Linguistics – Hungarian Academy of Sciences, RIL)
Bálint Sass (Research Institute for Linguistics – Hungarian Academy of Sciences, RIL)
László Tihanyi (Research Institute for Linguistics – Hungarian Academy of Sciences, RIL)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	x
C	Confidential, only for members of the consortium and the Commission Services	

Table of contents

Introduction

An unique property of the iTranslate4.eu site is that it uses several machine translation vendors, thus it provides several different translations of the input texts for certain language pairs. One important issue is how to display these alternative translations to the user. The main purpose of the evaluation framework (developed in WP4) was to establish a solid ordering between vendors serving a given language pair, allowing the site (and also API functions) to use this ordering when sorting the alternatives. Different kinds of machine translation evaluation methods were considered, namely automatic evaluation using some evaluation metrics, manual evaluation by human experts as part of an external campaign, and manual evaluation as part of normal operation, by collecting user votes on the translated sentences. Large scale manual evaluation is by far the most reliable method to evaluate the output of machine translations and have a clear preference over automatic evaluation metrics, which can only indicate the quality of an MT system and are mainly used during system development. The iTranslate4 project has also put much emphasis on manual evaluation and utilized crowd sourcing as well as the dynamically increasing user feedback in the evaluation framework developed in the first period of the project. The present report focuses on the results of the evaluation runs and describes the possible methodology that can be applied to produce a reliable ranking of translation outputs that is acceptable for all the partners. The automatic and crowd sourcing evaluation framework is only presented in a short summary, details are to be found in deliverables D4.1 and D4.2 (available at http://itranslate4.eu/project/deliverables.html).

1. Languages, language pairs, and scope of the evaluation

The 37 languages that currently appear on the iTranslate4.eu site are presented in Table 1. Considering the set of available translation services, five languages (Basque, Breton, Icelandic, Romanian, Welsh; marked with [S]) can only be source languages, and one (Persian; marked with [T]) can only be target language, the remaining 31 languages can be both source and target. Additionally, there are two languages (Norwegian and Norwegian Nynorsk; marked with [X]) which can be translated into each other but not into any other language. All languages except for the two Norwegian can be reached from all other either by one or more direct translation services, or by so called *indirect translation*, where due to lack of a direct service the translation is performed in two steps: the input text is translated into a *pivot language* first and then from the pivot languages to the target language using another direct service. The pivot language is mainly English but there are language pairs where a better suited alternative is applicable (e.g. between two Latin languages). In some cases (e.g. from Occitan to Japan) two indirect steps are needed.

Table 1: Languages of the iTranslate4.eu site (with abbreviations). Source-only languages are marked with [S], target-only languages are marked with [T], and excluded languages are marked with [X]. See text for details.

Arabic (ar)	Finnish (fi)	Korean (ko)	Russian (ru)
Basque (eu)[S]	French (fr)	Latvian (lv)	Slovenian (sl)
Breton (br)[S]	Galician (gl)	Macedonian (mk)	Spanish (es)
Bulgarian (bg)	German (de)	Norwegian (no)[X]	Swedish (sv)
Catalan (ca)	Greek (el)	Norwegian N. (nn)[X]	Turkish (tr)
Chinese (zh)	Hungarian (hu)	Occitan (oc)	Ukrainian (uk)
Danish (da)	Icelandic (is)[S]	Persian (fa)[T]	Welsh (cy)[S]
Dutch (nl)	Italian (it)	Polish (pl)	
English (en)	Japanese (ja)	Portuguese (pt)	
Esperanto (eo)	Kazakh (kk)	Romanian (ro)[S]	

Based on the above, the number of available language pairs iTranslate4.eu can serve can be calculated as $n_{lp} = n_s * n_t - n_{st} - n_x$, where n_s=*number of source languages,* n_t=*number of target languages*, n_{st}=*number of languages that can be both source and target* while n_x stands for excluded language pair members which cannot be connected at all. The result of the calculation is: $n_{lp} = 36*32-31-128=993$. Out of these there are 167 language pairs (so called *direct language pairs*) which are supported by 251 different direct services. Translation on all the other 826 language pairs (so called *indirect language pairs*) performed indirectly through a pivot language. In case of 38 language pairs (out of the above 167) more than one direct services are available as shown in Table 2.

Table 2: Language pairs for which more than one direct services are available. The number of direct services is indicated only when larger than two. Language pairs which are only partly evaluated are marked with [P], and the only language pair which is not evaluated is marked with [N].

de-en *6*	en-de *6*	es-de *3*	fr-de *6*	it-de[P]	pl-de	pt-en *3*	ru-de	zh-en[P]
de-es *3*	en-es *5*	es-en *5*	fr-en *5*	it-en *4*	pl-en *3*	pt-es[P]	ru-en *3*	
de-fr *6*	en-fr *5*	es-fr *4*	fr-es *4*	it-es *3*	pl-ru		ru-fr[P]	
de-it[P]	en-it *4*	es-it *3*	fr-it[P]	it-fr[P]			ru-pl	
de-pl	en-pl *3*	es-pt[N]	fr-ru					
de-ru	en-pt *3*							
	en-ru *3*							
	en-zh[P]							

Vendors of direct services are listed in Table 3.

Table 3: List of 14 vendors who provide (or provided) direct translation services for iTranslate4 together with their three letter abbreviations used throughout iTranslate4 deliverables.

Amebis	AME
Google	GOO
Grammarsoft	GRA
Lingenio	LIN
Linguatec	LGT
Microsoft Bing	MST
MorphoLogic	MOR
Prompsit	PRM
PROMT	PRO
pwn.pl	PWN
SkyCode	SKY
Sunda	SUN
SYSTRAN	SYS
Trident	TRD

The evaluation campaign has been carried out for direct language pairs, where there is a competition between direct services (i.e. the language pairs in Table 2). We decided to investigate only direct services for several reasons. First of all, the results for indirect translations in principle could be calculated from the evaluation results of the individual translation steps. Secondly, such results are of little relevance for the translation companies since the second translation step usually receives a linguistically degraded input from the first step. Developers of rule based systems still focus on the mapping of grammatically correct constructions and are not particularly interested in ungrammatical text especially if produced by machine translation. With the above considerations the enormous evaluation task of all 1041 language pairs is reduced to evaluating 38 language pairs. It should be noted that 38 is still a rather large number in terms of language pairs in MT evaluation campaigns. For example EuroMatrix evaluation campaign in 2009 (Callison-Burch *et al.*, 2009b) dealt with 9 language pairs and also the 2010 annual report of the current EuroMatrix+ project reports a MT evaluation campaign on only 6 language pairs.

The automatic and crowd sourcing evaluations have been carried out for 36 languages: for the requested 38 except for es-pt and pt-es, because the partner providing the second direct service joined the project later. The number of available languages and services continuously changed during the project. In the beginning, Google and Microsoft services were also included but later got dropped when they started to charge their services. SYSTRAN's statistical engines have very recently been stopped as well. These changes decreased the number of translator engines but three newly joined partners compensated for this decrease. In summary, automatic evaluation has been applied to 53 (first run) and then to 50 (second run) language pairs, and crowd sourcing based evaluation to 55 language pairs, both including the mentioned 36 language pairs. Although the number of user votes collected from the site is constantly growing, there are seven language pairs for which the necessary amount of user votes is not available yet, so the complete evaluation data (consisting of all the three evaluation strategies) is present only for 29 languages at the time of the preparation of the report.

There are language pairs which should not have been evaluated based on the current services, but some evaluation data has still been accumulated for them. This is because there was a competition between at least two direct services for the language pair in question at the time of the particular evaluation run, and/or there are simply enough votes collected for the service. Table 4 shows the statistics of evaluated language pairs together with the above discussed numbers.

Table 4: Language pairs in evaluation. AU stands for automatic evaluation, HE stands for human MTurk evaluation, and UF for user feedback evaluation.
For individual language pairs see Table 2 and text below.

	evaluated				not evaluated	total
	fully	AU+HE	HE	UF		
LPs to evaluate (1< direct services currently)	29	7		1	1	38
LPs not to evaluate (1 or 0 direct service curently)	7	5	7		936	955
subtotal	36	12	7	1		
total	56				937	993

To cover all evaluated language pairs Table 4 also lists language pairs which only have one direct engine currently but had more than one at an earlier stage of the project and so were included in the evaluation to some extent. According to line "LPs not to evaluate" of Table 4 there are 7 such LPs evaluated fully: bg-en, en-fi, en-hu, fi-en, hu-en, sl-en, uk-en; there are 5 LPs having automatic and human MTurk evaluation: en-lv, en-tr, lv-en, pl-fr, tr-en; and finally there are another 7 LPs which only have human Mturk evaluation: da-en, en-bg, en-da, en-sl, en-sv, no-en, sv-en.

2. Summary of the automatic evaluation framework

The automatic evaluation framework is built around the IQMT toolkit (Giménez, 2007), which is a common workbench integrating a number of standard evaluation methods and metrics. Evaluation results were calculated as a normalized average of 5 individual metrics: BLEU (Papineni et al., 2002), NIST (Doddington, 2002), GTM (Melamed et al., 2003), METEOR (Banerjee and Lavie, 2005) and ROUGE (Lin and Och,2004).

Language resources have been collected from the EU news parallel corpus at ec.europa.eu/news, with the following domains covered: *agriculture, business, external relations, culture, justice,*

economy, regions, employment, science, energy, transport, environment. The size of the resource is about 1900 paragraphs per language of which only a random selection of 25 were used for the evaluation.

Results

The full results from the first run (9th March 2011) of automatic evaluation are presented in the AU-1 column of Table 5 in the Appendix. At that time Google and Microsoft also offered a free service, which was included in the early version of the iTranslate4.eu portal, so these results were calculated with these services incorporated, which amounts to 252 different translators for 53 language pairs. As a tendency, also noted in Callison-Burch et al., (2010), automatic evaluation metrics tend to favor statistical MT systems with respect to rule based ones, and this finding seems to be confirmed by our results as well. It is important to emphasize, however, that for language pairs with well developed rule based systems other evaluation components can give rather different results (see Section 4.1), giving some empirical justification for the unreliability of automatic metrics and indicating the limitations of their usability.

The second run of automatic evaluation (21st March 2012) could only focus on the partner translators, the out-of-consortium services are no longer available and therefore are now out of the offered portal services as well. In the current scenario there are 144 different translators for 50 language pairs. For the sake of comparison the results of the first run have been renormalized with the Google/Microsoft services excluded (AU-1re column in Table 5); taking these scores and the scores from column AU-2 makes some comparative evaluation possible. Note that the scores do not express some absolute value because of the normalization, which takes into account the maximum value and span of a specific metric during an evaluation run to ensure a fair comparison (for example, a score of 0.1 should be weighted more with respect to a score of 0.05 if the best result of all translators in the given metric is 0.2 and much less if the best score is 0.8). Consequently, it is the changes in the ranking positions and the ratio of differences in the scores that can be considered for inspection and not the absolute values (to examine how a particular translator in itself performs now with respect to the first run one can look at the individual output results before normalization). The overall tendency of getting lower absolute scores in the second run is simply caused by higher global maximums for the given metric — the best translator performed better for that metric in the second run and so the respective values for the other translators were slightly demoted. (The score of 1.000 for the APE engine in the *pt-e*s language pair means for example that this translator for this language pair received the best (global maximum) values for each and every metric in the whole second evaluation run.)

In view of the above, generally there are very few differences in the results of the two runs. The rankings significantly changed only in the language pair en-de (English-German), other differences are minor and insignificant. The new APE engine performed relatively well for the pairs it provides. It is again of key importance to underline that these automatic measures and results are extremely unreliable and have therefore only a limited role in the final evaluation scenario (see Section 6.3).

3. Human evaluation using Mechanical Turk

The second evaluation component in iTranslate4 evaluation campaign was human *MTurk* evaluation using the online marketplace for tasks that need human intelligence called Mechanical Turk (www.mturk.com). A corpus containing medium-length (min. 5 max. 30 words long) test sentences for every source language was collected representing a range of different topics. The task of the evaluators was to rank the translations of 30 sentences from 1 (best) to 5 (worst). The 30 sentences were divided up in to 6 groups, each containing 5 sentences and allowing the evaluator to work only

with 5 source sentences at a time. The process of selecting the evaluators as well as further details about the campaign are described in deliverable D4.1.

The task for Swedish-English language pair is illustrated in Figure 1. During the campaign, every sentence was evaluated with 3 different evaluators. The final score was calculated in two ways, first, as a simple arithmetic mean and second as in the method used in the EuroMatrix project (Callison-Burch et al., 2009b), where a translator gained a point when it was evaluated better than (or equal with) another translator, and the translator with the most points won. In most cases these two metrics gave the same ordering.

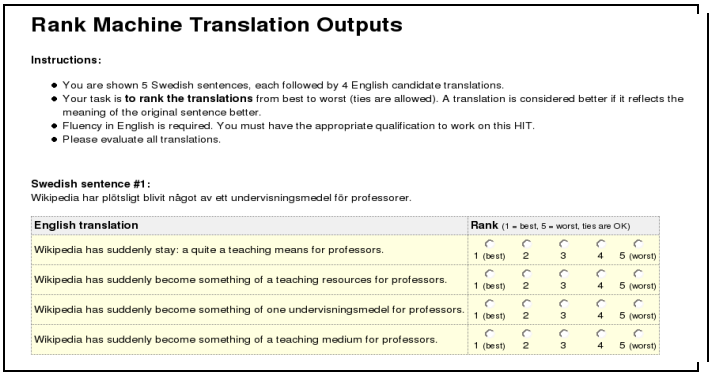

Figure 1: Interface for the Mechanical Turk evaluation.

Results

Similarly to the other methods, only those language pairs were considered for which there was a direct translation and there were more than one translation engines available. When the iTranslate4 human evaluation was performed (summer 2011) the free translation services from Google and Microsoft were still available, and thus these two services were also included in the human evaluation campaign. We evaluated 55 language pairs and 255 different translators using Mechanical Turk. All numerical results are shown in the HE column of Table 5.

As we can see in the AU-1 column, Google Translate (GOO) and Microsoft Bing (MST) showed superior performance compared to iTranslate4 partners in all cases according to the automatic evaluation. Human manual evaluation, which is considered much more reliable, shows a different picture drawing attention to the important property of the iTranslate4.eu portal, namely, that it mostly uses rule-based MT systems. There are 13 language pairs where an iTranslate4 partner proved to be the best and in some cases several partners performed better than Google Translate and Bing. This result shows that the task of machine translation has not yet been solved in general by the statistical machine translation paradigm, development of individual (rule-based or hybrid) systems can still be justified by their excellent performance. This result also highlights the known drawback of automatic machine translation evaluation, namely the preference for statistical systems. The standard measures for the correlation between the results in the AU-2 and HE columns, namely the Spearman rank correlation coefficient and Kendall tau give an average value of 0.4 and 0.37, respectively, which are fairly low so clearly we need a better alternative for reliable evaluation.

The winning language pairs together with the partner achieving the result are showed in Figure 2.

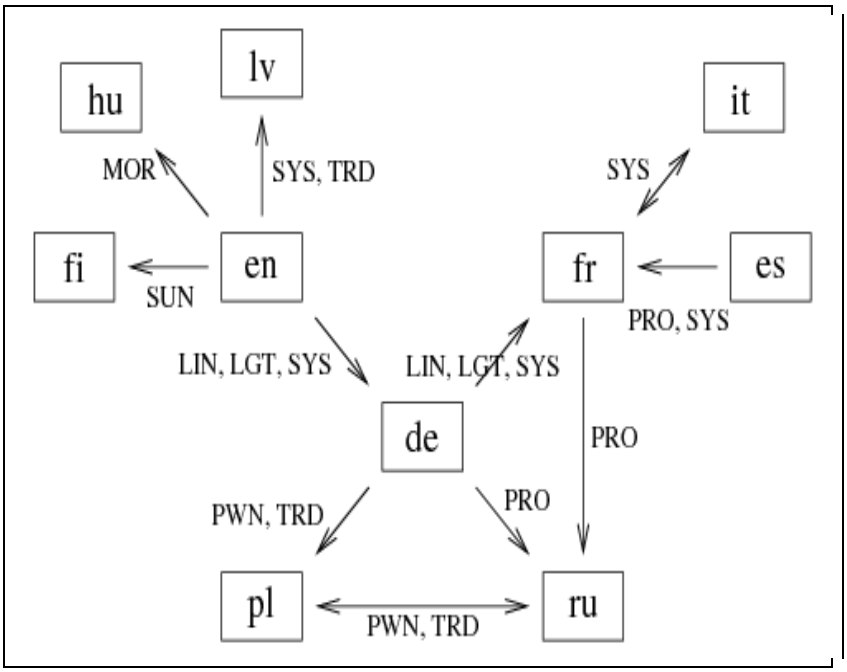

Figure 2: Language pairs and translators with better performance than statistical systems.

If we leave out English for an instant, we can see that the big European languages do appear in Figure 2. These are the languages which are important in Europe and there is a great demand for quality machine translation between them. These languages generate a significant traffic on the website although an order or magnitude less than English. Big machine translation sites often implement translation services between two non-English languages using English as the pivot language, here we show that thoroughly developed direct services between such languages do perform better. The overall good performance on non-English languages is further reinforced by the following data. While it is only in 4 from 34 (11.8 percent) language pairs containing English that iTranslate4.eu performs better than Google and Bing, this ratio is significantly higher in language pairs not containing English: 9 from 21 (42.9 percent).

iTranslate4 partners can win even in both directions where the source and target languages are fairly similar and at the same time different from English (see pl-ru (Polish-Russian) and fr-it (French-Italian)). Some smaller languages (Finnish, Hungarian, Latvian) are also there, probably because there is a serious effort in the respective country to develop a quality machine translation system for their own language.

4. Evaluation based on user feedback

The third component of the evaluation is based on the user feedback collected as votes for the translation outputs. More than one alternatives can be selected if available for a specific language pair. User feedback has been collected ever since the opening of the translation portal, however, the amount of votes has only been increasing moderately. As more and more data of this type has been accumulated, this evaluation component can be considered as primary indicator of the quality of the particular engines and can serve as the most reliable base for any kind of ranking implemented for the translation outputs.

A threshold of 10 votes (per language pair) has been set up in order to take into consideration these results for a certain language pair. This way valid data has been accumulated for 37 language

pairs. Out of the 38 direct language pairs to be evaluated (see Table 4) there were 29 which could be extended by voting results. The remaining language pairs belong to the newly joined partners or newly installed direct services so the votes have been collected only for a shorter time period.

User feedback is also available for 129 language pairs served by a single direct engine. These language pairs in many cases are extended with indirect language pairs resulting in multiple solutions, which can in principle be ranked by the user votes. There were 44 (non-English) language pairs with a direct translator and a single engine, which were extended with indirect solutions, this way offering alternatives to the users. This is an additional aspect of the evaluation protocol which has to be investigated further, i.e. whether user feedback should be used here at all to influence the ranking or the superiority of the direct solution over the indirect one(s) should be taken for granted.

Results

Numerical results, which are simply measured as the ratio of votes given to a particular engine for a particular language pair to the total number of votes given to all engines for a particular language pair, are presented in the UF column of Table 5. Even in a causal comparison with the second run results of the automatic evaluation, it is obvious that there are major differences in the rankings, which underlines again the importance of this kind of continuous user feedback about the translations and its importance in the calculation of the final rankings. Comparing the MTurk evaluation with user feedback in numerical terms, the correlation of the results in the HE and UF columns is 0.68 for the Spearman rank correlation coefficient and 0.65 for Kendall tau, which is significantly higher than in the case of the automatic method. This suggests that user feedback could indeed be a valuable replacement for the unreliable automatic evaluation and also for the expensive and unsustainable MTurk evaluation.

5. Current services

5.1. Evaluation strategies

In a practical application, a predefined ranking among the translator engines that is fixed on the basis of a one time evaluation run is hardly a viable solution. As illustrated in Figure 3, when a weighted average of the three components (automatic evaluation (c_{AU}), human expert evaluation at Mechanical Turk (c_{HE}), user feedback (c_{UF})) is calculated for example according to the following formula:

(1) score= $w_1 * c_{AU} + w_2 * c_{HE} + w_3 * c_{UF}$

there can be minimal differences in the results. In this case a specific partner will have every right to protest for a permanently demoted position in the ranking on the basis of a never 100% reliable evaluation run. We used w_1=0.1, w_2=0.3, and w_3=0.6 as an example setting representing the priority of manual evaluation over the automatic one and also the priority of the opinions of real users over results from artificial evaluation settings. It should be noted that formal justification for the exact coefficients would need extensive experiments to measure the correlation of the calculated score with manual expert evaluation. This, however, was beyond the scope of the project.

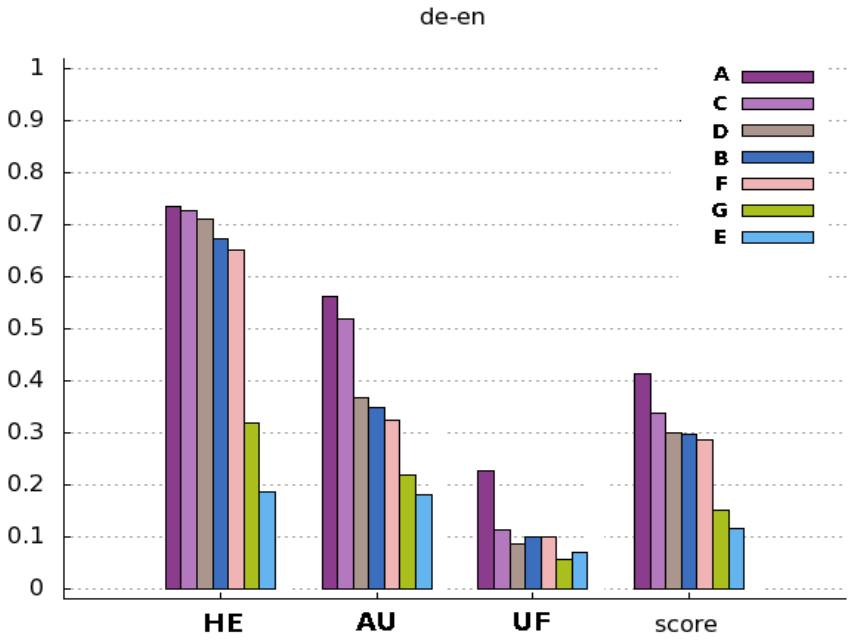

Figure 3: Results from components and unified score.

5.2. Proposed methodology

As we mentioned above, ranking only makes sense if the evaluation shows substantial differences between engines for a given language pair. To circumvent this problem the following methodology is proposed. On the basis of the overall results from the evaluation the translator engines are clustered into quality classes with the help of a density-based clustering method which does not require the number of clusters predefined (an efficient algorithm is for example DBSCAN (Ester et al., 1996)). This process is illustrated in Figure 4. Within each quality cluster random ordering is applied for each translation run. The input to the clustering algorithm can either be the weighted average of the results from three evaluation components according to (1), or an ordered selection of only the most reliable component that first provides sufficient evaluation data about a particular language pair (where Mechanical Turk evaluation is considered as the most reliable, user feedback less and the automatic component as the least reliable, and as more and more user data is collected, this can overtake the most prominent role since the Mechanical Turk evaluation is not sustainable in the long run and its data becomes obsolete pretty soon.)

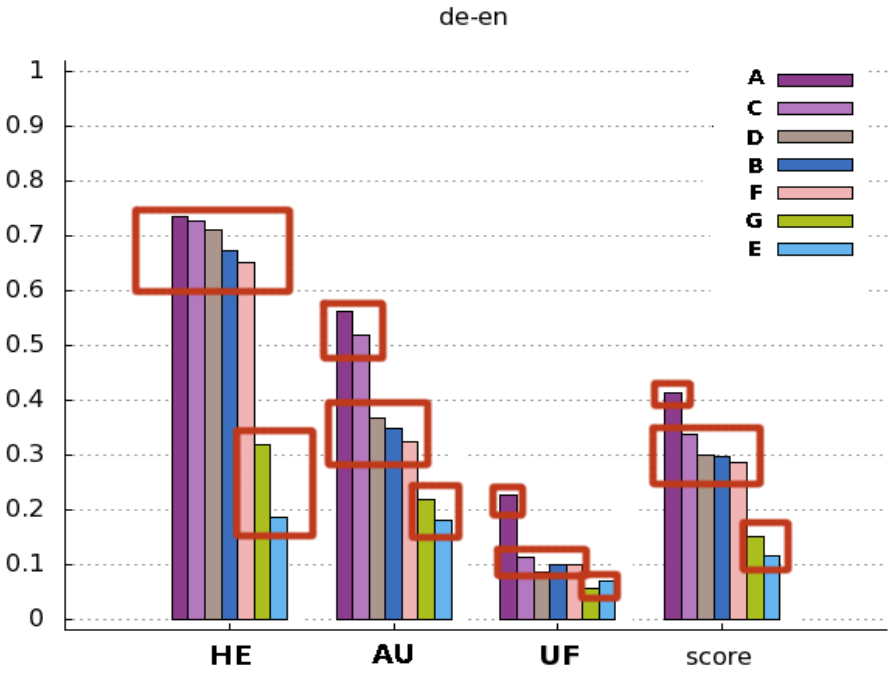

Figure 4: Clustering the evaluation results.

5.3. Current implementation

The direct services of the evaluated language pairs are listed together with their evaluation result (a number between 0 and 1) in a database. At present, final evaluation scores are calculated according to Formula (1) in Section 5.1. The clustering step is not yet applied at this stage due to the practical considerations mentioned below. The database of scores is consulted by the server program. If the database contains data for the current language pair then the order of the translations is presented in accordance with the evaluation result. If no data is found a random order is provided. Current scores are presented in the last but one column of Table 5 in the Appendix.

The evaluation results can easily be displayed in simple bar plots, which are automatically generated and illustrated in Figure 5, 6 and 7. In Figure 5, user feedback nicely correlates with the crowd sourcing result, in Figure 6 it is user feedback data that determines an ordering since the other two methods give very similar results for the first two engines, while in Figure 7 user feedback further emphasizes the minor differences already present in the results from the other methods.

Unfortunately the above principle (see Section 5.2.) had to be modified in practice because the speed of the translators proved to be sometimes significantly different. For an average length input in general a 1 second latency was considered as acceptable so the ordering algorithm is applied only to those translators which arrive within this time interval. Translations arriving later are simply displayed at the end of the list. This is somewhat contradictory to simple intuition that the longer the engine works the better the result it will produce, but this constraint had to be introduced to provide a convenient service for the end user. Occasionally, this strategy can go against the declared promise and goal of the whole evaluation process, i.e. the better the translation the higher it should appear in the list, but practical considerations overruled the ideal objective here.

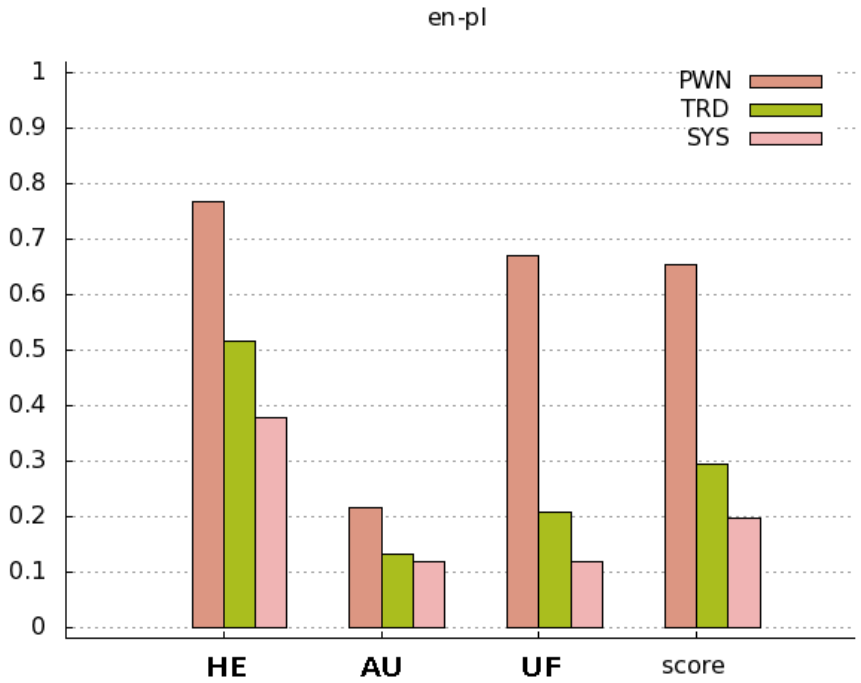

Figure 5: User feedback reinforces crowd-sourcing result.

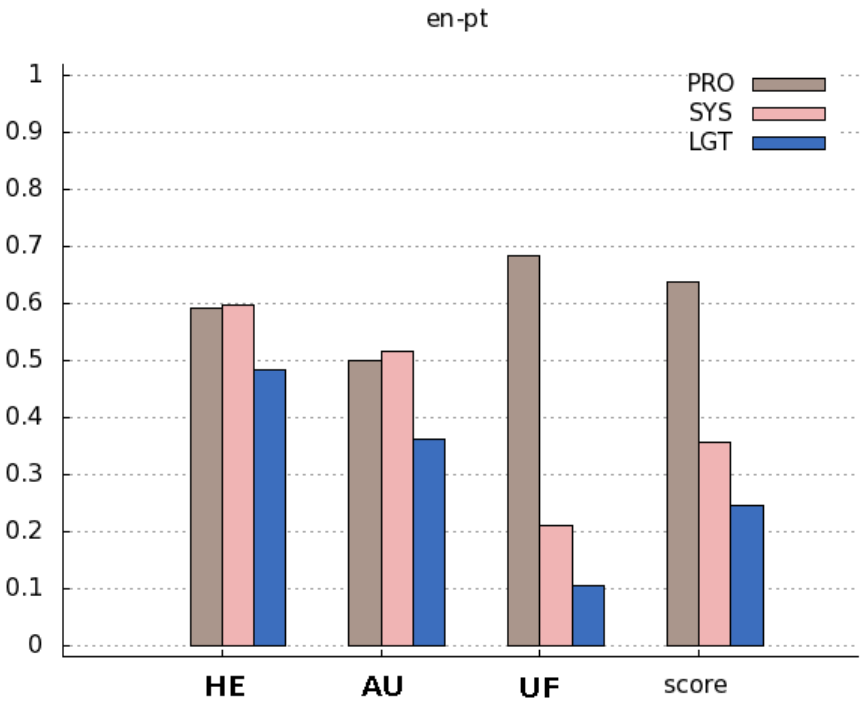

Figure 6: User feedback decides in a tie.

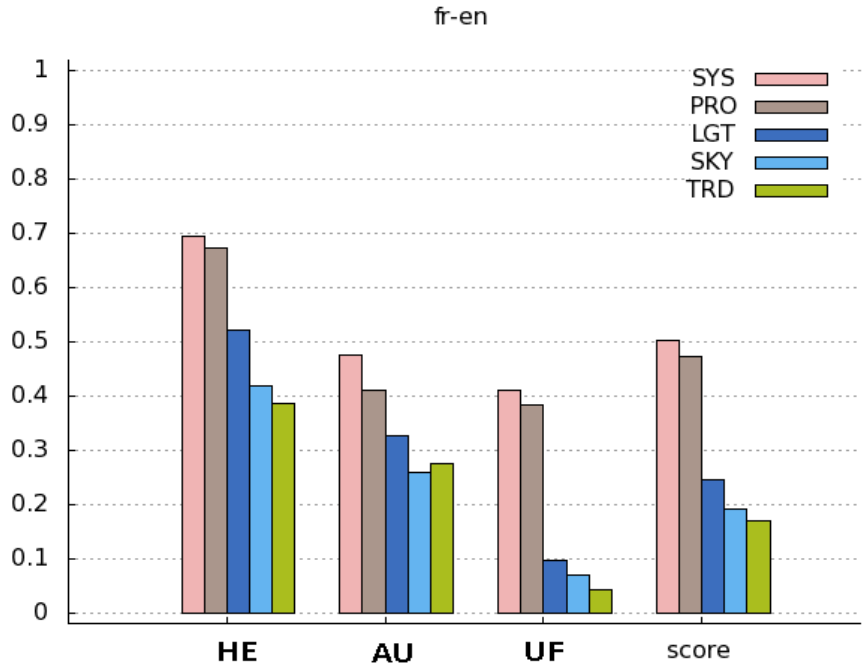

Figure 7: User feedback increases distinctions

6. Conclusion

The original aim of the work package on evaluation was to work out a sustainable framework which is able to order translation engines according to their quality for a given language pair. Additionally, the results of the iTranslate4 evaluation campaign supported the known fact that automatic machine translation measures being biased towards statistical translation engines are not as reliable as human evaluation. The results also showed that rule-based systems of the iTranslate4.eu portal can outperform statistical ones, underlying the importance of rule-based systems in machine translation in general. In some cases the difference between two engines is not significant, so a clustering-based solution was proposed to ensure the fairness of evaluation. At the end, due practical considerations about the usability of the site the proposed method is not fully implemented but it takes also the response time of the particular engine into account.

The sustainability of the evaluation campaign in the form of a three component evaluation method is an important issue. Automatic evaluation can be run at any time provided that independent test data is available for the given language pair but its reliability still remains a problem. The crowd-sourcing component is expensive and resource intensive in such a large scale that is required for reliable results. It is therefore the user feedback that could be utilized most efficiently although a periodic controlling experiment by crosschecking its correlation with expert manual evaluation is always useful and some filtering is also worth being introduced to eliminate salient outliers from the user votes.

References

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz and Josh Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In: *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70-106, Columbus, Ohio, June. Association for Computational Linguistics.

Chris Callison-Burch. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286-295, Singapore, ACL.

Chris Callison-Burch, Philipp Koehn, Christof Monz and Josh Schroeder. 2009a. Findings of the 2009 Workshop on Statistical Machine Translation. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1-28, Athens, Greece.

Chris Callison-Burch, Philipp Koehn, Christof Monz and Josh Schroeder. 2009b. *Evaluation Campaign*. EuroMatrix Deliverable 1.2b.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki and Omar Zaidan. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17-53, Uppsala, Sweden, ACL.

Jésus Giménez. 2007. *IQMT. A Framework for Automatic Machine Translation Evaluation based on Human Likeness*, Technical Manual, TALP Research Center.

NIST. 2010. *The NIST Metrics for MAchine TRanslation 2010 Challenge*. (MetricsMaTr10) Evaluation Plan, Appendix A.

Banerjee, Satanjeev and Lavie, Alon. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 2005, 65-72.

Doddington, George. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: *HLT-01*, 2002.

Ester, Martin, peter Kriegel, Hans, S, Jörg and Xu, Xiaowei. *A density-based algorithm for discovering clusters in large spatial databases with noise*. AAAI Press, 1996, 226-231.

Fort, Karën, Adda, Gilles and Cohen, K. Bretonnel. Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, 2011, 37(2):413-420.

Lin, Chin-Yew and Och, Franz Josef. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL '04,* Stroudsburg, PA, USA. Association for Computational Linguistics, 2004.

Melamed, I. Dan, Green, Ryan and Turian, Joseph P. Precision and recall of machine translation. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003 short papers - Volume 2, NAACL-Short '03*, Stroudsburg, PA, USA. Association for Computational Linguistics, 2003, 61-63.

Papineni, Kishore, Roukos, Salim,Ward, Todd and Zhu,Wei-Jing. Bleu: A method for automatic evaluation of machine translation. In: *ACL-02*, Philadelphia, PA. 2002.

Appendix

Table 5: Detailed numerical data of evaluation. Results of automatic evaluation are presented in three data columns. AU-1*,* AU-1re *and* AU-2 *contain data from the first run, data from the first run renormalized, and data from the second run respectively. Renormalized data (see Section 2) consisting only vendors also available in the second run is presented to allow direct comparison between the two runs. While 53 LPs were included in the first run (48 of which are automatically evaluated according to Table 4, as well as* da-en, en-da, en-sv, no-en*, and* sv-en *which had more than one direct services at the time of the first run), the second run contained 50 LPs (the same 48 supplemented by* es-pt *and* pt-es*). The rankings significantly changed only in* en-de *LP between the two runs. The fourth data column (labelled* HE*) shows the results of the manual MTurk evaluation for 55 language pairs (cf. Table 4). EuroMatrix scores (see Section 3) are presented normalized to 1 using the possible maximum score value for each language pair separately. Thirteen LPs where an iTranslate4 partner turned to be best according to MTurk evaluation (cf. Figure 2) are marked in the comment column. The fifth data column (labelled* UF*) presents the results of evaluation based on user feedback for 37 LPs which have enough user votes (cf. Table 4:* fully *and* UF *columns). Comments indicate if AU and UF evaluation provide different orderings. The last data column (labelled* final*) gives the final evaluation scores currently used by the iTranslate4.eu site for the 37 LPs which have more than one direct services at present and for which at least one kind of evaluation data is available (cf. Table 4: first row). Vendors (in the second column) are sorted according to final score. Since Google Translate and Microsoft Bing are not included in iTranslate4 currently, they are listed at the end. Actual ordering of the figures for a given LP and a given evaluation method is indicated by small numbers next to the data.*

LP	vendor	AU-1		AU-1re		AU-2		HE		UF		final		comment
bg-en	SKY	0.2993	3	0.4384	1	0.3923	1	0.5741	3	0.8120	1			
	SYS	0.1614	4	0.2358	2	0.2218	2	0.1815	4	0.1880	2	–		
	GOO	0.7416	1					0.8407	1					
	MST	0.6525	2					0.6741	2					
da-en	GRA	0.4562	3					0.7037	2					
	SYS	0.4074	4					0.3667	4					
	GOO	0.6076	1					0.8593	1			–		
	MST	0.5898	2					0.6259	3					
de-en	LGT	0.3496	5	0.5358	3	0.4730	3	0.6722	4	0.2936	1	0.4251	1	AU $\neq$ UF
	PRO	0.3679	3	0.5804	1	0.5395	1	0.7111	3	0.2385	2	0.4104	2	
	SYS	0.3248	6	0.4942	4	0.4441	4	0.6519	5	0.1560	4	0.3336	3	
	LIN	0.3616	4	0.5590	2	0.5136	2			0.1881	3	0.2346	4	
	TRD	0.2195	7	0.3065	5	0.2795	5	0.3185	6	0.0780	5	0.1703	5	
	SKY	0.1800	8	0.2446	6	0.2256	6	0.1852	7	0.0459	6	0.1057	6	
	GOO	0.5630	1					0.7352	1					
	MST	0.5188	2					0.7278	2					
de-es	PRO	0.3125	3	0.4866	1	0.4500	1	0.5944	3	0.9231	1	0.7772	1	
	SYS	0.2613	4	0.3905	2	0.3524	2	0.4417	4	0.0000	3	0.1678	2	
	SKY	0.1398	5	0.1893	3	0.1893	3	0.0917	5	0.0769	2	0.0926	3	
	GOO	0.4933	1					0.7833	1					
	MST	0.4266	2					0.6889	2					

LP	vendor	AU-1		AU-1re		AU-2		HE		UF		final		comment
de-fr	LIN	0.2937	4	0.4476	2	0.3991	2	0.7810	1	0.3103	1	0.4604	1	MTurk
	PRO	0.2782	5	0.4273	3	0.3820	3	0.7302	4	0.2414	2	0.4021	2	winner
	LGT	0.2721	6	0.4078	4	0.3119	4	0.7635	2	0.1724	3	0.3637	3	
	SYS	0.3158	3	0.4911	1	0.4365	1	0.7444	3	0.1379	4	0.3497	4	AU $\neq$ UF
	TRD	0.1246	7	0.1694	5	0.1531	6	0.2302	7	0.0690	5	0.1258	5	
	SKY	0.1225	8	0.1654	6	0.1930	5	0.1968	8	0.0690	5	0.1197	6	
	GOO	0.4287	1					0.7302	4					
	MST	0.3809	2					0.6580	6					
de-it	SYS	0.2047	3	0.2938	1	0.2678	1	0.5259	3			0.1846	1	
	SKY	0.0867	4	0.1178	2	0.1463	2	0.1222	4			0.0513	2	
	MST	0.3610	1					0.7037	2					
	GOO	0.3542	2					0.7926	1					
de-pl	PWN	0.1051	3	0.1579	1	0.1423	1	0.7556	1	0.6250	1	0.6159	1	MTurk
	TRD	0.0720	4	0.0945	2	0.0871	2	0.7296	2	0.3750	2	0.4526	2	winner
	GOO	0.1980	1					0.5296	4					
	MST	0.1189	2					0.6222	3					
de-ru	PRO	0.1839	3	0.2570	1	0.2402	1	0.8111	1	0.7054	1	0.6906	1	MTurk
	TRD	0.1362	4	0.1837	2	0.1812	2	0.3963	4	0.2946	2	0.3138	2	winner
	GOO	0.2122	1					0.6778	2					
	MST	0.1993	2					0.5889	3					
en-bg	SKY							0.4889	2					
	GOO							0.7722	1			–		
	MST							0.4111	3					
en-da	GRA	0.3471	3					0.6296	3					
	SYS	0.2571	4					0.1815	4					
	GOO	0.5542	1					0.8630	1			–		
	MST	0.4472	2					0.6926	2					
en-de	LGT	0.3025	4	0.4676	2	0.3655	3	0.7556	2	0.2961	1	0.4409	1	AU
	PRO	0.2676	6	0.4080	4	0.3733	2	0.6778	5	0.2697	2	0.4025	2	ranking
	LIN	0.3042	3	0.4702	1	0.3523	4	0.7778	1	0.2105	3	0.3949	3	changed
	SYS	0.2921	5	0.4466	3	0.4146	1	0.7302	3	0.1513	4	0.3513	4	
	TRD	0.0901	8	0.1312	6	0.1278	6	0.3857	7	0.0329	6	0.1482	5	MTurk
	SKY	0.1401	7	0.1940	5	0.1791	5	0.2730	8	0.0395	5	0.1235	6	winner
	GOO	0.4051	1					0.7079	4					
	MST	0.3544	2					0.6365	6					AU $\neq$ UF
en-es	PRO	0.4091	4	0.6652	2	0.5969	2	0.6200	4	0.4817	1	0.5347	1	AU $\neq$ UF
	SYS	0.4704	3	0.7727	1	0.7032	1	0.6222	3	0.3232	2	0.4509	2	
	LGT	0.3458	5	0.5434	3	0.4074	4	0.5000	5	0.0915	3	0.2456	3	
	APE					0.5440	3			0.0732	4	0.1404	4	
	SKY	0.1573	6	0.2155	4	0.2753	5	0.1333	6	0.0305	5	0.0858	5	
	GOO	0.6666	1					0.8178	1					
	MST	0.5820	2					0.7400	2					

LP	vendor	AU-1		AU-1re		AU-2		HE		UF		final		comment
en-fi	SUN	0.2341	3	0.3537	1	0.3594	1	0.6944	1	0.9278	1			MTurk
	SYS	0.1736	4	0.2454	2	0.2234	2			0.0722	2	–		winner
	GOO	0.3179	1					0.6556	2					
	MST	0.2533	2					0.3222	3					
en-fr	SYS	0.3985	3	0.6463	1	0.5717	1	0.7574	2	0.4390	1	0.5478	1	
	PRO	0.3851	4	0.6244	2	0.5549	2	0.7278	3	0.3415	2	0.4787	2	
	LGT	0.3077	5	0.4758	3	0.3591	3	0.6074	5	0.1707	3	0.3206	3	
	SKY	0.1961	6	0.2795	4	0.2544	4	0.2370	7	0.0244	4	0.1112	4	
	TRD	0.0768	7	0.1020	5	0.0968	5	0.2389	6	0.0244	4	0.0960	5	
	GOO	0.5035	1					0.8389	1					
	MST	0.4466	2					0.7111	4					
en-hu	MOR	0.2056	3	0.2971	1	0.2655	2	0.7500	1	0.9669	1			MTurk
	SYS	0.1952	4	0.2880	2	0.2692	1			0.0331	2	–		winner
	GOO	0.3188	1					0.6278	2					
	MST	0.2581	2					0.3167	3					AU $\neq$ UF
en-it	SYS	0.3596	3	0.5698	1	0.5118	1	0.7778	2	0.3441	2	0.4910	1	AU $\neq$ UF
	PRO	0.2936	5	0.4547	3	0.4106	3	0.6133	5	0.3925	1	0.4606	2	
	LGT	0.3283	4	0.5182	2	0.4589	2	0.6244	4	0.1882	3	0.3461	3	
	SKY	0.1216	6	0.1644	4	0.1980	4	0.1933	6	0.0753	4	0.1230	4	
	GOO	0.5640	1					0.8200	1					
	MST	0.5162	2					0.7511	3					
en-lv	SYS	0.3088	3	0.4898	1	0.4456	1	0.5593	1					MTurk
	TRD	0.1528	4	0.2104	2	0.2008	2	0.4519	2					winner
	GOO	0.4160	1					0.4407	3					
	MST	0.3826	2					0.4222	4					
en-pl	PWN	0.1402	3	0.2183	1	0.2164	1	0.7694	2	0.6719	1	0.6556	1	
	TRD	0.0932	4	0.1360	2	0.1313	2	0.5167	4	0.2086	2	0.2933	2	
	SYS	0.0866	5	0.1253	3	0.1183	3	0.3778	5	0.1194	3	0.1968	3	
	GOO	0.3009	1					0.7806	1					
	MST	0.1589	2					0.5722	3					
en-pt	PRO	0.3480	4	0.5518	2	0.4995	2	0.5917	4	0.6842	1	0.6380	1	AU $\neq$ UF
	SYS	0.3605	3	0.5717	1	0.5166	1	0.5972	3	0.2105	2	0.3571	2	
	LGT	0.3205	5	0.4997	3	0.3609	3	0.4833	5	0.1053	3	0.2443	3	
	GOO	0.5712	1					0.7806	1					
	MST	0.5676	2					0.6972	2					
en-ru	PRO	0.1603	3	0.2290	1	0.2413	1	0.7000	2	0.5363	1	0.5559	1	
	TRD	0.1324	4	0.1790	2	0.1993	2	0.6111	3	0.2787	2	0.3705	2	
	SYS	0.1300	5	0.1735	3	0.1909	3	0.4278	5	0.1849	3	0.2584	3	
	GOO	0.2287	1					0.7361	1					
	MST	0.2069	2					0.5083	4					
en-sl	AME							0.4056	3					
	GOO							0.8611	1			–		
	MST							0.6278	2					

LP	vendor	AU-1		AU-1re		AU-2		HE		UF		final		comment
en-sv	SYS	0.2541	3					0.3889	3					
	GRA	0.1786	4					0.3630	4			–		
	GOO	0.5692	1					0.8926	1					
	MST	0.5041	2					0.7111	2					
en-tr	SYS	0.1855	3	0.2750	1	0.2456	1	0.5407	3					
	SKY	0.0998	4	0.1354	2	0.1275	2	0.3519	4			–		
	GOO	0.3155	1					0.8667	1					
	MST	0.2595	2					0.7704	2					
en-zh	SYS	0.1340	4	0.1700	2	0.1610	2	0.3000	3			0.1061	1	
	LGT	0.1432	3	0.1817	1	0.1721	1	0.2037	4			0.0783	2	
	MST	0.1541	1					0.4222	2					
	GOO	0.1501	2					0.4556	1					
es-de	PRO	0.1974	3	0.2818	1	0.2583	1	0.6111	4	0.7368	1	0.6512	1	
	SYS	0.1937	4	0.2779	2	0.2581	2	0.6139	3	0.2632	2	0.3679	2	
	SKY	0.1212	5	0.1630	3	0.1500	3	0.1583	5	0.0000	3	0.0625	3	
	GOO	0.3265	1					0.6500	2					
	MST	0.3138	2					0.6722	1					
es-en	PRO	0.3637	3	0.5625	1	0.5236	1	0.6533	3	0.4574	1	0.5228	1	AU $\neq$ UF
	SYS	0.3436	4	0.5201	3	0.4756	3	0.6133	4	0.2021	2	0.3528	2	
	LGT	0.3430	5	0.5223	2	0.4809	2	0.5867	5	0.1543	3	0.3167	3	
	APE					0.4256	4			0.1436	4	0.1839	4	
	SKY	0.2043	6	0.2850	4	0.2779	5	0.1978	6	0.0426	5	0.1127	5	
	GOO	0.6096	1					0.6822	2					
	MST	0.5745	2					0.7622	1					
es-fr	PRO	0.3487	4	0.5509	2	0.4860	3	0.7194	1	0.5789	1	0.6118	1	MTurk
	SYS	0.4199	2	0.6900	1	0.6083	1	0.7111	2	0.2632	2	0.4321	2	winner
	APE					0.5244	2			0.1053	3	0.1652	3	
	SKY	0.1953	5	0.2830	3	0.2823	4	0.0889	5	0.0526	4	0.0865	4	AU $\neq$ UF
	GOO	0.4926	1					0.6500	3					
	MST	0.4181	3					0.6056	4					
es-it	APE									0.7500	1	0.7500	1	
	SYS	0.2866	3	0.4336	1	0.3894	1	0.6222	3	0.2500	2	0.3756	2	
	SKY	0.1511	4	0.2138	2	0.2105	2	0.2148	4	0.0000	3	0.0855	3	
	GOO	0.4323	1					0.6556	2					
	MST	0.4283	2					0.6815	1					
es-pt	APE					0.8035	1							
	SYS					0.5875	2							
fi-en	SUN	0.2484	3	0.3503	1	0.3187	1	0.5667	2	0.9648	1			
	SYS	0.2267	4	0.3247	2	0.2962	2	0.2333	4	0.0352	2	–		
	GOO	0.4096	1					0.9222	1					
	MST	0.3489	2					0.5370	3					

LP	vendor	AU-1		AU-1re		AU-2		HE		UF		final		comment
fr-de	SYS	0.2589	3	0.3893	1	0.3462	1	0.7524	2	0.2791	1	0.4278	1	AU ≠ UF
	PRO	0.2138	4	0.3140	2	0.2825	2	0.7444	3	0.2791	1	0.4190	2	
	LGT	0.2118	6	0.3067	4	0.2516	4	0.6556	5	0.2326	3	0.3614	3	
	LIN	0.2134	5	0.3099	3	0.2703	3	0.6714	4	0.1628	4	0.3261	4	
	TRD	0.1184	8	0.1644	5	0.1502	5	0.2873	7	0.0233	5	0.1152	5	
	SKY	0.1223	7	0.1631	6	0.1500	6	0.2206	8	0.0233	5	0.0952	6	
	GOO	0.3455	1					0.7635	1					
	MST	0.2832	2					0.6524	6					
fr-en	SYS	0.3413	3	0.5192	1	0.4751	1	0.6963	3	0.4110	1	0.5030	1	AU ≠ UF
	PRO	0.2987	4	0.4484	2	0.4105	2	0.6722	4	0.3836	2	0.4729	2	
	LGT	0.2505	5	0.3556	3	0.3257	3	0.5204	5	0.0959	3	0.2462	3	
	SKY	0.2057	7	0.2809	5	0.2588	5	0.4185	6	0.0685	4	0.1925	4	
	TRD	0.2201	6	0.3102	4	0.2751	4	0.3870	7	0.0411	5	0.1683	5	
	GOO	0.5055	1					0.8130	1					
	MST	0.4754	2					0.7352	2					
fr-es	PRO	0.4286	4	0.7059	2	0.6586	2	0.5389	3	0.5172	1	0.5379	1	AU ≠ UF
	SYS	0.5148	2	0.8660	1	0.8115	1	0.6028	2	0.4138	2	0.5103	2	
	APE					0.6172	3			0.0690	3	0.1473	3	
	SKY	0.1887	5	0.2729	3	0.2488	4	0.0417	5	0.0000	4	0.0374	4	
	GOO	0.5467	1					0.8278	1					
	MST	0.4702	3					0.5333	4					
fr-it	SYS	0.3658	3	0.5830	1	0.5266	1	0.7926	1			0.2904	1	MTurk
	SKY	0.1492	4	0.2078	2	0.2195	2	0.3444	4			0.1253	2	winner
	GOO	0.4292	1					0.7185	2					
	MST	0.4112	2					0.6630	3					
fr-ru	PRO	0.1305	4	0.1714	2	0.1572	2	0.5148	1	0.6923	1	0.5855	1	MTurk
	TRD	0.1478	3	0.1987	1	0.1875	1	0.1667	4	0.3077	2	0.2534	2	winner
	MST	0.1776	1					0.3333	3					
	GOO	0.1653	2					0.5074	2					AU ≠ UF
hu-en	SYS	0.2525	3	0.3676	1	0.3416	1	0.3630	4	0.0674	2			AU ≠ UF
	MOR	0.2177	4	0.3012	2	0.2686	2	0.7778	2	0.9326	1	–		
	GOO	0.3539	1					0.8037	1					
	MST	0.3288	2					0.6333	3					
it-de	SYS	0.2063	3	0.3014	1	0.2721	1	0.4296	3			0.1561	1	
	SKY	0.1123	4	0.1499	2	0.1453	2	0.1370	4			0.0556	2	
	GOO	0.3463	1					0.7741	1					
	MST	0.3160	2					0.6148	2					
it-en	PRO	0.2532	5	0.3741	3	0.3389	3	0.5489	4	0.3657	1	0.4180	1	AU ≠ UF
	SYS	0.3462	3	0.5311	1	0.5010	1	0.6422	3	0.2500	3	0.3928	2	
	LGT	0.3256	4	0.4891	2	0.4425	2	0.5311	5	0.2639	2	0.3619	3	
	SKY	0.1913	6	0.2634	4	0.2529	4	0.2244	6	0.1204	4	0.1649	4	
	MST	0.6237	1					0.7111	2					
	GOO	0.6143	2					0.7644	1					

LP	vendor	AU-1		AU-1re		AU-2		HE		UF		final		comment
it-es	APE									0.7000		0.7000	1	
	SYS	0.3446	3	0.5387	1	0.4918	1	0.5074	3	0.3000		0.3814	2	
	SKY	0.1664	4	0.2312	2	0.2477	2	0.2000	4	0.0000		0.0848	3	
	GOO	0.5649	1					0.7889	1					
	MST	0.5417	2					0.6519	2					
it-fr	SYS	0.3642	3	0.5809	1	0.5236	1	0.7741	1			0.2846	1	MTurk
	SKY	0.1185	4	0.1685	2	0.2631	2	0.0667	4			0.0463	2	winner
	GOO	0.4887	1					0.7222	2					
	MST	0.4404	2					0.6704	3					
lv-en	SYS	0.2978	3	0.4479	1	0.4135	1	0.1852	4					
	TRD	0.2169	4	0.3023	2	0.2684	2	0.3333	3			–		
	MST	0.5257	1					0.4630	2					
	GOO	0.4793	2					0.5667	1					
no-en	SYS	0.5167	3					0.3370	4					
	GRA	0.4831	4					0.5630	3			–		
	GOO	0.9781	1					0.8481	1					
	MST	0.6816	2					0.6444	2					
pl-de	PWN	0.1264	3	0.1869	1	0.1757	1	0.6481	2	0.7833	1	0.6820	1	
	TRD	0.0789	4	0.1101	2	0.1072	2	0.3667	4	0.2167	2	0.2508	2	
	GOO	0.3284	1					0.7778	1					
	MST	0.1662	2					0.4815	3					
pl-en	PWN	0.2264	4	0.3541	2	0.3262	2	0.6528	2	0.5463	1	0.5562	1	AU $\neq$ UF
	SYS	0.2660	3	0.3820	1	0.3484	1	0.5167	4	0.2209	3	0.3224	2	
	TRD	0.1607	5	0.2365	3	0.2396	3	0.5278	3	0.2328	2	0.3220	3	
	GOO	0.5179	1					0.8028	1					
	MST	0.3018	2					0.4583	5					
pl-fr	SYS	0.1179	3	0.1681	1	0.1536	1	0.2407	4					
	TRD	0.0742	4	0.0977	2	0.0902	2	0.4259	3			–		
	GOO	0.3932	1					0.9111	1					
	MST	0.2220	2					0.5963	2					
pl-ru	TRD	0.0287	3	0.0618	2	0.0858	2	0.6963	2	0.5577	1	0.5521	1	MTurk
	PWN	0.0371	1	0.0796	1	0.0881	1	0.7556	1	0.4423	2	0.5009	2	winner
	GOO	0.0302	2					0.6630	3					
	MST	0.0090	4					0.3111	4					AU $\neq$ UF
pt-en	SYS	0.3355	3	0.5023	2	0.4664	2	0.5278	3	0.4211	1	0.4576	1	AU $\neq$ UF
	PRO	0.3355	3	0.5095	1	0.4703	1	0.5083	4	0.2895	2	0.3732	2	
	LGT	0.2889	5	0.4213	3	0.3934	3	0.4472	5	0.2895	2	0.3472	3	
	MST	0.6802	1					0.7611	1					
	GOO	0.6737	2					0.6694	2					
pt-es	APE					1.0000	1			0.8333		0.8571	1	
	SYS					0.7149	2			0.1667		0.2450	2	

LP	vendor	AU-1		AU-1re		AU-2		HE		UF		final		comment
ru-de	PRO	0.1344	3	0.2091	1	0.1552	1	0.6222	2	0.7826	1	0.6717	1	
	TRD	0.0876	4	0.1339	2	0.0831	2	0.2963	4	0.2174	2	0.2276	2	
	GOO	0.2602	1					0.7852	1					
	MST	0.2304	2					0.6185	3					
ru-en	PRO	0.1710	4	0.2669	2	0.2460	2	0.5528	3	0.4308	1	0.4489	1	AU $\neq$ UF
	TRD	0.1560	5	0.2373	3	0.2243	3	0.4583	5	0.3651	2	0.3790	2	
	SYS	0.2345	2	0.3452	1	0.3173	1	0.6083	2	0.2041	3	0.3367	3	
	GOO	0.3232	1					0.6528	1					
	MST	0.2131	3					0.5222	4					
ru-fr	PRO	0.1632	3	0.2574	1	0.2325	1	0.6407	2			0.2155	1	
	TRD	0.0922	4	0.1363	2	0.1208	2	0.1481	4			0.0565	2	
	MST	0.2768	1					0.6074	3					
	GOO	0.2645	2					0.8074	1					
ru-pl	TRD	0.1272	2	0.1990	2	0.1868	2	0.7296	1	0.5172	1	0.5479	1	MTurk
	PWN	0.1588	1	0.2570	1	0.2568	1	0.6667	2	0.4828	2	0.5154	2	winner
	GOO	0.1245	3					0.6000	3					
	MST	0.0786	4					0.2778	4					AU $\neq$ UF
sl-en	SYS	0.3104	3	0.4746	1	0.4409	1	0.2778	4	0.1519	2			AU $\neq$ UF
	AME	0.2325	4	0.3250	2	0.3473	2	0.5815	3	0.8481	1			
	GOO	0.4670	1					0.8852	1			–		
	MST	0.4007	2					0.6259	2					
sv-en	SYS	0.3338	3					0.4185	4					
	GRA	0.2149	4					0.5185	3					
	MST	0.5444	1					0.6963	2			–		
	GOO	0.5103	2					0.7963	1					
tr-en	SYS	0.1926	3	0.2791	1	0.2570	1	0.4593	3					
	SKY	0.1529	4	0.2055	2	0.1878	2	0.3889	4					
	GOO	0.3040	1					0.8926	1			–		
	MST	0.2511	2					0.6815	2					
uk-en	TRD	0.2643	3	0.4330	1	0.4252	1	0.5926	2	0.5909	1			
	SYS	0.2481	4	0.3790	2	0.3633	2	0.4778	3	0.4091	2			
	GOO	0.4478	1					0.6556	1			–		
	MST	0.4474	2					0.4000	4					
zh-en	SYS	0.1747	3	0.2378	1	0.2162	1	0.5630	3			0.1905	1	
	LGT	0.1622	4	0.2173	2	0.1993	2	0.4815	4			0.1644	2	
	GOO	0.2266	1					0.7963	1					
	MST	0.2115	2					0.6444	2					

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision	Date	Author	Organisation	Description
v1.4	3th July 2012	Bálint Sass, Csaba Oravecz	RIL	update
v1.3	27th June 2012	Csaba Oravecz	RIL	update
v1.2	4th April 2012	László Tihanyi	RIL	update, evaluation based on user feedback
v1.1	1th April 2012	Bálint Sass	RIL	update, human evaluation
v1.0	28th March 2012	Csaba Oravecz	RIL	core, automatic evaluation

Statement of originality:

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