Title Page

Authors: Nora Aranberri, Gorka Labaka, Arantza Díaz de Ilarraza, Kepa Sarasola

Title: *Ebaluatoia*: crowd evaluation for English-Basque machine translation

Affiliation and address: IXA Group, Faculty of Computer Science, University of the Basque Country, Paseo Manuel de Lardizabal 1, Donostia-San Sebastian, Spain

Corresponding author: Nora Aranberri

Email: nora.aranberri@ehu.eus,
Phone: +34943018397

Abstract

This work explores the feasibility of a crowd-based pair-wise comparison evaluation to get feedback on machine translation progress for under-resourced languages. Specifically, we propose a task based on simple work units to compare the outputs of five English-to-Basque systems, which we implement in a web application. In our design, we put forward two key aspects that we believe community collaboration initiatives should consider in order to attract and maintain participants, that is, providing both a community challenge and a personal challenge. We describe how these aspects can comply with a strict methodology to ensure research validity. In particular, we consider the evaluation set size and the characteristics of the test sentences, the number of evaluators per comparison pair, and a mechanism to identify dishonest participation (or participants with insufficient linguistic knowledge). We also describe our dissemination effort, which targeted both general users and interest groups. Over 500 people participated actively in the *Ebaluatoia* campaign and we were able to collect over 35,000 evaluations in a short period of 10 days. From the results, we complete the ranking of the systems under evaluation and establish whether the difference in quality between the systems is significant.

Keywords

machine translation, crowd evaluation, pair-wise comparison, English, Basque

1 Introduction

Machine translation (MT) is considered one of the key technologies to help preserve and promote linguistic diversity within the emerging information society (META)[1]. System development requires many natural language processing (NLP) tools and/or vast quantities of parallel texts of the working languages. Developing MT systems is hard and becomes even more challenging for under-resourced languages.

With the scarce resources invested in development, there's little left for MT evaluation, let alone human evaluation. Development heavily relies on automatic scores, which allow for quick and relatively cheap evaluation rounds. But human evaluation is unavoidable if the systems are to be made available for human use. Whether for system quality comparisons, usability tests or intensive error identification, human assessment is indispensable to check the view (prospective) users have of the developing systems.

Ultimately, human evaluation remains the most reliable source to check progress on translation quality. If we leave purpose-oriented methods aside and focus on checking progress, large MT evaluation campaigns have tried several approaches over the years. For example, the annual Workshop on Statistical Machine Translation (WMT[2]) shared-tasks (2006-2014) have tested a good number of evaluation methodologies: ranking of translated sentences relative to each other (2007-2014); contrastive adequacy and fluency evaluation (2006-2007); ranking and correct/incorrect allocations of syntactic constituent translation (2007-2008); assessment of edited versions (2009-2010). These large-scale campaigns are able to collect over 20,000 evaluations per language pair. They engage participants in the shared task, interested volunteers, trusted friends of the community and sometimes a small number of paid annotators to perform the assessments and gather expert responses. Again, this ensures, to a certain degree, the reliability of the responses (although it does not necessarily reflect the views of end users). As it is apparent from the experience of the different campaigns, these put great emphasis in ensuring the reliability of the responses. In the editions 2012-2013 assessments were posted in Amazon's online marketplace, Mechanical Turk, for anyone to access and get paid for contributing to the evaluation. This was stopped in following campaigns given the low inter-annotator results obtained by "Turkers".

Although a couple of exceptions have cropped up over the years (see Urdu-English and Haitian Creole-English in WMT2011), WMT-type campaigns do not include small languages within their translation pairs. Researchers working with small languages, therefore, find themselves having to put an evaluation environment in place and recruit evaluators when limited resources are available to cover such costs. Besides, it is often the case that finding suitable evaluators is difficult for these languages. To put forward just a couple of examples of how small languages deal with MT evaluation, in a seminal work on Estonian-English translation, researchers report one human evaluator providing assessment for 250 segments on a 3-point scale for each of 6 systems trained on different corpora (Fishel, Kaalep and Muischnek 2007). This is quite a considerable work load to handle by a single person, and also, the results are based on the opinion of that single judge. For English-Latvian, a research paper addressing free word ordering reports a pair-wise system comparison performed by a single evaluator for 790 output pairs and an error analysis of 100 sentences (Khalilov, et al. 2010a). Again, we are faced with the assessment of one single person, who has taken the task to evaluate a very large set of segments. Evaluation of Spanish-Basque systems so far has focused on rather limited usability testing within a science and technology QA and CLIR contexts (Arrieta et al. 2008), on obtaining 2 to 3 responses for about 100 segments by in-house linguists or professional translators (Labaka et al. 2014) or on human-targeted translation error rate calculations of 100 segments (Labaka 2010).

We can also find a number of more robust studies which have managed to gather resources to embrace evaluation for small languages. Such is the case of a recent paper on statistical machine translation of Latvian[3], Lithuanian and Estonian, which reports a costly error classification of 1,000 sentences per

[1] META - Multilingual Europe Technology Alliance http://www.meta-net.eu/

[2] WMT workshops http://www.statmt.org/wmt14/

[3] Latvian - 2 million speakers; Lithuanian - 3 million; Estonian - 1.2 million. Data from: http://ethnologue.com

language together with automatic metrics (Khalilov, et al. 2010b). Also, Aranberri et al. (2014) performed a preliminary productivity test with professional translators and regular users for the English-Basque pair by embedding their system within a translation management workflow at Elhuyar[4].

Overall, in all cases, evaluations are performed by a very limited number of evaluators, often by members of the research teams themselves, and for very limited segments. This is by no means to imply that these efforts should be dismissed, but rather to emphasize the need to come up with new affordable ways to put MT evaluation in practice.

In this work we propose using a web-based evaluation application for crowd-based MT system comparison. Given a mindful, dynamic community, it is a relatively fast method to obtain assessments within the accepted reliability range without the need for a huge investment. Communities from small, under-resourced or endangered languages tend to display a marked awareness and willingness to honestly contribute to initiatives that will allow their languages to survive the technological age. A regular user who is a speaker of a small language can perform a pair-wise comparison with a certain amount of reliability, and mechanisms can be put in place to identify those who do not. This evaluation method allows researchers to discriminate between techniques that render noticeable progress and those which do not. Additionally, it provides an unprecedented opportunity to collect feedback from prospective users.

We report an experiment that emerged from the need to evaluate a number of English-to-Basque MT systems developed during the ENEUS project (FP7-PEOPLE-2011-IEF-302038). Aiming for a free generalist system for public use, we were specifically interested in checking whether evaluators perceived differences in quality between the various approaches developed. Given the characteristics of a comparison evaluation, the large amount of evaluations required for meaningful results, and considering the advantage of having prospective users serve as evaluators, we decided to involve the community in the evaluation. We opted for a large-scale crowd-based human evaluation campaign, *Ebaluatoia*, where we collected regular users' opinions and tested the reaction of a small community towards MT evaluation.

We believe that the evaluation initiative, which has allowed us to gather invaluable data that fall within the accepted reliability standards in the field, is highly reproducible for other small language communities. The web application is designed to attract and monitor the performance of the crowd while collecting comparison data from prospective users. The initiative exploits the high degree of implication communities from minority languages show and lays out the foundations for general community-involvement in MT evaluation.

The remaining is organized as follows: Section 2 describes the experimental setup where considerations for the evaluation method, test and control sets and evaluators are discussed, as well as the MT systems evaluated during the *Ebaluatoia* human evaluation campaign. Section 3 describes the web application and the user experience. Section 4 presents the evaluation campaign results, including inter-annotator agreements on overall *Ebaluatoia* results. Section 5 summarises the conclusions drawn from the crowd-based evaluation experience and suggests avenues for improvement and future work.

2 Experimental setup

Alegria et al. (2013) describe a crowd-based collaborative initiative to enrich the Basque Wikipedia by post-editing original Spanish articles. They asked volunteer participants to download a tuned version of the OmegaT [5]translation memory, which included access to their MT system, and to provide Basque post-edited versions. While they did collect valuable translation resources during the nine months of the campaign, they report a total of 30 participants, with only 20 completing substantial work. They link this low contribution to the inconvenient set-up and the intensive work the task required for regular users. Their experience highlights the importance of considering the effort involved in the task, as well as the difficulty of attracting and maintaining an active community.

[4] Elhuyar Language Services http://www.elhuyar.org/EN

[5] OmegaT: http://www.omegat.org/

In the experimental setup described in the following sections, we ensure that research requirements, integrity and validity are met while considering a simple and attractive setup for participants. We first focus on the evaluation method chosen and the compilation of the test set and control sentences, and then detail the evaluators' profile to finally present the MT systems compared during the campaign.

2.1 The evaluation method: pair-wise comparison

Considering the lessons from Alegria et al. (2013), we aimed to present as simple a task as possible that would meet our research goals and opted for the pair-wise comparison method. In this evaluation method, participants are presented with a source sentence and two machine translations. The only thing they need to decide is which of the two is better. This method requires lower cognitive effort than other methods and we therefore expect higher inter-annotator agreements. For example, the ranking of a higher number of translations involves remembering and comparing several outputs and this was thought too hard for participants. Having hundreds of people evaluate an attribute, be it fluency, adequacy or suitability, on a scale was also discarded. Each person might have different expectations and standards that may influence their responses even if an exact definition is provided for each scale point. Also, there would be no guarantee that the evaluators actually read the instructions and paid detailed attention to them. A targeted usability test was also rejected. Usability tests work best when a specific context of usage is exploited during the evaluation. However, we aim for a more general quality overview and do not intend to test the systems in a particular domain or context.

The pair-wise comparison provides a simple setup from the evaluators' perspective. With just one simple question "Which of the two translations below is better?" and three segments – the source and two machine translations – we obtain a straightforward answer. The evaluators can choose between three different answers, that is, they can vote for any of the two translations or claim that both are of equal quality. This last option was discouraged (an explicit note was made right next to the option to remind them of it) as we prefer evaluators to take a stance and do not equivocate whenever possible. Yet, this option is necessary as two MT outputs might effectively be of equal quality or even exactly the same.

A simple evaluation method, however, should not be detrimental to our research needs, and clearly our choice could be criticised for being less informative than other methods. The set goal for the evaluation, however, is not that of establishing the quality or usefulness of the translations, but rather that of checking whether there is noticeable difference from one system to the other. And the pair-wise comparison is sufficient to fulfil our goal. The evaluation will reveal which systems output higher quality translations.

2.3 The test set

The test sets in industry-based experiments tend to be representative texts of the companies involved and similarly, usability tests also lead to representative texts for the task at hand (Aranberri and O'Brien 2009; Plitt and Masselot 2010; Mitchell, Roturier and Silva 2014). In general MT research, however, not much thought is usually put into the test set content. For example, as is the norm, the WMT campaigns from 2006 and 2007 used a set put aside from the training corpus for evaluation. This changed in the campaign of 2008, where news stories from the previous months were also included in the test set. From 2009 onwards the test sets consist of updated news stories alone.

While the domain of news is certainly attractive for users, no further constraints are set to select the evaluation segments. For a crowd evaluation to be successful, however, sentences should be attractive to keep participants engaged but they should also be manageable, that is, not excessively long, complete and understandable on their own so that evaluators do not feel confused when rating them.

The candidate sentences for the evaluation test were selected based on the following premises:[6]

[6] Candidate sentences to be included in the final test set were manually reviewed to ensure compliance with the premises.

i. Sentences should have between 5 and 20 tokens (both inclusive). This would ensure manageable pieces of texts for evaluators while covering a range of sentence lengths for research analysis.
ii. Sentences should be full sentences with at least one verb. This excludes software paths, formulae, verbless headlines and incomplete bullet points.
iii. Sentences should be grammatical.
iv. Sentences should not include code or hidden variables.

From a more research-oriented perspective, we decided to include both in-domain (sentences from the training corpus set aside for evaluation purposes) and out-of-domain data (sentences covering topics different from those in the training corpus) in the evaluation. This would allow us to compare the corpus-based systems' performance under both scenarios and also check the stability of the rule-based system across domains.

To compile the test set, we first turned to our English-Basque parallel corpus. Made available by Elhuyar for research purposes, it consists of around 14 million English tokens and 12 million Basque tokens of IT software and documentation, academic books and entertainment web data. Over 85% of the content was obtained from translation memories (TM), hereafter the Elhuyar subcorpus, and the remaining 15% was automatically crawled from the Web using PaCo2 (San Vicente and Manterola 2012), hereafter the Paco subcorpus. As described in Section 2.5, this data was used to train our statistical machine translation systems after putting aside a test set for evaluation purposes. Based on the premises listed above, we extracted a total of 225 sentences from the Paco and Elhuyar subcorpora test sets, 200 and 25, respectively.

The remaining sentences were out-of-domain data. We collected them from the BBC News website and online magazines (BBC's Capital, Hello!, MTV), again, following the above-listed premises. We chose these sources in an attempt to collect well-formed appealing sentences.

The final test set consisted of 500 sentences. It included the following subsets:

- 200 sentences from the test set of the Paco subcorpus

  The Kukuxumusu Drawing Factory launches its first collection of suitcases and travel bags.
  Both are ideal starting points for excursions towards Mount Gorbeia.

- 25 sentences from the test set of the Elhuyar subcorpus

  We often lose sight of the fact that air has mass and exerts pressure.
  Beneath the epithelium is a lamina propria rich in elastic fibers.

- 50 sentences from the BBC news website. The first sentence (which met the premises listed above) of three pieces of news under each of the 12 headings on the main menu were included, as well as sentences on the sports and weather sections.

  Eleven students have been expelled from a school in southern California for allegedly hacking teachers' computers and changing their grades.
  A fragile ceasefire is now in place in the capital Kiev.

- 25 sentences from magazines (Hello!, MTV).

  Miranda Kerr is the new face of H&M's SS 14 campaign.
  Here's another chance to catch Lady Gaga in London as she brings her artRave tour to town.

- 200 sentences from the BBC's Capital – complete articles excluding sentences that did not meet the listed premises

  In a handful of countries, it's legal.
  A young giraffe at Copenhagen Zoo has been euthanised to prevent inbreeding.

2.4 The control sentences

Pinpointing outliers is necessary to ensure the reliability of responses, especially when working with non-experts. As much as possible, we should identify dishonest performance or insufficient linguistic knowledge of participants to discard their contribution. There are different ways to address this, which can be implemented before, during or after the evaluation is completed. Participants with insufficient knowledge can be detected through a qualification test before starting the evaluation. If a number of correct answers are known, participants who fail to provide such answers can be discarded during the evaluation task. Or, if correct answers are not known, responses diverging considerably from the average answer can be discarded during data analysis. To mention an example, the approach taken by the 2013 WMT campaign incorporates two methods. First, Turkers complete a qualification test in order to be admitted to the evaluation campaign. Secondly, evaluations that had previously been responded by experts with a high consensus were included in the crowd task and HITs (basic working unit with three ranking tasks) from Turkers who encountered at least 10 of these controls and failed more than 50% of them were discarded.

Discarding evaluations after the campaign was over was considered too risky for our setup. Not knowing what the response of the community would be, it might be the case that, after filtering outliers, the remaining valid responses be too few to ensure research validity. We opted for a way to discard participants while performing the evaluation. We presented control sentences in every fifth contribution of an evaluator, who was dismissed if over a third of the responses were incorrect (see further details in Section 3). To do so, we needed some pre-established answers for a number of control evaluation units. Control sentences do not ensure that the answers to the evaluation sentences are honest, but at least they monitor, to a certain extent, whether the evaluators are reading the source and translations when completing the task.

Source sentences for the controls were gathered from the training corpus and the web and followed the same premises as the test set sentences. The two translation alternatives were created as follows: one was a manually created translation, a correct translation that followed the source sentence structure as closely as possible; the other was the translation given by our RBMT system (see Section 2.5.3) worsened with negations, antonyms or unrelated words (see Table 1). Any evaluator with a basic level of English and Basque who reads both translation alternatives can clearly see that the human translation is better.

Control sentences serve a double purpose. First, as mentioned, they monitor participant performance. Additionally, they provide participants “time to breathe”. The machine translations that participants judge will most probably include a good number of mistakes and will often be difficult to read. Also, it might be the case that two outputs are very similar. Deciding between them will be difficult, even more so when the translations include many mistakes. This puts a considerable strain on participants. Encountering sentences where the answer is clear from time to time makes the task more bearable.

Source:	Imagine you're at your doctor's surgery.
Better:	Imagina ezazu zure medikuaren kontsultan zaudela.
Worse:	Irudi ezazu zu zarela zure mediku kirurgian.
Source:	Stick on a fake moustache, add some glasses, dye your hair and perhaps pop on a hat.
Better:	Jarri gezurrezko bibote bat, gehitu betaurreko batzuk, tindatu ilea eta again jantzi kapela bat.
Worse:	Bibote sintetiko batean jar ezazu, betaurrekoak gehi itzazu, zure ilea tinda ezazu eta beharbada eztanda egin ezazu txapel batean.

Table 1: Two of the control sentences shown to evaluators.

2.5 The evaluators

With a few exceptions (ACCEPT project: Roturier, Mitchell and Silva 2013; Roturier, Mitchell and Silva 2014), machine translation evaluators tend to be the researchers themselves, linguists or translators, often students of such disciplines, which might be hired to complete the task. When budgets are tight and a large set of evaluations needed, hiring a couple of experts for the job is out of the question. In this work, we propose to engage the prospective user communities of the MT systems at hand. We believe that the language awareness of small language communities makes them more prone to getting involved in this type of initiatives, even more so with the excitement machine translation usually generates among regular users. Also, because of the importance attributed to the language, participants are likely to take the task seriously, increasing the reliability of the responses.

In our pair-wise comparison, we ask participants to give their opinion about the difference in quality between two translations. Each person has his own set of standards and expectations, and this increases the subjectivity of the responses. But it is precisely the opinions of the general public that we aim to collect. For the crowd evaluation to be robust then, it is necessary to evaluate a large set of sentences and, given the characteristic of our evaluators (unlimited number of non-expert volunteers), to collect more than one response per evaluation unit. Deciding on the number of responses to be collected for reasonable results it tricky. Therefore, we analysed the volumes collected during the well-established WMT campaigns as reference. The 2014 WMT campaign, for instance, collected an average of 3,000 responses per system, with no clear reference as to the number of source segments used and the number of evaluations obtained per segment. We finally decided on a set of 500 sentences, which needed to be evaluated for 10 system pairs (5 systems), which meant a total of 5,000 evaluations. To compensate for subjectivity, we collected at least 5 responses per source sentence per system pair. As a result, we needed the crowd to complete 25,000 evaluations (with over 5,625 additional control evaluations).

Over 30,000 responses is quite a substantial amount considering that the target crowd is limited when working with small languages. We targeted Basque speakers with knowledge of English that accessed the web. The Basque speaking community is quite limited, with Eustat reporting 789,430 full Basque speakers as well as 541,562 inhabitants with diverging levels of knowledge (data from 2011).[7] We believe that an initiative like *Ebaluatoia* will mainly attract full Basque speakers. To this number, we need to subtract those who do not have sufficient knowledge of English for the task, those who do not access the web regularly, young children and elderly people (even if we did not set any age restrictions), those who are not interested and/or those who we do not reach. The resulting target crowd is clearly not huge. To this, we need to add that the evaluation task, per se, was not particularly pleasant. Most of the translations had mistakes and they were often difficult to read. Therefore, judging the difference in quality between two candidates might prove hard in many occasions.

Expecting regular web users of such a limited community to voluntarily contribute to a tiresome task of considerable proportion is a strong bet. We considered two key aspects to strike a chord with potential participants and boost collaboration, that is, we presented the campaign as both a community challenge and a personal challenge. To address them, we tried giving the evaluation task a game-like feel. We presented *Ebaluatoia* under the motto "*Help us technologize Basque*" and appealed to the language awareness of the community to engage them in an effort to advance in MT development. Also, a dynamic bar chart was displayed on the evaluation page which showed the overall number of evaluations performed up to that moment.

We believe that creating a sense of community helps maintain and even attract new participants. People tend to get involved in an initiative more easily when they see that others are also engaged. For this reason, we did not keep each participant's contribution hidden, but rather openly showed the progress of the evaluation. We incorporated a ranking of contributors that kept updating live within the main evaluation page. It displayed the top 20 contributors, the positions of the current participant and the last comer, together with the username and number of evaluations performed. The ranking shows new participants that other people are engaging in the campaign and returning participants see the changes since they were last active. This ties in with the personal challenge mentioned above. Because the ranking is updated with each contribution the participant provides, we hoped that this would create some rivalry among them and entice them to keep evaluating. Moreover, the top 5 contributors would receive a small token.

[7] Data for the Basque Autonomous Community, which covers the provinces of Bizkaia, Gipuzkoa and Araba – Spain, and excludes other Basque speaking territories such as Nafarroa and the French Basque Country. Report available at: http://www.eustat.es/elementos/ele0000400/ti_Poblacion_de_2_y_mas_a%C3%B1os_de_la_CA_de_Euskadi_por_nivel_global_de_euskera_territorio_historico_y_a%C3%B1o_1996-2011/tbl0000487_c.html#axzz31VXv0z6a

An additional aspect we considered was the effort/compensation balance. We expect people to join the initiative based on their good will for the language, but the prospect of winning a prize is very tantalising – as well as a way to show appreciation for their effort. To put this into practice, we ran a raffle. To every participant, we gave a raffle number for every 10 evaluations. They could see the number of evaluations they had performed and the raffle numbers collected at all times in the evaluation page. Every time they won a new number, a message would display with a notification. The advantage of the raffle is that all participants are included regardless of the extent of their contribution. Those who contribute more will have more chances of winning, but with just 10 evaluations, a participant is already in. A main prize was raffled. Three prize options were offered for the winner to choose from, all within the same price range, to appeal to a wide range of profiles and ages. From a research perspective, prizes (both the small tokens and the raffle numbers) help not only attract evaluators but also obtain a larger set of answers by the same evaluator.

Setting up the evaluation task as a game does not come without its risks. In a rushed attempt to collect more raffle numbers or outperform a rival, participants might overlook their performance – they might race through the source and translations and/or opt for a middle ground "both are of equal quality" answer rather than taking a stand. Yet we expected the control sentences to compensate for this, as well as the institution logos displayed in the evaluation page, which would hopefully remind participants that they were participating in a research activity.

2.5.1 Dissemination

Dissemination is key for the success of a crowd-based initiative. The evaluation campaign has to be publicized properly if it is going to reach regular users and convince them to volunteer to participate. Communication channels also have to be established with the community for a proper interaction and monitoring during the campaign and to distribute follow-up information. We used several channels to disseminate information about the initiative: social networks, mailing lists and direct communication with relevant players.

Two social network applications were targeted: a new Facebook account was created for *Ebaluatoia* and the IXA research group's Twitter account was used to publicize *Ebaluatoia* information. Both services were used to provide up-to-date information during the campaign.

People reached through the Facebook account were general users not specifically targeted for their profiles or interests. People reached through the Twitter account were specialists that may have a specific interest in language technology initiatives and included both developers and users. The Basque Twitter account had 233 followers and the English Twitter account had 82 followers at the time of the campaign. Among them are journalists from different local newspapers and scientific publications; the group for the dissemination of science of technology of the University of the Basque Country; a number of associations for the promotion of Basque in the Administration and online use of Basque; translators, philologists and language centres; staff from different Schools from the University of the Basque Country (Polytechnic School, Faculty of Humanities, Faculty of Computer Science), staff from the Basque Centre on Cognition Brain and Language, the Summer Basque University, the Association of Basque Schools in France; language technology companies; the Basque Foundation for Science (Ikerbasque); and Donostia 2016.

A post publicizing the campaign was sent to the University's on-line news board, a daily announcements mailing list that reaches academic and administrative staff, researchers and students on the three campus of the University of the Basque Country. Several lecturers of Technical Basque at different Faculties also helped spread the initiative. Additionally, groups with a special interest in languages and translation were targeted directly such as EIZIE (Association of Basque Translators, Proofreaders and Interpreters) and the School of Translation of the University of the Basque Country.

Langune, the Basque Association of Language Industries, and Sustatu, an online news weblog, also helped promote *Ebaluatoia* through news entries and the publication of a blog entry, respectively.

2.5.2 Participation and profiles

The *Ebaluatoia* campaign was officially run February 14-25, 2014. It attracted 551 participants. Out of those, 34 (6.17%) did not perform any evaluation and 52 (9.44%) did not pass the control sentences and were therefore not allowed to continue with the task. 465 participants (84.39%) provided valid answers and a total of 26,283 responses were collected, excluding answers from control sentences (Table 2).

The contribution per participants varies significantly. We find 14 super-users, who contributed over 600 evaluations each. Another 16 participants are found in the 250-600 range. 52 evaluated 100-250 sentences whereas another 127 range between 26 and 100 evaluations. Close to half of the participants are found in the 1-25 range, 256 to be precise.

	Number of participants	%
Total participants	551	
Thrown out	52	9.44
With no evaluations	34	6.17
Valid and active participants	465	84.39
Median of evaluations for valid and active participants	17	
Average evaluations for valid and active participants	71.88	

Table 2: *Ebaluatoia* participation summary.

With respect to participants profile, we observe that the dissemination channels had great impact. In terms of age-group (Table 3), the three age-groups covering the 18-45 age range have 25-30% of evaluators each, with the younger group accounting for a slightly larger set. Almost 10% of evaluators are below 18 and just above 10% are older than 45, with 2 in the over 65 range.

Age-group	Number of participants	%
<18	55	9.98
19-25	166	30.13
26-35	134	24.32
36-45	138	25.04
46-55	46	8.35
56-65	10	1.81
>66	2	0.36

Table 3: Number of participants per age-group,

The vast majority of participants (81.30%) have university-level education. 12.70% have secondary-level education, 4.35% report having pursued vocational training and 1.63% gave no response (Table 4). The participants reached by the campaign remain mainly highly educated population.

Level of study	Number of participants	%
University	448	81.30
Secondary School	70	13.70
Vocational Training	24	4.35
Other	9	1.63

Table 4: Number of participants per level of study.

Participants were also asked to specify the field of studies they were pursuing or their job (Table 5). 30.85% of the records belong to the technical field, with humanities following with 18.15%. A specific section was provided for translators, linguists and philologists, which accounted for 17.06% of evaluators. This bias is probably due to the fact that the campaign emerged from the Faculty of Computer Science and it has close links with the Faculty of Humanities and the Association of Basque Translators, Proofreaders and Interpreters.

Field of studies/work	Number of participants	%
Technical Studies	170	30.85
Humanities	100	18.15
Translators, linguists and philologists	94	17.06
Others	75	13.61
Experimental Sciences	49	8.89
Health Services	22	3.99
Services	21	3.81
Social Sciences and Law	20	3.63

Table 5: Number of users per field.

The reported level of English is intermediate for 54.26% of participants (Table 6). An advanced level was reported by 30.85% and an elementary level by 14.88%. These data agree with the overall level reported for Spain, where the population has a B1 overall level according to the English Proficiency Index of Education First (Europa Press, 29th January 2014).

The level is expectedly higher for Basque with 84.21% proficient speakers and 14.15% intermediate-level speakers, and only 1.64% low-level speakers (Table 7). The nature of the task attracts mainly full Basque speakers and therefore the high number of proficient speakers comes as no surprise. Still, the diverse community has also attracted speakers with lower levels of knowledge. According to the Basque Institute of Statistics Eustat (2010/2011 report), 60% of school students pursued their studies fully in Basque (model D) and 22% pursued them following the half Basque-half Spanish model (model B). Students who pursue second-level studies under model D are automatically awarded the B2 level certificate in Basque. Model B students obtain the B1 certificate. Completing a university degree in Basque provides students with the C1 certificate.

Level of English	Number of participants	%
A1-A2	82	14.88
B1-B2	299	54.26
C1-C2	170	30.85

Table 6: Number of users per level of English.

Level of Basque	Number of participants	%
A1-A2	9	1.63
B1-B2	78	14.16
C1-C2	464	84.21

Table 7: Number of users per level of Basque.

2.5 MT systems

The English-Basque MT systems developed during the ENEUS project covered the mainstream approaches in research nowadays. They include two statistical systems, a rule-based system and a hybrid system that combines all the three previous systems. A fifth system was added to this list to include the only publicly available English-Basque MT system at the time, the state-of-the-art Google Translate.[8]

2.5.1 SMT baseline (SMTb)

Our SMT baseline system was a standard phrase-based statistical machine translation system based on Moses (Koehn et al. 2007). As mentioned in Section 2.3, the parallel data to train the system was collected from different sources and formats. The Elhuyar subcorpus (over 85% of the content) was obtained from TMs, and the Paco subcorpus (15%) was automatically crawled from the Web.

We implemented two techniques to clean the corpus automatically. Both subcorpora were filtered for sentence length (we discarded all pairs which exceeded 75 words) and the Paco subcorpus was further

[8] On April 2, 2014, the Basque Government launched *Itzultzailea en-eu*, a publicly accessible online English-Basque system developed by Lucy. Unfortunately, this was weeks after the *Ebaluatoia* was completed and we could not include it among the evaluated systems. Google Translate is available at https://translate.google.com/#en/eu/

cleaned through translation likelihood (TL) filtering based on Khadivi and Ney (2005). After filtering, the final training corpus consisted of 1,296.501 sentences, with 14.58 million English tokens and 12.50 million Basque tokens.

The system was fed with the tokenized corpus for training. It was trained on both subcorpora but optimized on the Elhuyar subcorpus only. Optimization is nowadays a standard final step in SMT building. It was first proposed by Och (2003) and it exploits the automatic metrics that emerged in previous years. His minimum error rate training (MERT) aims to efficiently optimize model parameters with respect to word error rate and BLEU. The models' parameters are automatically tuned for weights to maximize the system's BLEU score on the development set.

Optimization is a way to refine the translation models to translate a specific data set. The Paco subcorpus was thought to be more spurious and noisy than the Elhuyar subcorpus, which is a clean corpus built with manual translations of formal texts. We included the Paco subcorpus for coverage purposes but considered that it would be safer to optimize the system on text that was unmistakeably well-formed and aligned.

2.5.2 SMT with segmentation (SMTs)

STM systems work best with language pairs that are similar, that is, languages that share grammatical features and tend to use similar expressions to communicate meaning. The more similar two languages are, the easier it will be for the system to learn equivalences automatically, and the better an almost word-for-word translation will look. However, when dealing with dissimilar languages, as is our case, things get a little more complex.

In short, languages can express semantic and morphosyntactic information using separate words or joined morphemes. In the case of joined morphemes, languages vary in that in some, each morpheme carries one single piece of information, and therefore, they are used in sequences to express complex meanings, and in others, different morphemes exist for different combinations of information. English is a predominantly analytic language, with distinct words for each morpheme, whereas Basque is a predominantly agglutinative language, with words consisting of a number of morphemes, each expressing a distinct piece of information.

Any effort made towards reconciling the source and the target languages should, in principle, help the word-aligner perform better and thus achieve a better translation. When opposing a predominantly analytic language to a predominantly agglutinative language in SMT, an approach used to draw the source and target languages closer is segmentation (Al-Haj and Lavie 2010, Naradowsky and Toutanova 2011). Segmentation involves splitting a word into its component morphemes. This is usually applied to the agglutinative language, which is the one that tends to join pieces into one word. This will create morpheme sequences that correspond better to the units in the source language, and consequently, make the alignment process easier.

Several segmentation options exist (Habash and Sadat 2006). According to the work by Labaka (2010) on Basque, we can isolate each morpheme, or break each word into lemma and a bag of suffixes; we can establish hand-written rules for segmentation, or let an automatic tool define and process the words unsupervised. Based on his results, we finally opted for the second option and joined together all the suffixes attached to a particular lemma in one separate token. Thus, on splitting a word, we generate, at most, three tokens (prefixes, lemma and suffixes).

The second MT system, SMTs, was built using this technique to address the token mismatch between English and Basque tokens. Following the baseline SMT, we built a standard phrase-based statistical machine translation system based on Moses using the same parallel corpus of 14.58 million English tokens and 12.50 million Basque tokens (now up to 19.22 million token after segmentation). This time, the aligner was fed with segmented words for the agglutinative language.

When using segmented text for training, the output of the system is also segmented text. Real target words are not available to the statistical decoder. This means that a generation postprocess is needed to obtain real word forms. We incorporated a second language model (LM) based on real word forms to be used after the morphological postprocess. We implemented the word form-based LM by using an n-best list, as was done in Oflazer and El-Kahlout (2007). We first ask Moses to generate a translation candidate ranking based on the segmented training explained above. Next, these candidates are postprocessed. We then recalculate the total cost of each candidate by including the cost assigned by the new word form-based LM in the models used during decoding. Finally, the candidate list is re-ranked according to this new total cost. This somehow revises the candidate list to promote the ones that are more likely to be real word form sequences. The weight for the word form-based LM was optimized at Minimum Error Rate Training (Och 2003) together with the weights for the rest of the models.

2.5.3 RBMT (Matxin ENEUS)

Matxin ENEUS is an English-Basque rule-based machine translation system developed at IXA during the ENEUS project. It is an adaptation of the original Spanish-Basque Matxin system (Mayor et al. 2011) to work with English as source language.[9] The system follows the classical transfer architecture, which involves three main components: analysis of the source language, transfer from source to target, and generation of the target language (Figure 1). It has a modular design that makes the three main components, as well as the linguistic data and programs within each component be clearly distinguishable and independent. At the current stage of development, the Matxin ENEUS prototype can address most simple sentence structures and several complex sentences in their simplest forms, namely, relative clauses, completives, conditionals, and a number of adverbial clauses (time, place and reason).

Analysis component

During analysis, semantic and morphosyntactic information is extracted from the text to be translated. Analysis packages are used in this process. Matxin ENEUS uses the Stanford coreNLP (Klein and Manning 2003; Manning, et al. 2014) for English analysis. Matxin ENEUS collects information about words (POS and morphological flexion), chunks (dependency relationships between chunks, that is, groupings of words that require a postposition or case-marker at different levels according to dependency relations), and sentence type.

Transfer component

The transfer component handles two types of information: lexical and structural knowledge. Lexical transfer is responsible for finding the lemma equivalences in the dictionaries, whereas structural transfer focuses on gathering morphosyntactic features and on moving them to the relevant chunks and words.

The first step in the transfer component is to collect lexical equivalences from the bilingual dictionary. This consists of 16,000 single-word entries and 1,047 multi-word units from the Elhuyar English-Basque dictionary made available for research purposes, which we have enriched with WordNet pairs, rising the number of entries to 35,000. It also avails of a semantic dictionary which includes attributes such as animate/inanimate, substance, vehicle, etc.

Next, a set of rules prepares the information extracted from the analysis component to perform the preposition equivalence selection. Among others, it moves the information about prepositions or case-markers to the chunk node, together with the morphological information of the nucleus of the chunk (number and definiteness in the case of Basque). Prepositions are processed using a purposely-built dictionary. Due to the partial equivalences of English prepositions and Basque postpositions, the equivalence list is enhanced with selection rules that identify the different uses and define contexts that will allow the correct preposition to be selected. In addition, Matxin ENEUS avails of two other sources of information which are used when no selection rule applies: verb subcategorization information and lexicalized syntactic dependency triplets, both automatically extracted from a monolingual corpus (Agirre et al. 2009).

[9] Matxin is an open-source architecture available for download at sourceforge http://sourceforge.net/projects/matxin/ under the GPLv2 license.

Then, the information necessary for the verb phrase transfer is extracted from the sentence. For Basque, this means information about the subject person, the indirect object person and the direct object number. Matxin ENEUS covers most of the tenses in the indicative, for all four paradigms (subject, subject-direct_object, subject-direct_object-indirect_object, subject-indirect_object), in the affirmative, negative and questions, for active and passive voices. The imperative is also included.

Although to a more limited degree, modals can also be handled by the system (one sense per form). It can identify ability (can, could, would), permission and prohibition (must, mustn't, can, have to), advice (should) and probability (may, might, will) for affirmative and negative cases. After verb transfer, a last information movement fixes disagreements or incompatibilities encountered in previous steps.

Generation component

Generation is divided into two main steps. The first sets the internal order of the chunk's elements, as well as that of the upper-level chunks. Next, the information gathered by chunk-nodes is moved to the word that needs to be flexed. In the case of Basque, it is the last element in the chunk that carries all the information about the chunk (postposition or case-marks, number and definiteness, among others). The remaining elements are usually used in their lemma forms. In the second step, morphological generation is performed. Thanks to a morphological dictionary, the tags are interpreted and the lemma is transformed into the appropriate surface form. This process is performed by the morphological dictionary built by the IXA group which uses knowledge from the Basque Lexical Database (EDBL according to its Basque initials).

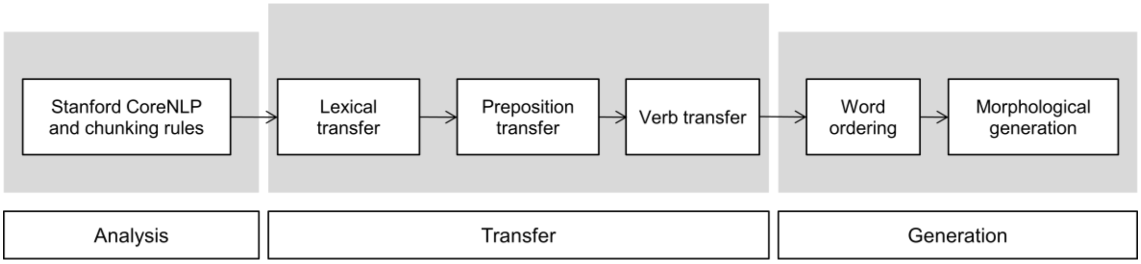


Figure 1: The general Matxin architecture.

2.5.4 Hybrid system (Hybrid)

SMTb, SMTs and Matxin ENEUS were hybridized following España-Bonet et al. (2011) and Labaka et al. (2014). Their method is based on the assumption that RBMT systems excel at syntactic ordering and that SMT systems are more fluent with respect to lexical selection (Figure 2).

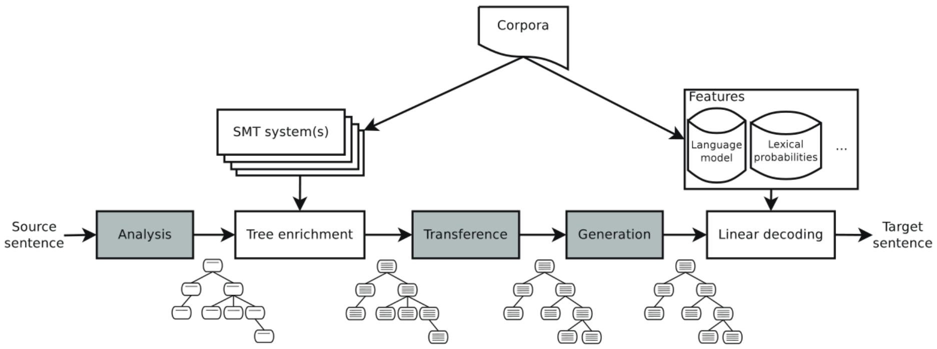


Figure 2: General architecture of Hybrid where the RBMT modules that guide the MT process are highlighted as grey boxes. Figure reproduced from España-Bonet et al. 2011: 3.

The hybrid architecture first uses the tree-structure (a dependency parse tree) from the RBMT analysis. Next, it collects translations for the different phrases from SMTb and SMTs, and after going through the transfer and generation modules, the translations of the RBMT system are also added to the list. For the SMT systems, two types of translations are gathered: the translation of the exact phrase and the translation of the entire subtree dependant on that phrase. Complete subtree translations are collected with the aim to address possible incorrect analysis by the RBMT system. Translation candidates for the exact phrase are collected using two methods (1) the SMT systems are asked for the translation of the exact phrase, and (2) first, the SMT systems are asked for the translation of the whole sentence, and next the source sentence and the translation are aligned; the translation candidates are extracted by collecting the alignments for the exact phrase. Both methods are used because SMT translations are highly dependent on the local context due to the n-gram translation model they use.

Once all the translation candidates are collected, the linear decoder selects the most appropriate fragments (Figure 3). The decoder implemented is a standard Moses decoder that has been modified to block rearrangements.

no se prevé el uso de armas antirreglamentarias, apuntó el consejero de interior

emanaldiak	**ez**	dituzte aurreikusten	**arauz kontrako armekin**	**,**	barne sailburua	baieztatu zuen
jarduera	**ez**	aurreikusten	antirreglamentarias armaz	**,**	barne sailburua	esan zuen
emanaldiak	**ez**	**dira espero**	antirreglamentarias armaz	**,**	herrizaingo sailburuak	esan zuen
					esan zuen barne sailburuak	

ez dira espero antirreglamentarias armaz emanaldiak , esan zuen herrizaingo sailburuak

Figure 3: Translation candidates collected based on the Matxin structure. The first three rows show phrase translations, the fourth row shows a longer phrase translation and the last row shows the translation of the entire sentence. The fragments in bold show the final selection expected from the lineal decoder.

2.5.5 Google Translate (Google)

Google Translate is Google's free online language translation service, one of the most widely used freely available online translation engine. Josh Estelle, a Google Translate engineering leader speaking at Google I/O 2013 revealed that they have reached the 1 billion translations for 200 million users per day barrier.[10]

From its launch in 2001 until around 2005-2006, Google Translate relied on a rule-based engine, Systran, to translate between English and other 8 languages. Starting around 2005, Google Translate begun to work on statistical systems. They participated in a NIST DARPA TIDES Machine Translation Evaluation for the first time in 2005 with their Arabic-English and Chinese-English statistical systems, winning the competition.[11][12] In 2007 Google switched completely to using statistical systems for all languages.[13] It makes use of European Union and United Nations parallel documentation for training, as well as parallel data crawled from the web and obtained under licence agreements.

On 13th May 2010, Basque, together with Azerbaijani, Armenian, Urdu and Georgian was launched as alpha language, bringing the total number of languages in Google Translate to 57.[14] It now supports 80 languages.[15] Since 2008, once a language is made available, one can select to translate between that language and any other that is listed. English is used as a pivot language for those pairs with scarce training data. Not surprisingly, little is known about the intricacies of Google Translate, with the company publishing just enough information to reveal its general approach and latest trends and updates.

Google Translate was, together with the systems built in-house, the only English-Basque MT system that was freely available to users online when the *Ebaluatoia* evaluation campaign took place. It was decided that including this system would give an indication of the relative distance of our systems with regards to

[10] Stephen Shankland for Cnet at http://www.cnet.com/news/google-translate-now-serves-200-million-people-daily/

[11] Ashley Taylor for The Connectivist. Breaking the Language Barrier: Technology Is The Great Equalizer. July 11, 2013. http://www.theconnectivist.com/2013/07/breaking-the-language-barrier-technology-is-the-great-equalizer/

[12] From NIST at http://www.itl.nist.gov/iad/mig//tests/mt/2005/doc/mt05eval_official_results_release_20050801_v3.html

[13] Adam Tanner for Reuters at http://www.reuters.com/article/2007/03/28/us-google-translate-idUSN1921881520070328

[14] From Google Translate Blog at http://googletranslate.blogspot.com.es/2010/05/five-more-languages-on.html

[15] From Google Translate at http://translate.google.es/about/intl/en_ALL/

the only existing reference in terms of quality. Google avails of huge parallel corpora and long experience in building SMT systems and was therefore considered a strong contender.

2.5.6 Overall automatic scores

We calculated system performance using automatic metrics to compare their behaviour against the human evaluation (Table 8). For automatic metrics to be calculated, a reference translation of the source sentences is necessary, as well as the machine translation output. In order to compile the reference translations, we collected existing translations where possible and manually translated the sentences who lacked a Basque version.

Ebaluatoia test set[16]

	BLEU	**NIST**	**TER**	**METEOR**[17]
SMTs	10.41	4.06	85.04	24.45
SMTb	09.84	3.97	86.77	23.65
Google	12.70	4.41	83.87	26.46
Hybrid	08.16	3.78	90.20	22.37
Matxin ENEUS	04.86	3.26	96.02	17.52

Table 8: Automatic scores for the MT systems under evaluation for the Ebaluatoia test set.

All four metrics agree on the system ranking. Google is the best-scoring system, followed by SMTs and STMb. The hybrid system lags behind all full SMT systems and Matxin ENEUS scores poorly. The large difference in scores between the RBMT system and the statistical systems might be due to two reasons. Firstly, the RMBT system's quality is expected to be low given its stage of development. Secondly, automatic scores tend to favour SMT systems over RBMT systems because they do not consider the correctness of the machine translation but rather compare the difference between the MT output and the reference translation.

The overall system ranking according to the automatic metrics is as follows:

Google > SMTs > SMTb > Hybrid > Matxin ENEUS

3 The web application and user experience

Although there are many web-based platforms that allow performing machine translation evaluation, they did not seem to fit with the requirements set out during the setup. Dedicated systems such as Appraise (Federmann 2012) and the Dynamic Quality Framework (DQF)[18] tools, or general systems such as Amazon's Mechanical Turk[19] and Crowdflower[20] have inconvenient login processes. The first two require that an administrator assigns a pre-set job to a pre-known evaluator. This is unworkable on a spontaneous volunteer-based crowd collaboration task where the extent of the contribution depends on the initiative of the participant. In the general systems users can create an account themselves but the validation processes takes long – over two days in certain cases. The delay between the time a user decides to participate and the time he can actually start contributing would cost us invaluable users, who would most certainly give up along the way.

In order to smooth the user experience, implement the game-like elements we hoped would attract users and control the user performance to ensure sufficient and valid responses for research, we developed our own evaluation platform (http://*Ebaluatoia*.org). The web application (also accessible from mobile phone devices) consists of 5 main stages participants follow during each contribution.

[16] The *Ebaluatoia* test set (see Section 2.3) includes 225 sentences set apart from the corpora used for training, which might benefit our SMT systems.

[17] METEOR scores were calculated using its basic setup, that is, with the language set as "other", without stemming and no link to WordNet.

[18] DQF: https://evaluation.taus.net/tools

[19] Amazon Mechanical Turk: https://www.mturk.com

[20] Crowdflower: http://www.crowdflower.com/

The Homepage or Login page of the site welcomes participants to *Ebaluatoia*. Once in the Homepage, participants can log in directly or register, if accessing the site for the first time. A link to the instructions page is also provided for them to be able to read the details of the campaign without having to register. Additionally, the functionality to reset a forgotten password is offered. The page includes the logo of the initiative as well as the logos of the supporting institutions (University of the Basque Country, the IXA research group, FP7 and the Marie Curie Actions).

When participants decide to get involved in the initiative, they first need to register. This step provides us with contact details as well as information to create participant profiles. It is not our intention to create profile-specific experiences, but rather understand the configuration of the evaluators. The registration form gathers the following information:

- Name – real name of the participant
- Username – name to appear on *Ebaluatoia*
- Email –participant contact information. This is the only contact point with the participants. An authentication email is sent to each registered participant with a link to click on to confirm participation. This should be done before the end of the campaign. Participants can start contributing before they confirm participation. The participants who introduce a fake email address or fail to confirm participation are not included in the raffle.
- Age group – <18, 18-25, 26-35, 36-45, 46-55, 56-65, >65
- Level of studies – Second Level studies; Professional training; Third Level studies; Other.
- Domain of studies – Technical studies; Experimental sciences; Health sciences; Social sciences and law; Humanities; Services; Translators, linguists and philologists; Other.
- Password – to be used to access *Ebaluatoia*
- Level of English (elementary A1/A2; intermediate B1/B2; advanced C1/C2)
- Level of Basque (elementary A1/A; intermediate B1/B2; advanced C1/C2)

After logging in, participants reach the Welcome page. This page welcomes the participants and reminds them of the number of sentences they have evaluated as well as the numbers for the raffle they have collected so far. Participants click the button “Continue evaluating” to proceed.

Participants are next taken to the Instructions page for participation. Minimal instructions explain the objective of *Ebaluatoia*, that is, the evaluation of machine translated sentences. Participants are told about the pair-wise comparison method and that they should give their true opinions. They are warned that control sentences will be presented without notice to ensure that they perform honestly. Also, information about the prizes for top contributors and the raffle is provided: how to become a top contributor, how to obtain the raffle numbers, the prizes and raffle date.

Participants then click on “Show me the sentences” and access the Evaluation page (Figure 4). This is the main evaluation environment. The central part of the page presents the evaluation unit, namely, the evaluation question “Which translation is better?”, the source sentence, the two machine translations and the three possible answers “the 1st translation”, “the 2nd translation” and “both are of equal quality – only if truly necessary” as radial buttons. To the left, a bar showing the total amount of evaluations completed is displayed. To the right, the ranking of contributors is shown. It lists the top 20 contributors, specifies the position of the current participant, as well as the last comer. These two charts are updated every time the participant completes an evaluation. At the bottom of the page, the current participant’s total number of evaluated sentences and the raffle numbers collected is shown.

The platform is programmed to ensure the evaluation follows a number of conditions necessary for research validity.

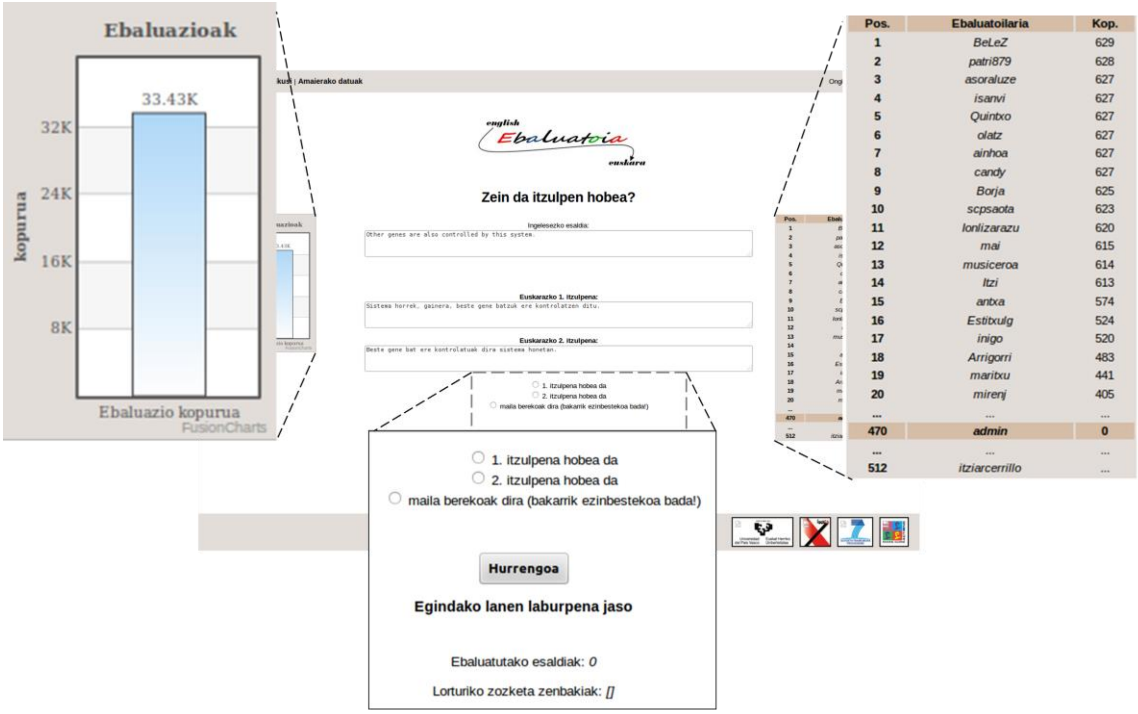
- Each source sentence is only shown to an evaluator once to avoid the response to be influenced by previously seen translation candidates.

- The two machine translations – or translation options in control sentences – are displayed randomly to avoid the order in which translations for each system pair are presented influencing the response.

- 5 evaluations per system-pair and source sentence must be collected. This means that 25 responses are necessary for a source sentence to be "complete". To ensure that as many sentences as possible are completed during the established period for the campaign, once a sentence is displayed for a first time, the system tries to fill this in before displaying a new one. In other words, when a participant asks for a new evaluation, the system displays the source sentence with the highest number of responses that the particular participant has not yet seen.

- When a participant evaluates for the first time, the 1st and 2nd sentences presented are control sentences. From then onwards, every 5th sentence is a control sentence. As with source sentences, the same control sentence is not to be shown to the same participant more than once.

- If a participant does not answer the control sentences correctly, he will not be allowed to continue collaborating. It is compulsory to successfully answer the first two control sentences. From there onwards, control sentence failure has to be kept below 1/3 for the platform to keep the participant in. The recount for success is only performed at every 10th sentence, that is, right before giving the participant a new raffle number. This avoids participants guessing when the control sentences are provided or identifying them. If a participant falls below the success threshold, the platform shows a message "Sorry but you have not passed the control sentences. Your level of English or Basque might not be adequate for this task. We cannot let you participate in *Ebaluatoia*". The evaluations completed by the participant are erased and are put on hold for a new participant to complete them.



Zein da itzulpen hobea?

Ingelesezko esaldia:
Other genes are also controlled by this system.

Euskarazko 1. itzulpena:
Sistema horrek, gainera, beste gene batzuk ere kontrolatzen ditu.

Euskarazko 2. itzulpena:
Beste gene bat ere kontrolatuak dira sistema honetan.

○ 1. itzulpena hobea da
○ 2. itzulpena hobea da
○ maila berekoak dira (bakarrik ezinbestekoa bada!)

Hurrengoa

Egindako lanen laburpena jaso

Ebaluatutako esaldiak: *0*

Lorturiko zozketa zenbakiak: *[]*

Pos.	Ebaluatoilaria	Kop.
1	*BeLeZ*	629
2	*patri879*	628
3	*asoraluze*	627
4	*isanvi*	627
5	*Quintxo*	627
6	*olatz*	627
7	*ainhoa*	627
8	*candy*	627
9	*Borja*	625
10	*scpsaota*	623
11	*lonlizarazu*	620
12	*mai*	615
13	*musiceroa*	614
14	*Itzi*	613
15	*antxa*	574
16	*Estitxulg*	524
17	*inigo*	520
18	*Arrigorri*	483
19	*maritxu*	441
20	*mirenj*	405
...	...	...
470	**admin**	0
...	...	...
512	*itziarcerrillo*	...

Figure 4: Screenshot of the Evaluation page with enlarged elements showing the overall contribution chart (left), contributor ranking (right), and response options and contribution information (centre).

To continue evaluating, participants click on the "Next" button. This action reloads the page and shows a new evaluation unit. They can log out at any moment by clicking on the "Log out" button at the top right corner. This takes them to the Logout page which includes a summary of their contribution and reminds them that they can return to the site and keep contributing any time.

4 Results

In this section we present the results from the *Ebaluatoia* campaign. We first report the inter-annotator agreement for experiment validity. We then outline the overall quantitative human evaluation results to establish a system ranking and compare this to the automatic metric scores.

4.1 Inter-annotator agreement

We provide the participant agreement scores for the evaluation as a measure of reliability of the comparison task. We measured pair-wise agreement among participants using Cohen's kappa coefficient (K) (Cohen 1960), which is defined as

$$\kappa = \frac{P(\mathrm{A}) - \mathrm{P(E)}}{1 - P(E)}$$

where P(A) is the proportion of occasions in which the participants agree, and P(E) is the proportion of occasions in which they would agree by chance. Note that k is basically a normalized version of P(A), one which takes into account how meaningful it is for participants to agree with each other, by incorporating P(E). The values for k range from 0 to 1, with zero indicating no agreement and 1 perfect agreement.

We calculate P(A) by examining all pairs of systems and calculating the proportion of time that participants agreed that A>B, A=B, or A<B. In other words, P(A) is the empirical, observed rate at which participants agree, in the context of pair-wise comparisons.

As for P(E), it should capture the probability that two participants would agree randomly. Therefore:

$$P(E) = P(A > B)^2 + P(A = B)^2 + P(A < B)^2$$

Note that each of the three probabilities in P(E)'s definition are squared to reflect the fact that we are considering the chance that two participants would agree by chance. Each of these probabilities is computed empirically, by observing how often participants considered two translations to be of equal quality.

Table 9 gives the K values for inter-annotator agreement in the *Ebaluatoia* campaign. The exact interpretation of the kappa coefficient is difficult, but according to Landis and Koch (1977), 0-0.2 is slight, 0.2-0.4 is fair, 0.4-0.6 is moderate, 0.6-0.8 is substantial, and 0.8-1.0 is almost perfect agreement. We see that the scores for all the system pairs range between 0.49 and 0.53, within the moderate agreement range.

System pair	Kappa score
SMT baseline vs SMT with segmentation	0.52
SMT baseline vs Google	0.50
SMT baseline vs Matxin ENEUS	0.52
SMT baseline vs Hybrid	0.50
SMT with segmentation vs Google	0.51
SMT with segmentation vs Matxin ENEUS	0.51
SMT with segmentation vs Hybrid	0.53
Google vs Matxin ENEUS	0.49
Google vs Hybrid	0.51
Matxin ENEUS vs Hybrid	0.51

Table 9: Inter-annotator kappa scores for the comparison results per system-pair.

These scores fall within the accepted ranges of kappa scores obtained in the WMT campaigns. As shown in Table 10, the kappa scores range between 0.168 and 0.494. The 5-output ranking method used in the WMT campaigns is bound to have lower agreement scores than a pair-wise comparison. Yet, we see that our kappa scores surpass the ones reported for the WMT tasks. Another thing to consider is the profile of the participants. In all campaigns, it was shared-task participants who performed the evaluations, i.e. experts and "trusted friends", to a higher or lower extent. WMT12 and WMT13 collected judgements from both shared-task participants and non-experts hired through Amazon's Mechanical Turk. As expected, experts obtained higher kappa scores than Turkers (see WMT13m scores). Despite having a number of experts within the *Ebaluatoia* participants, the majority of the contributors are non-experts, and the scores are considerably higher than those reported for the WTM13 crowd scores.

LANGUAGE PAIR	WMT11	WMT12	WMT13	WMT13r	WMT13m	WMT14
Czech-English	0.400	0.311	0.244	0.342	0.279	0.305
English-Czech	0.460	0.359	0.168	0.408	0.075	0.360
German-English	0.324	0.385	0.299	0.443	0.324	0.368
English-German	0.378	0.356	0.267	0.457	0.239	0.427
Spanish-English	0.494	0.298	0.277	0.415	0.295	—
English-Spanish	0.367	0.254	0.206	0.333	0.249	—
French-English	0.402	0.272	0.275	0.405	0.321	0.357
English-French	0.406	0.296	0.231	0.434	0.237	0.302
Hindi-English	—	—	—	—	—	0.400
English-Hindi	—	—	—	—	—	0.413
Russian-English	—	—	0.278	0.315	0.324	0.324
English-Russian	—	—	0.243	0.416	0.207	0.418

Table 10: Table reproduced from Bojar et al. (2014: 19). Kappa scores for inter-annotator agreement in the WMT shared-tasks11-14. The WMT13r and WMT13m columns provide breakdowns for researcher annotations and MTurk annotations, respectively.

The meaning of kappa scores is blurry and we should be cautious with their interpretation. Let alone if we compare scores for different tasks with different systems, test sets and evaluation methods. Yet we feel that the agreement we obtained in the crowd-based evaluation falls within the accepted range in current research and proves small language communities can act as valid evaluators for pair-wise MT comparisons.

4.2 Overall human evaluation scores

During the evaluation task, participants were presented with a source sentence and two machine translations. Their task was to compare the translations and decide which was better. They were given the options "1st is better", "2nd is better" and "they are both of equal quality". Participants were encouraged to decide for one system and avoid selecting the third option as much as possible.

No further definition of "better translation" was provided. Each participant set their own criteria, their own expectations and standards. It is participants themselves who decide which linguistic features and to what degree are relevant enough to make one translation better than another.

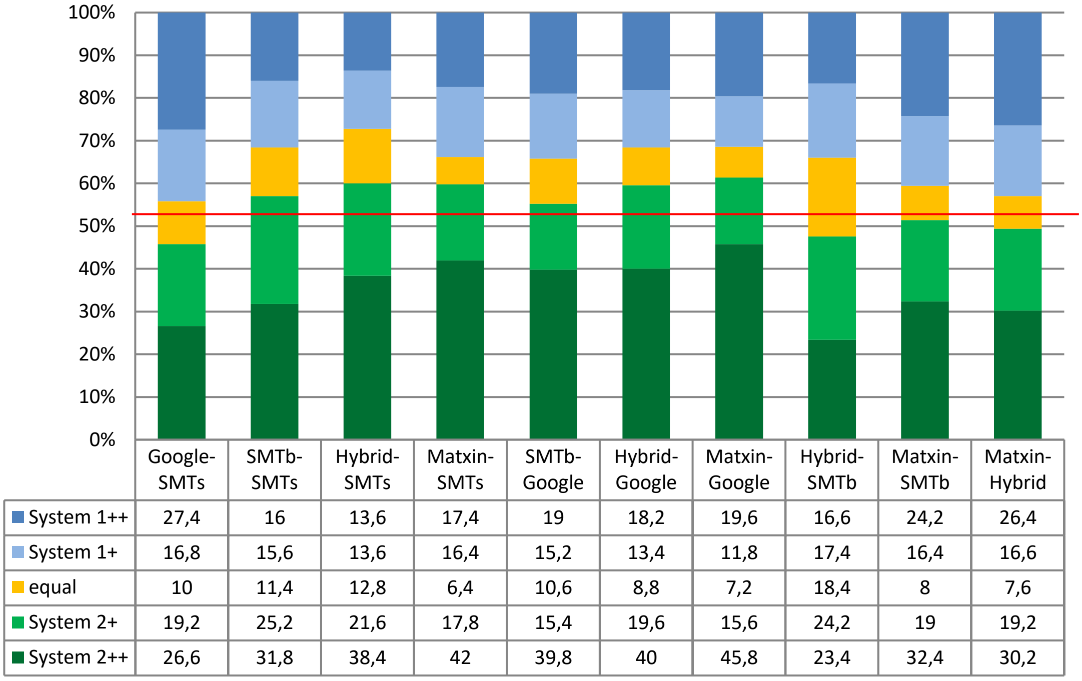
We aimed to collect 5 evaluations per source sentence for each system-pair (2,500 evaluations per pair). However, up to 7 evaluations were collected for some of the sentence/system-pair combinations while waiting for the required evaluations for the whole set to fill in completely (Table 11). Because these are all valid answers, we will consider all evaluations when reporting the results.

	SMTb-SMTs	SMTb-Google	SMTb-Matxin	SMTb-Hybrid	SMTs-Google	SMTs-Matxin	SMTs-Hybrid	Google-Matxin	Google-Hybrid	Matxin-Hybrid
Total evaluations	2635	2632	2660	2653	2600	2630	2623	2616	2618	2616

Table 11: Total evaluations collected per system pair.

We adopted the following strategy to decide on a winning system for each evaluation sentence in each system-pair comparison: if the difference in the number of votes between two systems is larger than 2, we consider the system with the higher number of votes to be the undisputed winner (we code this as "System X++"). If the difference in votes between two systems is 1 or 2, we still consider the system scoring higher to be the winner (we code this as "System X+"). If both systems score the same amount of votes, we consider the result to be a draw (we code this as "equal"). We calculated the statistical significance of the difference in the number of sentences allocated to each cluster for each system pair. The difference is statistically significant at p>0.05 for all systems pairs except for the SMTs-Google pair (p=0.59612) based on a Z-test.[21]

From the evaluations collected during *Ebaluatoia* (Figure 5), we see that the SMTs and Google are the preferred systems against the other competitors. When compared against each other, the difference in sentences allocated to each system is not statistically significant, with only 8 additional sentences allocated to SMTs (229 sentences for SMTs and 221 for Google, 50 equal).



	Google-SMTs	SMTb-SMTs	Hybrid-SMTs	Matxin-SMTs	SMTb-Google	Hybrid-Google	Matxin-Google	Hybrid-SMTb	Matxin-SMTb	Matxin-Hybrid
■ System 1++	27,4	16	13,6	17,4	19	18,2	19,6	16,6	24,2	26,4
■ System 1+	16,8	15,6	13,6	16,4	15,2	13,4	11,8	17,4	16,4	16,6
■ equal	10	11,4	12,8	6,4	10,6	8,8	7,2	18,4	8	7,6
■ System 2+	19,2	25,2	21,6	17,8	15,4	19,6	15,6	24,2	19	19,2
■ System 2++	26,6	31,8	38,4	42	39,8	40	45,8	23,4	32,4	30,2

Figure 5: Percentage of winning sentences allocated to each system in *Ebaluatoia* per system pair.

SMTb lags behind SMTs (158 and 285 sentences, respectively, 57 equal), showing that the techniques to improve statistical MT of morphologically rich languages has been successful, and well noticed and welcomed by participants. It is preferred over Matxin ENEUS (257 and 203 sentences, respectively, 40 equal) and Hybrid (238 and 170 sentences, respectively, 92 equal). The proportion of translations rated as equal for the SMTb-Hybrid pair (18.4%) is the highest across all system-pairs. If we add the high proportion of “System X+” rating obtained to this (59%), we could conclude that the quality difference between these systems is the hardest to decide upon.

Matxin ENEUS is never the preferred system of participants. This is not surprising, as Matxin ENEUS, the rule-based prototype included in the evaluation, currently covers a considerable number of structures but is still far from being a high-coverage high-quality system. However, we see that its output is still considered better than its competitors’ 31-43% of the time. This is a considerable proportion and one that is worth further investigation, in particular for hybridization purposes. It would be invaluable to pinpoint the specific structures in which this system succeeds and its specific strengths against our statistical systems to try to guide future hybridization attempts.

Hybrid is the preferred system only when paired against Matxin ENEUS (247 and 215 sentences, respectively, 38 equal). We see that the hybridization attempt succeeded in improving the RBMT system’s output but did not surpass the statistical system. It is Matxin ENEUS that guides the hybrid

[21] Although primarily a test used for non-parametric variables, a Z-test can be used with parametric variables if it is possible to assume that (1) the probability of common success is approximately 0.5, and (2) the total population is very high (under these assumptions, a binomial distribution is close to a Gaussian distribution).

translation process. Because this is an early prototype with considerable coverage constrains, we can assume that the RBMT foundation of Hybrid will probably be of low quality, and this is detrimental to the SMT systems. However, thanks to the phrase candidates collected from SMTb and SMTs, and their recombination with Matxin ENEUS's output, the final translation is enhanced with respect to the pure Matxin ENEUS translation.

The overall ranking of the systems can be summarised as follows, from better to worse:

SMTs ≈ Google > SMTb > Hybrid > Matxin ENEUS

If we compare the ranking obtained from the manual evaluation and the one proposed by automatic metrics, we see that the statistical systems are assessed differently. Whereas manual results suggest that the difference in quality between SMTs and Google is not significant, automatic scores place Google as the best-scoring system, over 2 BLEU points ahead SMTs and almost 3 points ahead of SMTb. This reveals human evaluators have identified and welcome the changes introduced by the segmentation technique for processing morphology in SMTs but automatic metrics have not been able to recognize and account for the improvements. This indicates that automatic metrics do not always recognize improvements that are relevant for humans and thus demonstrates that human evaluation is necessary and indispensable for reliable MT evaluation.

Ebaluatoia ranking	SMTs ≈ Google > SMTb > Hybrid > Matxin ENEUS
Automatic metrics ranking	Google > SMTs > SMTb > Hybrid > Matxin ENEUS

5 Conclusions

In this work we set to explore the feasibility of running a crowd-based pair-wise comparison evaluation to get feedback on machine translation progress for under-resourced languages. Pair-wise translation comparison, that is, deciding on a better translation between two candidates given a source sentence, is a relatively simple task that a speaker of the target language with knowledge of the source can perform without excessive cognitive effort. From a research perspective, this method can identify differences in translation quality among systems as well as obtain groups of sentences which have been translated better for each MT system. Besides, for generalist systems, the feedback comes directly from prospective users.

We put forward two key aspects that we believe community collaboration initiatives should consider in order to attract and maintain participants. Firstly, the initiative should provide a common goal to achieve as a community (community challenge). In our design we addressed this aspect by appealing to the language awareness of the Basque community and by presenting the initiative as a contribution to help technologize Basque. Also, a chart displaying the overall contribution of the community was embedded in the evaluation platform for participants to follow the progress. Secondly, the initiative should provide an element for self-achievement where each individual participant can improve their own performance (personal challenge). We addressed this by giving the campaign a game-like feel. In the main evaluation page, we displayed the number of evaluations of the participant and a ranking of contributors which included the participants' username, position and number of evaluations. This encouraged rivalry among participants, enticing them to continue collaborating. Additionally, we set a reward mechanism. We ran a raffle for which numbers were obtained according to the evaluations performed, and we also rewarded the top 5 contributors.

All of the considerations mentioned above must comply with research validity. To ensure this, a sizeable evaluation set must be compiled, a set consisting of participant-friendly sentences, and multiple answers for the same comparison pair collected. Also, a mechanism to identify dishonest participation (or participants with insufficient linguistic knowledge) must be put in place. A way to do so while minimally disrupting the user experience is to use frequent control sentences and let go participants who do not pass a success threshold.

Our dissemination effort targeted both general users and interest groups (language service providers and translation associations, language instructors and research centres). The channels used varied from social networks to mailing lists, on-line news boards and weblogs. The participant profiles clearly correspond with the specific efforts.

The *Ebaluatoia* campaign achieved the set goals. The response of the community was phenomenal, exceeding our expectations. Over 500 people participated actively in the evaluation and we were able to collect over 35,000 evaluations in a short period of 10 days. A key aspect to the success was pointed as the use of quick and simple work units.

From the *Ebaluatoia* results, we completed the ranking of the five English-Basque systems under evaluation. According to participants' preferences, Google Translate and the SMT system that uses segmentation score best. The third preferred system is the SMT baseline, followed by the hybrid system, and with Matxin ENEUS holding the last place. The results suggest that the difference in quality between the top two systems is not significant, whereas the quality difference between the remaining systems is noticeable. This shows that humans appreciated the effect of the segmentation technique used over the SMT baseline in contrast to the automatic scores. In the case of Matxin ENEUS, it wins in 31-43% of the sentences, showing that it can contribute to better translation quality.

The comparison results will now be used to guide further research. We aim to perform an exhaustive error analysis to reveal the exact strengths and weaknesses of each system. Also, a structural analysis of successful sentences will show us which type of constructions each system prefers. With the combination of this data, we will not only be able to address specific weaknesses for each of the approaches, but also pursue system combination techniques for hybridization that maximizes the strengths of each approach.

Even when the objectives of the *Ebaluatoia* where met, we learnt that further tuning of the design might improve user experience. This entails both usability considerations for application design and the optimization of the game-like elements. For instance, spontaneous feedback from participants indicate that minor details such as displaying the information of a number of neighbouring contributors in the contributor ranking would encourage further rivalry and provide participants with a clearer view of their overall progress.

The success of the initiative proves that the design was adequate for our community and we believe it is highly reproducible for other small language communities with a similar profile. Also, we aimed for a setup that can be easily replicated for various evaluation cycles, which allows measuring improvement progressively. In this respect, we expect to repeat the campaign to evaluate future systems. Besides, we would also like to extend the methodology to other NLP-related tasks, such as text simplification evaluations or annotation.

Acknowledgments

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007/2013) under REA grant agreement nº 302038 and the Spanish Ministry of Science and Innovation (Tacardi project, TIN2012-38523-C02-01). We also thank the anonymous reviewers for their feedback and valuable suggestions.

References

Agirre E., Atutxa A., Labaka G., Lersundi M., Mayor A., Sarasola K. (2009). Use of rich linguistic information to translate prepositions and grammar cases to Basque. In Ed. Lluís Màrquez and Harold Somers: Proceedings of the XIII Conference of the European Association for Machine Translation (EAMT 2009), pages 58-65. Barcelona, 14-15 May 2009.

Alegria, I., Cabezon, U., Fernandez de Betoño, U., Labaka, G., Mayor, A., Sarasola, K., and Zubiaga, A. (2013). Reciprocal Enrichment between Basque Wikipedia and Machine Translators. In I. Gurevych and J. Kim (Eds.), *The People's Web Meets NLP: Collaboratively Constructed Language Resources*, Springer, Book series *Theory and Aplications of Natural Language Processing*, E. Hovy, M. Johnson and G. Hirst (eds.).

Al-Haj, H. and Lavie, A. (2010). The impact of Arabic morphological segmentation on broad-coverage English-to-Arabic statistical machine translation In Proceedings of the Ninth conference of the Association for Machine Translation in the Americas, Denver, Colorado, October 31 – November 4, 2010.

Aranberri, N. and O'Brien, S. (2009). Evaluating MT output for -ing forms: a study of four target languages. W. Daelemans & V. Hoste (Eds.) *Linguistica Antverpiensia New Series*, Themes in Translation Studies, *8*, 105-122.

Aranberri, N., Labaka, G., Diaz de Ilarraza, A. and Sarasola, K. (2014). Comparison of post-editing productivity between professional translators and lay users. In Proceedings of the Third Workshop on Post-Editing Technology and Practice (WPTP 2014), Vancouver, Canada, October 22 – 26, 2014, pages 20-33.

Arrieta, K., Díaz de Ilarraza, A., Hernáez, I., Iturraspe, U., Leturia, I., Navas, E., Sarasola, K. (2008). AnHitz, development and integration of language, speech and visual technologies for Basque. In Second international symposium on Universal communication, Osaka, Japan, pages 338-343.

Bojar, O., Buck, C. Federmann, C., Haddow, B., Koehn, P., Machácĕk, M., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R. and Specia, L. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Balrimore, USA, June, pages 12–58. Association for Computational Linguistics.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R. and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the 8th Workshop on Statistical Machine Translation, WMT-2013*, Sofia, Bulgaria. pages 1–44, Sofia, Bulgaria.

Callison-Burch, C. Koehn, P., Monz, C., Post, M. Soricut, R. and Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montréal, Canada, June, pages 10–51. Association for Computational Linguistics.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. and Schroeder, J. (2008). Further Meta-Evaluation of Machine Translation. In Proceedings of the Third Workshop on Statistical Machine Translation, June 2008, Columbus, Ohio, pages 70-106.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. and Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. In Proceedings of the Second Workshop on Statistical Machine Translation, June 2007, Prague, Czech Republic, pages 136-158.

Callison-Burch, C., Koehn, P., Monz, C and Zaidan, O. (2011). Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of WMT*, pages 22-64, Edinburgh, Scotland, UK.

Callison-Burch, C., Koehn, P., Monz, C. and Schroeder, J. (2009). Findings of the 2009 workshop on statistical machine translation. In Proceedings of the Fourth Workshop on Statistical Machine Translation, March 30-31, 2009, Athens, Greece, pages 1-28.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocky, M. and Zaidan, O. (2010). Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, Uppsala, Sweden, 15-16 July 2010, pages 17–53.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurment*, 20(1), 37–46.

España-Bonet, C., Labaka, G., Diaz de Ilarraza, A., Màrquez, L. and Sarasola, K. (2011). Hybrid Machine Translation Guided by a Rule–Based System. In *Proceedings of the Thirteenth Machine Translation Summit [organized by the] Asia-Pacific Association for Machine Translation (AAMT2011)*, Xiamen, China. pages 554-561.

Federmann, C. (2012). Appraise: an OpenSource Toolkit for Manual Evaluation of MT Output. *The Prague Bulletin of Mathematical Linguistics*, 98, 130–134.

Fishel, M., Kaalep, H. and Muischnek, K. (2007). Estonian-english statistical machine translation: the first results. In Proceedings of the Sixteenth Nordic Conference of Computational Linguistics (NODALIDA-2007), Tartu, Estonia, May 2007.

Habash, N. and Sadat, F. (2006). Arabic preprocessing schemes for statistical machine translation. In Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, New York, NY, USA, June 2006; pages 49-52.

Khadivi, S. and Ney, H. (2005). Automatic Filtering of Bilingual Corpora for Statistical Machine Translation. In *Proceedings 10th International Conference on Application of Natural Language to Information Systems, NLDB 2005*, Springer Verlag, LNCS, Alicante, Spain, pp. 263-274, June 2005.

Khalilov, M., Fonollosa, J.A.R., Skadina, I., Bralitis, E. and Pretkalnina, L. (2010a).Towards improving english-latvian translation: a system comparison and a new rescoring feature. In Proceedings of the Seventh international conference on Language Resources and Evaluation (LREC'10), Valetta, Malta, May 2010, pages 1719–1725.

Khalilov, M., Pretkalnin, L., Kuvaldina, N. and Pereseina, V. (2010b). SMT of Latvian, Lithuanian and Estonian Languages: a Comparative Study. In Proceedings of the Fourth International Conference: Human Language Technologies - The Baltic Perspective. Riga, Latvia, October 7-8, 2010.

Klein, D. eta Manning, C.D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423-430.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation, *Annual Meeting of the Association for Computational Linguistics (ACL)*, demonstration session, Prague, Czech Republic, June 2007.

Labaka, G. (2010). EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language. Its use in SMT-RBMT-EBMT hybridation. PhD Thesis. University of the Basque Country.

Labaka, G., España-Bonet, C., Màrquez, L. and Sarasola, K. (2014). A hybrid machine translation architecture guided by syntax. *Machine Translation Journal*, 28(2), pages 91-125.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 55-60.

Mayor, A., Alegria, I., Díaz de Ilarraza, A., Labaka, G., Lersundi, M., and Sarasola, K. (2011). Matxin, an open-source rule-based machine translation system for Basque. *Machine Translation Journal*, 25(1), 53-82.

Mitchell, L., Roturier, J., Silva, D. (2014). Using the ACCEPT framework to conduct an online community-based translation evaluation study. In Proceedings of the Seventeenth Annual Conference of the European Association for Machine Translation (EAMT), June 2014, Dubrovnik, Croatia.

Naradowsky, J. and Toutanova, K. (2011). Unsupervised bilingual morpheme segmentation and alignment with context-rich hidden semi-Markov models. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, June 19-24, 2011; pages 895-904.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

Oflazer, K. and El-Kahlout, I. D. (2007). Exploring Different Representation Units in English-to-Turkish Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 25-32, Prague, Czech Republic.

Plitt, M. and Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context, *Prague Bulletin of Mathematical Linguistics*, 93, 7–16.

Roturier, J., Mitchell, L., Silva, D. (2013). The ACCEPT Post-Editing Environment: a Flexible and Customisable Online Tool to Perform and Analyse Machine Translation Post-Editing. In Proceedings of Machine Translation Summit XIV Workshop on Post-editing Technology and Practice, September 2013, Nice, France.

San Vicente, I. and Manterola, I. (2012). PaCo2: A Fully Automated tool for gathering Parallel Corpora from the Web. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 23-25 May, 2012, Istanbul, Turkey.