eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Ebaluatoia: crowd evaluation of English-Basque machine translation

Nora Aranberri

Tutor: Gorka Labaka

hap

Dissertation submitted in partial fulfilment for the requirements of the
MSc in Language Analysis and Processing

September 2014

Departments: Computer Languages and Systems, Computer Architecture and Technology, Computation Sciences and Artificial Intelligence, Basque Philology, and Electronics and Telecommunications.

LABURPENA

Lan honetan Ebaluatoia aurkezten da, eskala handiko ingelesa-euskara itzulpen automatikoko ebaluazio kanpaina, komunitate-elkarlanean oinarritua. Bost sistemaren itzulpen kalitatea konparatzea izan da kanpainaren helburua, zehazki, bi sistema estatistiko, erregeletan oinarritutako bat eta sistema hibrido bat (IXA taldean garatuak) eta Google Translate. Emaitzetan oinarrituta, sistemen sailkapen bat egin dugu, baita etorkizuneko ikerkuntza bideratuko duten zenbait analisi kualitatibo ere, hain zuzen, ebaluazio-bildumako azpi-multzoen analisia, iturburuko esaldien analisi estrukturala eta itzulpenen errore-analisia. Lanak analisi hauen hastapenak aurkezten ditu, etorkizunean zein motatako analisietan sakondu erakutsiko digutenak.

Hitz gakoak: itzulpen automatikoa, ingelesa, euskara, ebaluazioa, bikotekako konparazioa, errore analisia

ABSTRACT

This dissertation reports on the crowd-based large-scale English-Basque machine translation evaluation campaign, Ebaluatoia. This initiative aimed to compare system quality for five machine translation systems: two statistical systems, a rule-based system and a hybrid system developed within the IXA group, and an external system, Google Translate. We have established a ranking of the systems under study and performed qualitative analyses to guide further research. In particular, we have carried out initial subset evaluation, structural analysis and error analysis to help identify where we should place future analysis effort.

Key words: machine translation, English, Basque, evaluation, pair-wise comparison, error analysis

Acknowledgements

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.

Table of contents

Appendices

List of Tables

List of Figures

List of Examples

1 Introduction

As the Multilingual Europe Technology Alliance (META)[1] claims, languages, primary means of communication between humans, are a symbol of identity and as such, language diversity is an invaluable heritage that needs to be preserved. With the advent of globalization, the new socioeconomic interactions have created the need for people from different cultures and languages to communicate, and have challenged its future. Language diversity has appeared as a barrier for successful cross-lingual communication which is often overcome through the use of a lingua franca.

Living in a connected world, however, should not lull us into neglecting and abandoning our native languages and identities in favour of a lingua franca. There are surely ways that can help us establish successful communication and stay connected while preserving our own language and, with it, our cultural identity.

Machine translation (MT) is considered one of the key technologies to help preserve and promote linguistic diversity within the emerging information society. MT is regarded as an indispensable tool in removing language barriers in an effort to achieve international inclusiveness, which allows people to share, access and contribute information across the globe.

Developing MT systems is hard and it becomes even more challenging for low-resourced languages such as Basque. MT system development requires many natural language processing (NLP) tools such as part-of-speech taggers, morphological analyzers and syntactic parsers, to mention but a few, and/or vast quantities of parallel texts of the working languages. The high investment required to build these tools often results in minority languages being neglected and unequipped to survive in the current globalized world. Similarly, the parallel data available for these languages is very limited because most of the text production is done in widely-spoken languages and the information that is imported into minority languages or written in parallel with other major languages is limited.

The META-NET White Papers Series[2] shows that languages within Europe differ substantially in the maturity of research and availability of language processing tools.

> "One of the major conclusions is that Basque is one of the EU languages that still needs further research before truly effective language technology solutions are ready for everyday use. At the same time, there are good prospects for achieving an outstanding position in this important technology area. This development of high-quality language

[1] META is an initiative co-funded by the 7th Framework Programme and the ICT Policy Support Programme of the European Commission that brings together researchers, commercial technology providers, private and corporate language technology users, language professionals and other information society stakeholders to prepare an ambitious, joint international effort towards furthering Language Technology as a means towards realising the vision of a Europe united as one single digital market and information space. http://www.meta-net.eu/meta/about

[2] The Europe's Languages in the Digital Age White Papers Series covering 31 languages is available at http://www.meta-net.eu/whitepapers/overview

technology for Basque is urgent and of utmost importance for the preservation for a minority language as Basque."

The Basque Language in the Digital Age by META[3]

The work presented in this dissertation is part of an effort to maintain a healthy development of resources and NLP research to equip Basque to succeed through the current digital age while empowering speakers to join the emerging information society without losing their identity.

In this work we will deal with MT systems that translate from English into Basque. This language pair is important for two main reasons. Firstly, it allows Basque speakers to access information directly into their mother tongue without having to resort to English or Spanish, major languages for which translations or competitive NLP tools are often available. Moreover, to the extent to which the vast majority of information is nowadays produced in English, English-Basque translation allows Basque speakers to access information directly from the source language. Secondly, having an English-Basque MT system opens up a channel to access information from distant cultures as English being the most developed language in terms of NLP tools, it often acts as pivot language to connect languages that would otherwise not be able to do so.

In particular, this dissertation addresses MT evaluation. This is a step of great importance for guiding and monitoring development. And yet, it is a topic that remains controversial given its subjective nature and because no one-fits-all method is available.

The experiment we report emerged from the need to evaluate a number of English-Basque MT systems developed during the ENEUS project (FP7-PEOPLE-2011-IEF-302038). These were built using different approaches and an extrinsic evaluation was necessary to compare output quality across systems. Because the final users of our systems are to be regular people, it was decided that a large-scale crowd-based human evaluation campaign, a.k.a. Ebaluatoia, would be run to collect their opinion. The results from this initiative would then be analysed to guide further research. This dissertation lays out a number of high-level qualitative analyses (evaluation subset results, basic structural analysis and error analysis) that aim to help identify in which direction we should proceed with deeper analysis to direct future research.

Several side opportunities emerge from the crowd-based methodology chosen for the main evaluation initiative. Ebaluatoia being the first crowd-based evaluation campaign that is run in the area of NLP in the Basque Country, it will allow us to check the response of the community. This first contact will serve to gauge user response and expectations, and set a precedent for possible future initiatives. Secondly, we consider that the campaign can serve as a platform to raise awareness of the importance of research to society, expose the general public to science and research. Participants will face the translations of MT prototypes that will later be available to them online. At the same time, we will help

[3] Excerpt from The Basque Language in the Digital Age — Executive Summary by META-NET as part of META, A Network of Excellence forging the Multilingual Europe Technology Alliance. White papers. Available at http://www.meta-net.eu/whitepapers/volumes/basque-executive-summary-en

promote the IXA research group within the society as well as research funding bodies such as the EU and its Marie Curie Actions.

The remaining work is organized as follows: Section 2 outlines the different approaches to machine translation and the different MT evaluation methodologies used nowadays pointing out their advantages and weaknesses. Section 3 describes the experimental setup where the MT systems evaluated during the Ebaluatoia human evaluation campaign are described as well as considerations for the evaluation method, test and control sets, evaluators, and the web application. Section 4 presents the results. These include overall Ebaluatoia results and automatic metric scores, as well as finer results per evaluation subtest and a qualitative error analysis of the output of all the systems evaluated. Section 5 summarises the conclusions drawn from the evaluation experiment and following analyses, and suggests avenues for future work.

2 Background

In this section we first outline the basic architecture of the different approaches to MT, in particular rule-based machine translation (RBMT) systems and statistical machine translation (SMT) systems, as well as their hybridization possibilities. This will help us better recognize the investment made to build the different systems that took part in the evaluation campaign. Also, the nature of the errors made by each system and the possible ways to fix them will be clearer. Secondly, we present an overview of the evaluation methodologies used within the current task-oriented approach to evaluation, which rejects the previous approach which mustered all efforts in obtaining a high score in a particular method. We briefly revisit attribute evaluation, system ranking, usability testing, error analysis and post-editing productivity. Finally, we address automatic evaluation and briefly present the most popular string-based metrics, namely, BLEU, NIST and TER.

2.1 Approaches to MT systems

MT systems can be defined as "computerized systems responsible for the production of translations from one natural language into another, with or without human assistance" (Hutchins & Somers, 1992: 3). Research on the idea of automatic translation started as early as the 1950s, and various approaches have been proposed throughout the years in the quest for a successful system. Figure 1 shows the periods when each approach was developed and most prevalent.

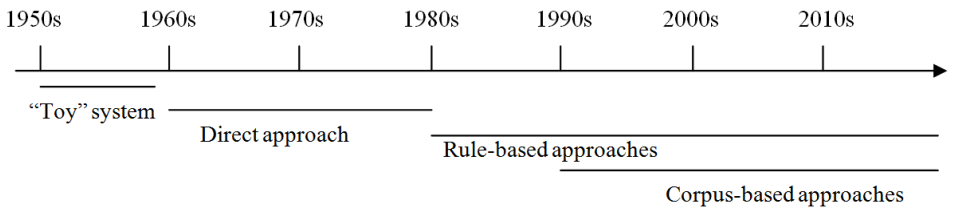


Figure 1: Chronology of machine translation development (from Quah, 2006: 58).

Both rule-based and corpus-based MT systems survive nowadays. RBMT systems rely of manually crafted grammatical rules and lexical equivalences to obtain a translation. Corpus-based systems extract the knowledge required for translation from corpora, without the need for grammatical rules and bilingual dictionaries. We distinguish two architectures within this approach: Statistical (SMT) systems and Example-based (EBMT) systems. The basic principle behind SMT is that resources for translation are extracted from corpora using statistical probabilities of distribution and estimation calculated from words. In EBMT, the goal is to reuse examples of already existing translation chunks as the basis for a new translation. The systems evaluated in Ebaluatoia belong to the RBMT and SMT paradigms, and therefore, those are the ones we focus on in the previous sections.

2.1.1 Basic structure of RBMT systems

Although each RBMT system has its own peculiarities, they most often consist of the following three modules: analysis, transfer and generation. The first module analyses the

source text. This module is of paramount importance as correct analysis of the source is inherently difficult due to ambiguity at different level of analysis, and because often errors at this stage are carried out through the whole translation process, resulting in bad output. The developers of Systran[4], one of the most popular commercial RBMT systems, report that, for their system, 80% of the code belongs to the analysis module, while transfer accounts for 10% and generation takes up the remaining 10% (Surcin et al., 2007). This clearly shows the effort required in identifying the correct syntactic structure and vocabulary.

Analysis

This module performs a grammatical analysis where information about part-of-speech (POS), clause dependencies and relationships between entities of the sentence as well as their functions are extracted (Surcin et al., 2007). Several processes complete a representation of a source segment.

Tokenization and sentence splitting

The first thing an RBMT system does with the input text is to tokenize it, that is, split the text into words or tokens (each of the linguistic units of a text – words, punctuation, numbers, alphanumerics). Next, this information is used to identify sentence units. Tokenization is usually a relatively easy task for languages where words are delimited by whitespaces and punctuation (Mikheev, 2004). The tokenizer considers a word the sequence of characters separated by whitespace.

Sentence splitting is usually a simple process as sentence boundary markers, e.g. a period, an exclamation mark or a question mark, can be used as clues. Sophisticated segmentation programs make use of clues such as lower and upper cased forms, common and proper nouns, during the disambiguation process.

Part-of-speech tagging

Once the tokenization and sentence splitting are completed, POS tagging is performed, that is, part-of-speech descriptors or tags are assigned automatically to the input tokens. The grammatical analysis starts here with the identification of nouns, adjectives, verbs, etc.

Parsing

In the context of natural language processing (NLP), parsing is defined as "using a grammar to assign a (more or less detailed) syntactic analysis to a string or words" (Carroll, 2004: 233). There are different approaches to parsing (shallow parsing, dependency parsing, context-free parsing) and different algorithms to put each of the approaches into practice (cf. Carroll, 2004). The choice of approach depends on the nature of the translation language pair and the resources available for each language.

Parsers have difficulty with structural ambiguity. One of the possibilities for disambiguating structures is to learn the preferences of a language using the probabilities

[4] http://www.systransoft.com/

extracted from corpora. For example, the likelihood of a pair of words to have a head-modifier relationship can be extracted (Collins, 1996).

Transfer

As the name suggests, the transfer stage contains rules to transform source language (SL) structures and lexis into target language (TL) structures and lexical equivalents. The analysis module is specific to the SL, regardless of the TL. This stage, however, is language pair-specific. A comprehensive comparative study of the source and target languages is carried out, which results in an enormous amount of hand-written grammar transfer rules. The formalism used for the rules depends directly on each system's programming requirements.

Generation

The third and last module of an RBMT system addresses generation. It is responsible for all the necessary syntheses, and word-order rearrangements. The TL lexical equivalents extracted from the dictionaries and the TL grammatical structure obtained in the transfer module and joined together to form a well-formed target text. A specific sub-module modifies the resulting segment to ensure correct TL morphology, agreements and word-order.

2.1.2 Basic structure of SMT systems

SMT systems are trained using corpora, from where they extract the statistical knowledge to perform new translation. The statistics learnt, the system is ready to translate new text automatically. During the training cycle, a translation model and a language model are built. The new translations are then produced by the decoder, which uses a search model algorithm.

The translation model

The aim of the translation model is to create a bilingual dictionary or phrase table which includes the most probable word/phrase pairings, together with their probabilities. These probabilities are calculated automatically from a large parallel corpus using statistical algorithms. Figure 2 shows an example of the phrases selected by the algorithm as potential German alignments for the English in europe. Note how the potential phrases are ordered according to the probabilities, from in europa with 0.8290 to der europaeischen with 0.0034 probability scores.

```
in europa ||| in europe ||| 0.829007
europas ||| in europe ||| 0.0251019
in der europaeischen union ||| in europe ||| 0.018451
in europa , ||| in europe ||| 0.011371
europaeischen ||| in europe ||| 0.00686548
im europaeischen ||| in europe ||| 0.00579275
fuer europa ||| in europe ||| 0.00493456
in europa zu ||| in europe ||| 0.00429092
an europa ||| in europe ||| 0.00386183
der europaeischen ||| in europe ||| 0.00343274
```

Figure 2: Phrase table sample (from Moses statistical machine translation system at http://www.statmt.org/moses/?n=FactoredTraining.ScorePhrases)

Word alignment

The success of the translation model depends on word alignment, which is not a straightforward task. First, large training data, that is, the initial parallel corpus, is required to estimate the pairings (Brown et al., 1990; Manning and Schütze, 2000). Secondly, languages do not correspond one-to-one 1:1 and often one word in a language corresponds to zero 1:0, two 1:2, or more words 1: 2+n, in another. The number of target words that correspond to one source word is called fertility. Thirdly, languages do not only differ in the amount of words used to convey the same meaning, but also in the distribution of these words. The difference in word ordering is called distortion. The alignment algorithms are quite complex, as they try to address all these features.

In a paper written in 1993, Brown et al. described the so-called IBM Models as possible alternatives for carrying out efficient word alignment. Model 1 pairs the selected TL string with the SL string assuming that all positions are equally likely. Model 2 assigns the pairings depending on word order. Model 3 also selects the number of words in the TL string that are to be paired to each of the SL words in the string. In Model 4, the pairing depends on the paired TL and SL words and on the positions of other TL words also paired to a particular SL word. However, Models 3 and 4 present some deficiency by losing part of the probability assigning it to strings that are not TL. Finally, Model 5 overcomes this problem. These IBM Models are widely used nowadays, often through their application in GIZA++ (Och & Ney, 2003).

Other algorithms have also been proposed for word alignment. Frequently they are variations of techniques used to align sentences that deal with a first step of creating bilingual dictionaries for further sentence alignment (Manning & Schütze, 2000).

The language model

The aim of the language model is to gather target language knowledge. It is concerned with the TL only, and therefore statistics are calculated from monolingual corpora. The statistics learnt during this training process are consulted during translation to calculate the most probable order in which words should appear, and whether any should be deleted or new ones introduced. This is how the system becomes fluent.

Several methods can be followed to model the TL. The n-gram model (Brown et al., 1990) is one of the most widespread. The term n-grams is applied to sequences of words, n being the number of words in the sequence. In other words, a 2-gram is a two-word sequence and a 3-gram is a three-word sequence. This model addresses the learning process as a word prediction task. According to the Markov assumption, knowing the last few words in a chain is enough to predict the next word (Manning & Schütze, 2000). Based on this, all the possible n-grams (usually 1 to 4) in the monolingual corpus are listed according to their probability of occurrence (ibid). It could be argued that the higher the n-gram level, the more fluent the output will be. However, two things should be kept in mind when using long n-grams: (1) long sequences get repeated less often than short sequences. This might lead to long sequences having lower probabilities, and therefore not being considered by the system despite being beneficial for the final translation. The algorithm

should compensate for that; (2) the number of calculations required for long strings is enormous and combinatorial explosion problems may arise (Arnold, 2003).

Other language models include clustering and probabilistic parsing methods (Manning & Schütze, 2000). The former groups similar words based on the distribution of neighbouring words. The latter performs an automatic grammatical analysis of the target sentences during the training process and replicates the structures during translation.

The search model

The goal of the search model is to find the best translation probability given a source probability following Bayes' theorem (Brown et al., 1990).

$$P(T \mid S) = \frac{P(T)\,xP(S \mid T)}{P(S)}$$

where P(T|S) is the probability of T-translation given S-source, P(T) is the probability of T translation (information obtained from the language model), P(S|T) is the probability of S given T (information obtained from the translation model) and P(S) is the probability of S-source. Because the source is given, P(S) = 1, and therefore, it can be omitted from the equation. Also, because the goal is to find the highest P(T|S), the final equation looks like this:

$$\arg\max \; P(T \mid S) = \arg\max(\; T, P(S \mid T)P(T))$$

Simply put, the process works as follows: the sentence to be translated S is searched for in the existing phrase table. If the sentence is not present in the table, which is the most likely case, shorter source phrases which cover the sentence offering the best probability P(S|T) are selected. From here onwards, the computation to obtain the best joint probability (phrase table x n gram table) considering different phrase table and n-gram options begins. In the process, previously fixed degrees of distortion and fertility are allowed (Somers, 2003).

Most current SMT systems both research prototypes and commercial systems are based on Moses (Koehn et al., 2007). It is an open-source SMT decoder that can be trained with any parallel data one might have. It has a strong and active research community behind it and it has received much attention from industry and funding bodies recently (MosesCore FP7- Grant Agreement 288487).

2.1.3 Hybrid systems

Hybrid systems combine different systems and/or approaches to exploit their strengths. We distinguish two methods for hybridization: system combination and system selection. System combination usually merges statistical and linguistic approaches, exploiting the strengths of both paradigms. Systems are combined either by modifying an SMT system with components of an RBMT system (Eisele et al., 2009) or by modifying an RBMT system with components of an SMT system (España-Bonet et al., 2011).

For example, Eisele et al. (2009) use a standard Moses SMT system and enrich its phrase table entries with data obtained from translating the corpus with several RBMT systems. The final phrase table includes phrase equivalences by the SMT system and RBMT systems. The new translation decoding is carried out by the SMT system as usual. In an opposite attempt, España-Bonet et al. (2011) rely on an RBMT system's dependency parse tree for the new translation's structure and then enrich the different chunks with translation candidates from an SMT system. A decoder then selects the RBMT or SMT phrase candidates based on a previously-defined set of features.

In the system selection method, we first translate a sentence using several MT systems and then decide which of the translations is of better quality to offer this as final output. For example, Hildebrand and Vogel (2008) select the candidate that best fits the target language model. They report an improvement of 2-3 BLEU compared to the best single system.

2.1.4 Advantages and disadvantages of MT approaches

The main advantage of SMT systems over RMBT systems is their ease of implementation. SMT systems come with a backbone of algorithms and only need for the developer to feed them with corpora. As opposed to RBMT systems, there is no need for exhaustive comparative linguistic analysis of the SL and TL, no need for rule-writing and dictionary coding, which involves a high investment in human resources and long development periods.

Collecting parallel texts for minority languages such as Basque, however, is a challenging task. SMT systems need vast amounts of data to have sufficient appearances of words to learn their equivalences and to ensure coverage. Research systems for major languages are trained on corpora that range in the 300 million words, unattainable for the English-Basque pair (the larger corpus gathered so far and that presented in this work, consists of around 14 million words). This weakness is accentuated with the agglutinative nature of Basque, which increases the types in the corpus and requires even larger volumes of data to properly learn equivalences.

Another important aspect to highlight about the systems if their ease for improvement. RBMT systems are incremental and deterministic (Senellart, 2007). Their output is consistent because the systems rely on fixed rules and dictionary entries. As a result, mistakes are easily pinpointed and a solution can be designed to correct them. SMT systems, however, are unpredictable (from a human intuitive perspective). Researchers have tried identifying the error types produced by the systems using different schemes. Some focus on the type of corrections (post-edition) the output requires (Dugast et al., 2007), others use a grammar-independent classification which identifies mainly word-level errors such as missing words, word order errors or incorrect words (Vilar et al., 2006; Font Llitjós et al., 2005) and even a combination of both has been suggested (Tatsumi & Sun, 2008). The problem, however, is that even when the errors are described, it is not clear how we should tackle them through corpora.

2.2 Evaluation methods

With the rapid advancement of machine translation in recent years and all, researchers and companies, developing and adopting the technology, a task-oriented approach to choosing the right evaluation model is finally gaining momentum. The players in the translation process, namely, developers, linguists, technical writers, translators and post-editors, managers and users, have varying needs and expect different answers. Here we briefly revisit the most widely-used methods divided into human and automatic metrics.

2.2.1 Human evaluation methods

Much as the different evaluation strategies try to overcome frailties, human evaluation is criticized for being subjective, inconsistent, time-consuming and expensive. The expertise of each individual evaluator appears to affect the judgements – training, experience, familiarity with MT, and personal opinion about MT. The quality of the previous segment might also affect the evaluators' perception of the current sentence; boredom and tiredness are also risky. Stricter guidelines only have limited impact on managing subjectivity and consistency. Yet, humans are still the most reliable source to obtain meaningful informative evolutions. Users are also humans, after all.

Different evaluation methods have been devised over the years that aim to collect information about different aspect of translation.

Error analysis

Evaluators thoroughly review a text to pinpoint errors. This is the most exhaustive of all approaches, as it identifies and locates all errors present in the text. It is also the most time-consuming and requires the most highly trained evaluators. It is an indispensable analysis to identify the exact linguistic problems in the text. Although the quantity (and severity) of errors might be used as an indicator, it does not provide information of the overall quality.

Attribute evaluation

This method is less costly and time consuming to implement than an error analysis and it can help to focus on assessing quality attributes that are most relevant for specific content types and purposes. The two attributes that are most prominently used for evaluation are adequacy and fluency. Adequacy, refers to the extent to which "*the meaning expressed in the gold-standard translation or the source is also expressed in the target translation*" (Linguistic Data Consortium, 2003). Fluency assesses to what extent the translation is "*one that is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker*" (Linguistic Data Consortium, 2003). Attributes are usually rated using a 4- orf 5-point scale.

This method offers a more global view of quality. No specific errors are identified but rather evaluators assess the overall quality of each sentence according to a particular attribute. The cognitive effort involved is lower than in error analysis, but evaluators still

need to decide on quality levels. Also, it is not clear how these attributes can be mapped to different user needs and usage contexts.

System ranking or comparison

In system ranking, evaluators order a number of translations for the same source from best to worst. This method aims to speed up human evaluation and to reduce the cognitive effort involved. Since 2011, this is the human evaluation method used in the annual WSMT shared tasks, where up to 5 translations are shown to an evaluator per source sentence (Callison-Burch et al, 2011; 2012; Bojar et al., 2013; Bojar et al., 2014). It is particularly useful when comparing translations but it does not provide any information on the actual quality of the output.

Usability testing

This method aims to take into consideration the value of the translations, that is, it measures whether a user thinks (by rating usability of a scale) or proves (by performing a task based on translated text) that a translation is of sufficient quality for a particular context. Usability tests are usually expensive to implement and highly depend on the skills and expertise of the evaluators.

Post-editing productivity

This is a very production-oriented metric. Here you compare the time a translator or post-editor spends translating a sentence from scratch with the time he spends post-editing a machine-translated sentence. This method is useful to decide whether a company should adopt the technology. Researchers also benefit from this metric as they can pinpoint the errors made by the MT system by analysing the post-editor's job. Post-editing, however, is a learnt skill and evaluators for this method should be carefully selected.

2.2.2 Automatic evaluation methods

Automatic metrics emerged to address the need of objective, consistent quick and cheap evaluations. According to Barnejee & Lavie (2005), the ideal metric should:

- correlate highly with human evaluations;
- be able to report minor differences in quality;
- be consistent;
- output similar scores for systems with similar performance; and
- be general so that it can be used for different MT tasks, domains and contexts.

These requirements are very hard to meet but a good number of attempts have been done and taken up by the MT community.

Bilingual Evaluation Understudy (BLEU)

In 2002, researchers at IBM launched BLEU – BiLingual Evaluation Understudy (Papineni et al. 2002). This metric is based on precision between MT output and several reference translations, assessing how many of the words in the MT output are contained in the reference translations. To assess the translation quality, BLEU counts the number of n-

grams of varying length (usually up to 4-grams) in the MT output that match the n-grams present in the reference translations. Then it divides each number by the total number of n-grams in the MT output and calculates their geometric average. In the cases where the score for a particular *n* is zero, the metric does not consider the values obtained for the other *n* and reports zero. BLEU, therefore, is not a suitable metric for sentence-level predictions, but rather a text-level scorer.

According to the authors, BLEU accounts for both "fidelity" – as it accounts for the words in the reference present in the MT output – and "fluency" – as higher n-gram matches account for word-order measurement.

We said that precision measures the number of words from the MT output that occur in the references. In order to measure the quality, however, it is also necessary to know the number of words present in the references that occur in the MT output, i.e. recall. This indicates how much of the information from the source is present in the translation, in other words, the degree of fidelity. Given the difficulty of computing this when several reference translations exist, a brevity penalty is introduced to penalise sentences that are too long. This is done at a text level to allow for certain freedom at sentence level.

The National Institute of Standards and Technology metric (NIST)

The advantages offered by a metric such as BLEU are undeniable. As a result, the Defense Advanced Research Projects Agency (DARPA) included it as a measurement in the Translingual Information Detection, Extraction and Summarization (TIDES) programme. The National Institute of Standards and Technology (NIST) reviewed the IBM metric and refined it to address two main issues that had been identified.

Firstly, the scores for different levels of n-grams were added together, rather than multiplied, to combine them. This meant that the metric could handle more variation between the MT output and the reference output, and also address varying sentence lengths. Not sharing a 4-gram would not directly result in a score of 0. Secondly, NIST assumed that less frequently occurring n-grams were more important, that is, more relevant, informative and specific to a particular text. Therefore, they were allocated more value than to recurrent n-grams. This new variant, since called NIST, was reported to obtain better correlations with human evaluations for adequacy and fluency (NIST report, 2002).

Translation Error Rate (TER)

By 2006 the initial enthusiasm for the BLEU and the likes of it started to fade. Researchers revisited the algorithms and questioned their capacity to assess quality and usefulness for an end-user (Callison-Burch et al. 2006). The idea of returning to an edit-distance approach re-emerged (Przybocki et al. 2006) even if it does not capture all the effort post-editing encompasses. Edit-distance metrics account for technical effort but neglect temporal and cognitive efforts (Krings, 2001; O'Brien, 2006).

Edit-distance measurement aims to count "*the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references, normalized by the average length of the references*" (Snover et al. 2006a: 225). Insertions, deletions, and substitutions

of single words and shifts of word sequences are considered edits, each with a penalisation of 1, similar to the incorrect use of punctuation marks or differences in capitalisation.

In order to calculate the edit-distance, Snover et al.'s Translation Error Rate (TER) score calculates the minimum number of edits that would convert the MT output into the reference. Next, the total number of edits is divided by the length of the sentence (see Figure 3). When more than one reference is available, the TER score is calculated separately for each reference and the best score considered.

$$TER = \frac{number\ of\ edits}{average\ number\ of\ reference\ words}$$

Figure 3: Equation for TER, where the number of edits is divided by the average number of words in the reference.

As all metrics that work with references, TER is highly dependent on the specific reference translation it is supplied because it penalises every single difference with regard to it. The score varies considerably depending on the closeness of the reference and the MT output. To explore this, Snover et al. (2006b) asked human evaluators to generate reference sentences that were as close as possible to the MT output. Then, TER was calculated using the MT-oriented references. This new way of calculating TER was called HTER for human-mediated TER. It was expected that the edit-distance between these MT-oriented references and the MT output would be much shorter than when using non MT-oriented references. They demonstrated that the TER score improved by 33% and obtained higher correlations with human judgements than BLEU. However, it was acknowledged that having MT-oriented references available was not workable in practice given the time and human resources required.

The metrics described above are the ones more widely used in MT evaluation campaigns such as CESTA, the ACL Workshops on Statistical Machine Translation or the NIST Metrics for Machine Translation Challenge. However, a great number of other metrics and variants are also used (GTM, Turian et al., 2003; METEOR, Banerjee and Lavie, 2005; PER, Leusch et al. 2003; ROUGE, Lin & Och, 2004; WER, Nießen et al. 2000; etc.). It is worth noting that all these metrics belong to the string-based strand of evaluation metrics. The problem of string-matching metrics is that they compare the machine translated words against a reference text. This means that (1) metrics fail to identify correct translation alternatives to those present in the references; (2) often costly reference translations are needed for the metrics to work (several if possible).

Work has also been done to promote other approaches although they have not managed to establish themselves as reference metrics. Some of these try to extract and compare syntactic and morphological information (see Liu & Gildea, 2005; Giménez & Màrquez, 2007; Owczarzak et al. 2007a, Owczarzak et al. 2007b). Others try to avoid the use of references (see C-score, X-score and D-score, Rajman & Hartley, 2001). More complex machine learning approaches are also being examined (see Russo-Lassner et al. 2005; Albrecht & Hwa, 2007).

3 Experimental setup

The main goal of this work is to run a large-scale human evaluation campaign to compare the English-Basque MT systems developed during the ENEUS project. We will show that the nature of the evaluators shaped many of the decision taken with regards to the experimental setup. In this section we discuss the numerous aspects that need to be taken into account to set up the Ebaluatoia initiative. We first present the MT systems and then describe evaluation method and the test set that was purposely compiled for the campaign. Next we turn to evaluators, describe their potential and limitations, and report on the profiles of Ebaluatoia participants. Finally, we describe the web application.

3.1 The MT systems

The English-Basque MT systems developed during the ENEUS project cover the most popular 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 has been added to this list to include a publicly available English-Basque MT system, the state-of-the-art Google Translate.[5] The Basque Government currently offers a publicly available online English-Basque system, Itzultzailea en-eu.[6] It is a proprietary system developed by Lucy. Despite rumours for an earlier launch, it was finally made public on April 2, 2014. Unfortunately, this was weeks after the Ebaluatoia was completed and we could not include it among the systems to be evaluated. In the following sections we describe each of the systems and outline the research questions that emerge from the evaluation of the different approaches covered by our MT selection.

3.1.1 SMT baseline (SMTb)

Our baseline SMT system is a standard phrase-based statistical machine translation system based on Moses (Koehn et al., 2007). The parallel data to train the system was collected from different sources and formats. Over 85% of the content was obtained from translation memories (TM) made available by Elhuyar Language Services, 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. Each source comes with specific pros and cons. TM pairs are advantageous in that the alignments are correct, as they have been confirmed by a translator during translation. However, each sentence is stored once only, and therefore, real word frequencies are lost. This reduces the variability of word frequencies, and therefore, renders the word alignment process more difficult. Crawled data, in turn, keeps frequency information but the alignment quality is uncertain, as sentence pairs have been matched automatically. Incorrect sentence pairings introduce noise for the word aligner and Moses phrase extractor (Zens et al. 2002), which decreases the accuracy of the translation model of the SMT system.

[5] Google Translate is available at https://translate.google.com/#en/eu/
[6] http://www.itzultzailea.euskadi.net/traductor/portalExterno/text.do

We consider problematic alignments (1) pairs that have been misaligned by the automatic aligner during collection, and (2) pairs that differ from source to target text to such an extent that they will add more noise than value during training (see Example 1). The differences in (2) might not be translation errors per se, but rather the result of content being added or removed from the translation to better suit the audience or as a consequence of transcreation. However, they introduce noise for the aligner to learn cross-lingual equivalences, and therefore, we chose to remove them from the training corpus.

EN: Which lessons can be learnt from Norway?
EU: Esate baterako, lege betekizunak ezartzekoa.

EN: With a lot of effort and good work, the "pastores" or bull-minders, with the help of some of the runners, managed to lead the bull towards the bullring and once in the arena, the bull was quickly led away to the pens, which it reached over a minute after the other bulls had already entered.
EU: Lasterkari eta unaiek ahaleginak eta bi eginez, plazaraino ekarri eta zezentokietan sartu dute atzenean, gainerako zezenak baino minutu eta gehiago geroago.

Example 1: Problematic alignments.

We implemented two techniques for the automatic filtering of problematic pairs: a purely length-based filtering and a translation likelihood (TL) filtering based on Khadivi and Ney (2005). Both subcorpora were filtered for sentence length and the Paco subcorpus was further cleaned through the TL filtering.

For the length-based filtering, we first discarded pairs with more than 75 words. This is a standard sentence length cut-off applied for Moses training, as longer sentences introduce too much variability for the alignment models and phrase extractor. At a second step, we applied an additional length-based filtering based on Khadivi and Ney (2005). They propose three-level length-based rules to control for differing source and target sentence lengths as follows:

- If the target length is shorter than 3 words and the source is more than six time the target's length (or vice versa), filter it out.
- If the target length is 4 to 9 words and the source length is more than 2.2 times the target's length (or vice versa), filter it out.
- If the target length is 10 words or over and the source length is more than 2 times the target length (or vice versa), filter it out.

These rules will allow incorrect alignments that are shown in the difference in sentence length to be identified and discarded. Remember that the source and target language difference ratio that you choose to apply should consider the natural difference between the language pair.

Khadivi and Ney's (2005) translation likelihood method focuses on the content of the pairs and addresses incorrect alignments that do not necessarily show discrepancy in sentence

length. Following their method, we used GIZA++[7] to train the IBM models for word alignment using the whole (uncleaned) corpus. The resulting translation probability dictionary was used to assign a sentence-level translation probability to each aligned pair. These values were then used to rank sentences and identify the weakest alignments.

The accuracy of the aligner is essential for the successful implementation of this method. However, the fact that English is a morphologically poor language (MPL), and Basque is a morphologically rich language (MRL) makes this task even more challenging. The more alike the source and target languages, the greater the chances for good alignments. However, for our working pair, the aligner is often faced with 1-to-many and many-to-1 patterns, as the number of types (different words that occur in the corpus) and singletons (words occurring once only in the corpus) is much higher for MRLs than for MPLs.

In an attempt to address this, the IBM models were trained on three different versions of the corpus. Training 1 used the original tokenized corpus. Training 2 was performed using a segmented corpus (both source and target). Training 3 only considered the alignments of nouns, verbs, adjectives and adverbs (see Table 1). The version that best correlates with the TL scores was used to establish the cut-off threshold for filtering.

Training 1	testua markatzeko atzealdeko modulua deskargatzen
	unloading text markup backend module
Training 2	testu +a markatze +ko atzealde +ko modulu +a deskargatze +n
	unload +ing text markup backend module
Training 3	testu markatu atzealde modulu deskargatu
	unload text markup backend module

Table 1: Examples of the different training versions.

Having a methodology for (quite) accurately ranking sentence alignments is a step towards the automatic filtering of erroneous pairings. However, we still need a strategy to establish the cut-off point. Khadivi and Ney (2005) artificially introduce different levels of noise in the corpus and use this as a pointer to establish the cut-off. However, this is usually unknown when dealing with an opportunistic corpus.

We proceeded as follows: we randomly collected 5 samples of 100 alignments from the Paco subcorpus. We manually evaluated the alignments by assigning correct or incorrect to each pairing. The evaluation revealed an alignment error rate of 84-89%. This is consistent with the reported accuracy of the corpus crawler, set at around 85%. We then analyzed the distribution of the manually evaluated pairings across the corpus to check the amount of incorrect and correct (evaluated) sentences that would be filtered by removing different fractions of the corpus (see Table 2).

[7] http://code.google.com/p/giza-pp/

Initial corpus	sentences		
	158415		
Manual evaluation	total	correct	incorrect
	500	425	75

	Training 1 (as is)				Training 2 (segmentation)				Training 3 (content words)			
% corpus removed (after length filtering)	sentences removed	good alignments	bad alignments	bad / total alignments removed (%)	sentences removed	good alignments	bad alignments	bad / total alignments removed (%)	sentences removed	good alignments	bad alignments	bad / total alignments removed (%)
0	6,250	5	9	64.29	5,051	4	7	63.64	4,161	11	4	26.67
5	13,858	19	25	56.82	12,719	16	32	66.67	11,874	24	**23**	48.94
10	21,467	38	**38**	50.00	20,387	31	42	57.53	19,586	46	40	46.51
15	29,075	68	41	37.61	27,876	51	**50**	49.50	27,299	64	43	40.19
20	36,683	79	45	36.29	35,484	73	52	41.60	35,012	77	48	38.40
25	44,291	105	50	32.26	43,092	92	54	36.99	42,725	99	49	33.11

Table 2: Distribution of 500 manually annotated alignments
for different filtering cut-off points

We first concluded that the TL scores correlated best with the manual evaluation when the training was performed with the segmented corpus (Training 2), as the number of correct alignments discarded per bad alignment is lower. The TL score is not a perfect scorer, and therefore, correct alignments will sometimes be allocated low confidence scores, and vice versa, bad alignments will also score high. In order to decide on a cut-off point to clean bad alignments, we needed to compromise one of two aspects, size or quality. Given the nature of the working languages, we opted for a relatively smaller but cleaner corpus. We considered that removing 15% (plus the additional 5k sentences filtered out through the length-based technique) provided the best good vs. bad alignment filtering ratio.

After filtering the Paco subcorpus with the TL technique, and both the Paco and the Elhuyar subcorpora with the length-based technique, the final training corpus consists of 1,296.501 sentences, with 14.58M English tokens and 12.50M Basque tokens. It includes texts from IT localization software and documentation, academic books and entertainment web data.

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.

3.1.2 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 start to get a little more complex.

Whaley (1997) introduced two indices to classify languages in terms of morphology. The Index of synthesis refers to the amount of affixation in a language, i.e., it shows the average number of morphemes per word in a language. Languages that use separate words to express different semantic and syntactic information are called isolating or analytic languages. The more a language joins together morphemes in a single word, the more it leans towards the synthetic type.

The Index of fusion refers to the extent to which each morpheme carries a distinct piece of information. Agglutinative languages have low index of fusion because they tend to use a separate morpheme for each piece of morphosyntactic information. This means that segmenting a word by piece of information is relatively easy. In contrast, fusional or inflexional languages have a high index of fusion because they tend to use morphemes that combine different information. In this case, it is usually not possible to split morphemes in a way that each sub-unit provides a single piece of information.

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 separate 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. 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: 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 (Labaka, 2010). Based on the results of Labaka (2010), 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 (analytic language) and Basque (agglutinative language) tokens. Following the baseline SMT, we built a standard phrase-based statistical machine translation system based on Moses using the same parallel corpus of 14.58M English tokens and 12.50M Basque tokens. 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 (unsegmentation step) is needed to obtain real word forms. We incorporate 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.

3.1.3 RBMT ENEUS (Matxin)

Matxin is an English-Basque rule-based machine translation system developed at IXA during the ENEUS project. It is a reimplementation of the original Spanish-Basque Matxin system (Mayor et al., 2011). It is an open-source architecture available for download at sourcefoge[8] under the GPLv2 license, which allows accessing and modifying the entire code. 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 (see Figure 4 at the end of the section). 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.

Linguistic data includes dictionaries and rule-sets. Dictionaries gather lexical equivalences, among others. Rule-sets, in turn, mainly gather rules for morphological and syntactic transfer. Programs are responsible for passing the new text to be translated through the dictionaries and rule-sets in an orderly manner to obtain a translation. This architecture makes the integration of new languages relatively easy, as a linguist can update or change the information in the dictionaries and rule-sets without programming knowledge. Dictionary and rule-set management, that is, what programs control, will be the same for every language pair. Needless to say, having an open-source license, it is also possible to change and improve the programs' code.

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 Parser (Klein and Manning, 2003; de Marneffe et al., 2006) for English analysis. The information Matxin collects from the analysis output is as follows:

[8] http://sourceforge.net/projects/matxin/?source=directory

1 Words: words or multi-word units (MWU) are identified and tagged with the following information: lexical form, lemma, part-of-speech (POS) and morphological flexion information. We call this POS and morphological analysis.

2 Chunks: in Matxin, a chunk is defined as a group of words that requires a postposition or case-marker. Groupings can appear at different levels according to dependency relations. Chunks are identified following a set of rules developed at IXA and syntactic information for them is extracted from the Stanford Parser. The analyzer usually considers the main verb to be the root of the sentence. Words are grouped into chunks and the relations between them are specified in a dependency tree. The dependency relationship of the chunk with its parent, and the dependencies of the words within the chunk are specified. This process stands somewhere between a syntactic and a dependency analysis.

3 Sentences: it is the largest translation unit, at this stage of development. This is the maximum context the system avails of to produce a translation. Given a larger text, Matxin splits it into sentences and treats them separately. The analyzer provides information about the type of sentence.

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.

Lexical transfer

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. It has been enriched with WordNet pairs, rising the number of entries to 35,000. The semantic dictionary is searched for additional information (attributes such as animate/inanimate, substance, vehicle, etc.). The bilingual dictionary covers both closed categories (pronouns, determiners, discourse markers, numbers) and open categories (nouns, verbs, adjectives, adverbs, both single words and MWUs).

Preposition transfer

Next, the first movement phase starts (move 1). This 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. This dictionary consists of English prepositions and their Basque postposition equivalences, where the lemmas and morphological tags are specified. The equivalence list includes 66 English prepositions.

But preposition equivalence is not straightforward. The difficulty lies in the partial equivalences of English prepositions and Basque postpositions, that is, different senses of an English preposition are translated using various postpositions in Basque. For example,

the English preposition “by” can be translated by 10 different postposition depending on the context. Therefore, 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. Rules include different types of knowledge. By default, the design of Matxin allows including attributes of the elements that are in direct dependency in the analysis tree (lemma, POS, morphological, syntactic and semantic features). At the time of write-up, Matxin ENEUS avails of 27 selection rules.

Rules are given full priority during selection but are not the only resource the system avails of for preposition selection. In addition, Matxin avails of two other sources of information, which are used when no selection rules apply: verb subcategorization information and lexicalized syntactic dependency triplets, both automatically extracted from a monolingual corpus (Agirre et al., 2009).

Lexicalized triplets contain very precise information, as they specify the exact word (and postposition) with which a verb appeared in the corpus. In the cases where selection rules are not sufficient to decide on an equivalent, this second resource is used. The verb is identified and the lemma to which the postposition needs to be attached is searched for. Should the lemma appear with the verb and carry one of the candidate postpositions, that is selected.

If the previous resource is not useful, verb subcategorization is used. This resource includes, ordered by frequency, a list of the most common postposition and case-marker combinations that appear for each verb, which identifies transitivity. Matxin collects a list of candidate equivalences for all the prepositions that depend on a verb. Next, it uses the subcategorization information to, taking the verb into account, select the combination that best matches and is more frequent.

Verb transfer

Once the equivalences for the prepositions are obtained, the second movement phase (move 2) extracts from the sentence the necessary information for the verb phrase transfer. Basque verbs carry information about the subject person, the indirect object person and the direct object number. All this information is not concentrated in English verbs, and therefore, when translating from English into Basque, information from different elements of the source sentence will have to be moved to the verb chunk. The verb transfer rule-set uses all these information to output the verb lemma and the data tags for the generation component to be able to build the appropriate surface form.

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. The prototype can respond to the following list of English tenses:

Verb tenses	
Present simple	Past simple
Present continuous	Past continuous
Present perfect	Past perfect
Present perfect continuous	Past perfect continuous
Future simple	Conditional simple
Future perfect	Conditional perfect
Imperative	

Table 3: English tenses covered in the Matxin ENEUS prototype.

Although to a more limited degree, modals can also be handled by the system. It can identify the most common modals: 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. Depending on the context, the modals can acquire a slightly different meaning. At the time of writing, only one sense per modal was covered by the system. After verb transfer, a last information movement step fix disagreements or incompatibilities encountered in previous steps.

Complex sentences

With regards to complex sentences, the current Matxin ENEUS prototype can address, in their simplest forms, relative clauses, completives, conditionals and a number of adverbial clauses (time, place and reason).

Generation component

Generation is divided into three main steps. The first sets the internal order of the chunk's elements, as well as that of the upper-level chunks. The internal word other is set by a reduced set of rules, establishing the canonical order of Basque. The order of upper-level chunks is performed by a rule-based recursive process. The default behaviour is to output the canonical order, and so grammatical information about the target language only is used. However, a separate set of rules controls the translation of non-canonical word orders (fronting, clefting) in the source. These mainly focus on identifying the topic of the sentence, which is located right to the left of the verb chain in Basque.

Secondly, the final information movements are carried out (move 4). These move the information gathered by chunk-nodes 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. The generation of verb phrases is more complex. The elements that make up the phrase can follow different patterns and they may have subordinate markers attached to them.

Finally, morphological generation is performed. During the translation process, lemma and morphosyntactic information has been collected in tag sequences. At this point, all the words that need to appear in the translation have been selected, and their lemma and the required morphosyntactic tag sequence have been assigned. 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).

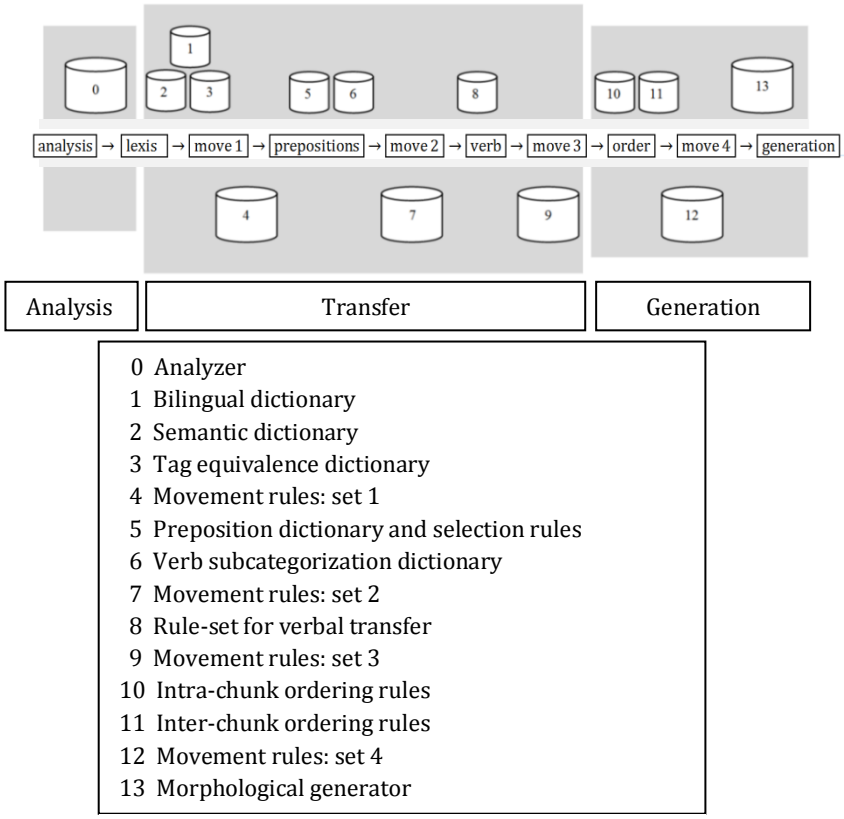


Figure 4: The Matxin architecture and the list of dictionaries and rule-sets it uses.

3.1.4 Hybrid system (SMTh)

SMTb, SMTs and Matxin were hybridized following España-Bonet et al. (2011). Based on the assumption that RBMT systems excel at syntactic ordering and that SMT systems are more fluent with respect to lexical selection, the hybrid translation process is guided by the rule-based engine and, before transfer, a set of partial candidate translations provided by SMT systems is used to enrich the different phrases. The final hybrid translation is created by choosing the most probable combination among the available phrases with a statistical decoder in a monotonic way (See Figure 5).

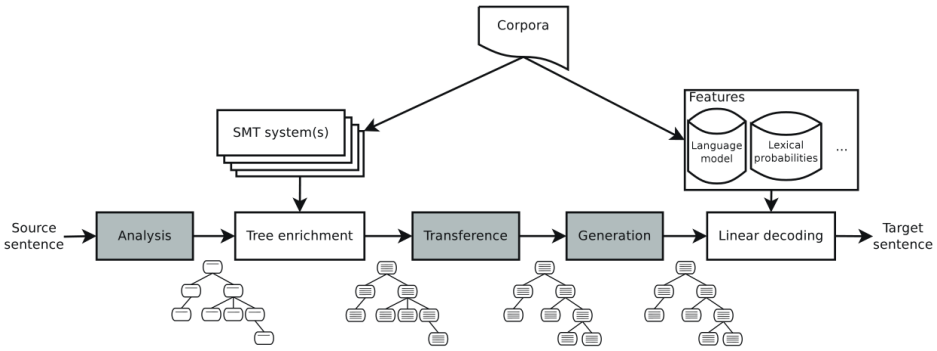


Figure 5: General architecture of SMTh where the RBMT modules that guide the MT process are highlighted as grey boxes. Figure reproduced from España-Bonet et al., 2011: 3.

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, also the translations of the RBMT system. 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 (see Example 2). 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

Example 2: 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.

3.1.5 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.[9]

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.[10][11] In 2007 Google switched completely to using statistical systems for all languages.[12] It makes use of European Union and United Nations parallel documentation for training, as well as parallel data crawled from the web.

[9] Stephen Shankland for Cnet at http://www.cnet.com/news/google-translate-now-serves-200-million-people-daily/
[10] 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/
[11] From NIST at http://www.itl.nist.gov/iad/mig//tests/mt/2005/doc/mt05eval_official_results_release_20050801_v3.html
[12] Adam Tanner for Reuters at http://www.reuters.com/article/2007/03/28/us-google-translate-idUSN1921881520070328

On 13th May 2010, Basque, together with Azerbaijani, Armenian, Urdu and Georgian was launched as alpha language, bringing the total number of languages on Google Translate to 57.[13] It now supports 80 languages.[14] 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 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 very strong contender.

3.1.6 System summary

The five MT systems we have described cover the most common approaches and techniques. Overall, three statistical systems, one rule-based system and a hybrid system will be evaluated (see Table 4). Firstly, SMTb is a pure statistical baseline. Secondly, SMTs is a statistical system trained on segmented target data to address morphologically rich languages. Matxin is the third system, the only purely rule-based system in the lot. The fourth system, SMTh, is a hybrid system that combines the previous three systems. Finally, the fifth system included in the evaluation is Google, a statistical system. Matxin, SMTb, SMTs and SMTh are research prototypes developed in the IXA group, while Google is the search engine company's translation system, which we use as benchmark to compare our systems.

	statistical system	rule-based system
SMTb	yes	no
SMTs	yes	no
SMTh	yes	yes
Matxin	no	yes
Google	yes	no

Table 4: Summary of the 5 MT systems to be evaluated in the Ebaluatoia campaign.

As well as comparing system quality, the evaluation of these systems will allow us to address a number of more specific questions:

- Is the SMT with segmentation better than the baseline SMT? Automatic metrics tend to overlook the contribution of segmentation (Labaka, 2010). We would like to test whether humans perceive the difference.
- Given the limited coverage of the RBMT system, does it always perform worse than the SMT systems? Automatic metrics are not a good option to study this, as they tend to favour SMT systems (Labaka et al., 2011; Bechara and Rubino, 2012), and they cannot be used at sentence-level. Humans will help us identify

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

in which structures, if any, the underdeveloped RBMT system can outperform the SMT systems.

- How does the hybrid system perform in contrast to the SMT systems and the RBMT system it combines? We will check the performance of the configuration proposed for the current hybrid system and compare its quality to the individual systems it combines.
- How far are the research prototypes from Google's engine?

3.2 The evaluation method: pair-wise comparison

Given that the evaluators would be volunteers who access the platform online (see Section 3.5), we aimed to present as simple a task as possible. Therefore, the pair-wise comparison method was chosen for the Ebaluatoia campaign. In this evaluation method, evaluators 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 was chosen because it requires lower cognitive effort than other methods and obtains higher inter-annotator agreements. For example, the ranking of a higher number of translations involves remembering and comparing several outputs and this was thought to be too much hard work for participants. Having hundreds of people evaluate an attribute, be it fluency, adequacy or suitability, on a scale was also rejected. 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 pay detailed attention to them. A targeted usability test was also discarded. Usability tests work best when a specific context of usage is exploited during the evaluation. However, we are aiming for a more general quality overview and do not intend to test the systems for 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. They can vote for any of the two translations or claim that both are of equal quality. This last option was unrecommended (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 machine translations might effectively be of equal quality or even exactly the same (see Example 3).

Question:	Which is better?
Source:	Over eight billion disposable carrier bags are used in England every year.
Translation 1:	Erabili eta botatzeko poltsen gainean zortzi milioi eramaile ingalaterran erabiltzen dira urtero.
Translation 2:	Botatzekoak garraiolari poltsak zortzi milioi Ingalaterran urtero erabiltzen dira.
Response options:	Translation 1, Translation 2, both are of equal quality

Example 3: Evaluation unit.

Our choice could be criticised for being less informative than other methods. The evaluation will reveal whether a system outputs higher quality translations, but we will not gain any insights into the actual quality level. However, we believe that the value of machine translation output should be tested on specific usage environments. Different levels of output quality might be useful for gisting, for post-editing or publication. To mention but a simple example, the quality needed by a general user to follow certain instructions and the quality needed by an expert might vary. Or, a user might be tolerant to not-perfect translations when looking for information online, but would certainly expect printed material to be of high quality. This is not the aim of the current study and therefore, we consider that the pair-wise comparison will help us collect the necessary information for our purposes.

The machine translations that evaluators will judge will most probably include a good number of mistakes and will often be difficult to read. Given that the SMT systems follow similar approaches and that the level of development of the RBMT system is not advanced, it could be the case that the quality of the outputs of two systems is difficult to judge upon. This may be because they are both very similar; or it may be because the poor quality of the outputs makes it difficult to decide which errors are more or less important. This will put a considerable strain on participants, which might result in lower performance. To compensate for this, we decided to introduce control sentences with clear pre-established answers mixed with evaluation sentences (see Section 3.4).

Even if the primary evaluation method for the system comparison is the pair-wise method, we will then use string-based automatic metrics to contrast the results. We will compare whether human evaluation scores match with automatic metrics. Also we will see whether automatic metrics have been able to distinguish the subtle differences between the systems as well as human evaluators do. Additionally, an initial qualitative analysis will be presented through an error analysis method to obtain a linguistically-oriented result that can guide further research.

3.3 The test set

Machine translation evaluation test sets in industry consist of texts that are representative of the type of material the company will be translating with the system. Since the introduction of SMT systems, researchers tend to use a part of the training corpus previously put aside for this purpose. This is referred to as in-domain data. Often, material that is completely foreign to the system is also used to assess the difference in performance between in-domain and unrelated or out-of-domain data in the Ebaluatoia set. We decided to include both in-domain and out-of-domain data. This will allow us to compare the corpus-based systems' performance under both scenarios and also check the stability of the rule-based system across domains.

Another important aspect to consider is the suitability of the sentences for crowd evaluation. Segments should be manageable; not excessively long, complete and understandable on their own so that evaluators do not feel confused. As much as possible, they should include attractive content.

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

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

We first turned to the evaluation set of the training corpora. We trained the SMT systems with text from two subcorpora: Paco and Elhuyar. Based on the premises listed above, we extracted a total of 225 sentences from these sources; 200 from Paco and 25 from Elhuyar (see discarded sentences in Example 4). Remember that the SMT systems were optimized on the Elhuyar subcorpus only and this step seems to affect significantly the final quality of the output. Therefore, although we will consider both sets in-domain data, we do expect variation in quality between those two subsets.

- To create a subjective effect.
- pipelining microinstruction execution in, A-46
- daDT = TimeValue (b2) - TimeValue (a2)
- Search results as from 07/01/2013 in ``Classical music''
- You are in: Home ''Pensioners'' Services ''Applications for Benefits'' Pensions /other national benefits'' Retirement
- 1 11 x 16 cm engraving on a 28 x 41 cm page
- lt; stronggt; Managing your wiki librarylt; /stronggt;
- The filter in the category of other XXXXX Calc filters loads the document in a XXXXX Calc spreadsheet .

Example 4: Discarded candidate sentences from the training corpus.

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 sentences appealing for the general public.

The final evaluation set consists of 500 sentences. It includes the following subsets:

- 200 sentences from the evaluation set of the Paco subcorpus used for SMT training (not MERT)

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 evaluation set of the Elhuyar subcorpus used for SMT training and MERT

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 (covering all news topic range, sports and weather)

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 random sentences from magazines (Hello!, MTV). The first sentence (which met the requirements 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.

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 site (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.

3.4 The control sentences

We used control sentences to monitor the performance of evaluators. Evaluators provide their opinion on the quality of the different systems (they compare system outputs). Therefore, we cannot use their responses as a basis to identify dishonest performance or insufficient linguistic knowledge to stop their contribution during the campaign. We decided to introduce control sentences for which a correct answer is pre-established. 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.

Control sentences were gathered from the training corpus and the web and followed the same premises as the evaluation 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 Matxin worsened with negations, antonyms or unrelated words (see Example 5). Any evaluator with a basic level of English and Basque who read both translation alternatives can clearly see that the human translation is better.

Control sentences served a double purpose. First, as mentioned, they monitored evaluator performance. Additionally, they provided evaluators time to breathe. Deciding between two very similar outputs is difficult, even more so when the translations include many mistakes. Encountering sentences where the answer was clear from time to time makes the task more bearable.

Source:	Humans, as a rule, hate poo.
Better:	Gizakiok, orokorrean, gorroto dugu kaka.
Worse:	Gizakiak gorroto dugu txiza erregela bat bezala.
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 agian jantzi kapela bat.
Worse:	Bibote sintetiko batean jar ezazu, betaurrekoak gehi itzazu, zure ilea tinda ezazu eta beharbada eztanda egin ezazu txapel batean.

Example 5: A number of control sentences shown to evaluators.

3.5 The evaluators

Human evaluation in general is criticized for being subjective. Adding to this, in our pair-wise comparison, we ask evaluators 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. We could perhaps opt for professional linguists or translators to perform the task and thus collect more educated responses. Yet, it is exactly that, the opinions of the general public, that we aim to collect. We aim to uncover whether the MT systems show a distinguishable qualitative difference. Also, for the evaluation to be solid, it is necessary to evaluate a large set of sentences. We decided on a set of 500 sentences, which needed to be evaluated for 5 system pairs. This means a total of 2,500 evaluations. Having one or two people evaluate the whole set was highly impractical and methodologically not sound for various reasons, including intense cognitive effort and familiarity with the evaluation set as the task progresses. Instead, we decided to try crowd collaboration. Participants are volunteers with a sufficiently high level of English and Basque who access the evaluation platform online.

Rather than having a complete set evaluated by a single person, we decided to collect responses by an unlimited number of volunteer participants. To compensate for subjectivity, we collected 5 responses per source sentence per system pair. As a result, we needed the crowd to complete 22,500 evaluations (with the additional >5,625 evaluations required as control sentences).

The target crowd is considerably limited. We target Basque speakers with knowledge of English that access the web. The Basque speaking community is quite limited, with Eustat reporting 789,430 Basque speakers and 541,562 inhabitants with diverging levels of knowledge (data from 2011).[15] 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

[15] 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

any knowledge of English, 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, is not particularly pleasant. Most of the translations will have mistakes and they will often be difficult to read. We expect that opting for a translation over another to 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. It is therefore necessary to take some steps to try to attract participants. We tried giving the evaluation task a game-like feel. To do so, 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 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. We decided to offer different prizes to try to include a wide range of profiles and ages.

Also, we incorporated a ranking of contributors that kept updating live within the main evaluation page. It displays the position, the username and the number of evaluations performed. We hoped that this would create some rivalry among participants and entice them to keep evaluating. Moreover, the top 5 contributors would receive a small token (a USB key). From a research perspective, prizes (both the small gifts 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 – 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 expect the control sentences to compensate for this, as well as the institution logos displayed in the evaluation page, which hopefully remind participants that they are participating in a research activity.

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. The ranking of contributors mentioned above is one of the measures taken. It displays the 20 top contributors, the current participant and the last comer, thus displaying the total number of participants and their activity. This shows returning participants the changes since they were last active and new participants see that other people are engaging in the campaign. Additionally, a bar chart is displayed showing the total number of evaluations performed so far.

3.5.1 Dissemination

Dissemination is key for the success of a crowd-based initiative such as Ebaluatoia. 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.

The Facebook account got 115 likes. People reached through this network are general users not specifically targeted for their profiles or interests. Even if the number of likes may seem rather low considering the amount of friends users tend to have on Facebook, it is quite significant considering that the campaign is only targeting Basque speakers who have a certain level of English, and therefore, the recommendations people make tend to be very restrictive. Also, it should be noted that the number of users who see a piece of information is much larger than the number of users who actually click to like it.

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 at 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. People reached through this network are specialists that may have a specific interest in language technology initiatives and include both developers and users.

A post publicizing the campaign was sent to the University 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.

3.5.2 Participation and profiles

The Ebaluatoia evaluation campaign was officially run February 14-25, 2014. It attracted 551 people who registered. 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 users (84.39%) provided valid answers and a total of 26,283 evaluation responses were collected, excluding control sentences (see Table 5).

The contribution per user varies significantly. We find 14 super-users, who contributed over 600 evaluations each. Another 16 evaluators 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 evaluators are found in the 1-25 range, 256 to be precise.

Total users	551
Thrown out	52
With no evaluations	34
Valid and active users	465
Median of evaluations for valid and active users	17
Average evaluations for valid and active users	71.88

Table 5: Ebaluatoia participation summary.

With respect to user profile, we observe that the dissemination channels have had great impact. In terms of age-group (see Figure 6), 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.

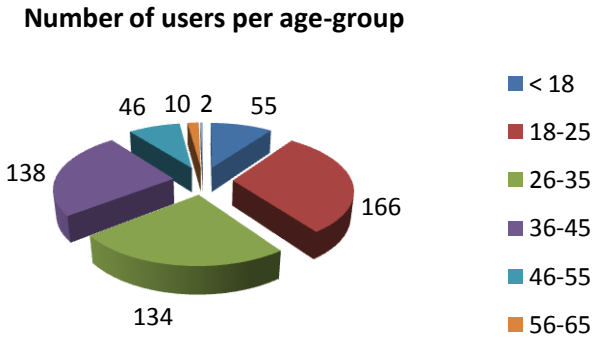


Figure 6: Number of users per age-group.

The vast majority of evaluators (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 (see Figure 7). The participants reached by the campaign remain mainly highly educated population.

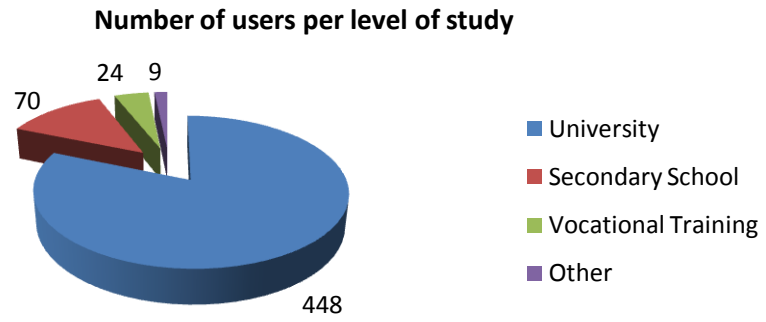


Figure 7: Number of users per level of study.

Participants were also asked to specify the field of studies they were pursuing or their job (see Figure 8). 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.

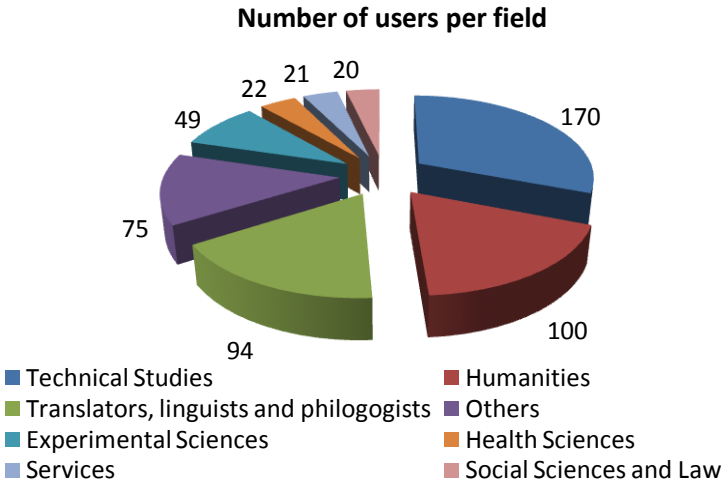


Figure 8: Number of users per field.

The reported level of English is intermediate for 54.26% of participants (see Figure 9). 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. Spain has a B1 overall level according to the English Proficiency Index of Education First (Europa Press, 29th January 2014). The Basque Country obtained the highest score among the autonomous regions with 57.90 points (2012).[16]

[16] 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

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 (see Figure 10). The nature of the task attracts mainly Basque native 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.

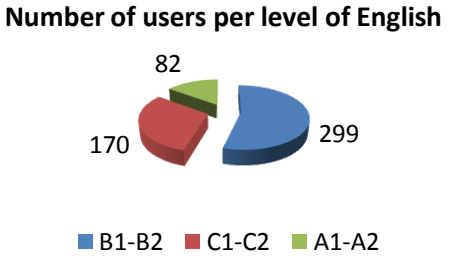


Figure 9: Number of users per level of English.

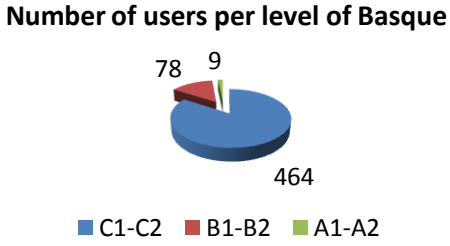


Figure 10: Number of users per level of Basque.

3.6 The web application and user experience

The web application (also accessible from mobile phone devices) was implemented by Elhuyar. It consists of 5 main stages participants follow during each contribution.

The web address www.ebaluatoia.org was distributed for volunteers to join the initiative. The Homepage or Login page of the site (see Figure 11) welcomes participants to Ebaluatoia. Once in the Homepage, participants can log in directly (or register, if accessed 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).

Figure 11: Screenshot of the Login page.

When participants decide to get involved in the initiative, they would first need to register (see Figure 12). 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. 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; Others.
- 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)

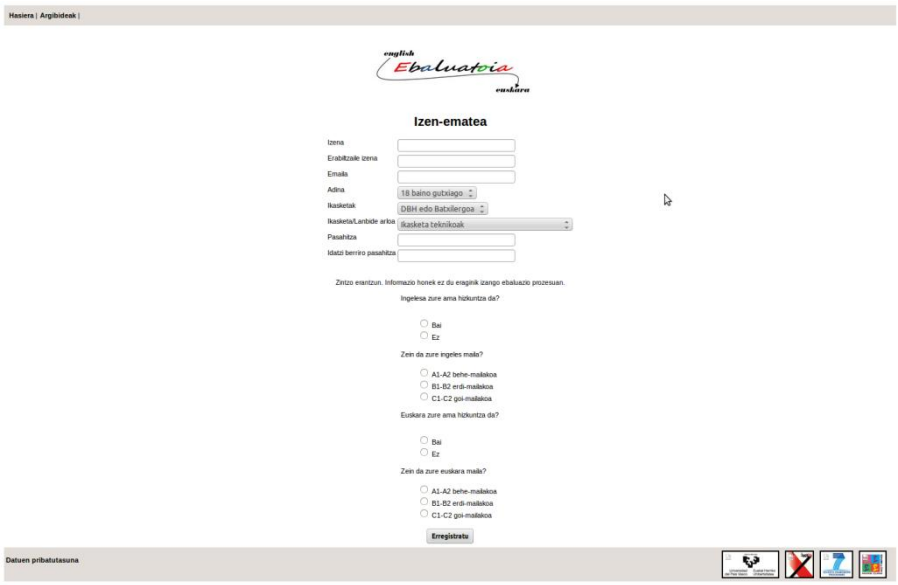


Figure 12: Screenshot of the Registration page.

After logging in, participants reach the Welcome page (see Figure 13). 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.

Figure 13: Screenshot of the Welcome page.

Participants are next taken to the page Instructions for participation (see Figure 14). 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.

Figure 14: Screenshot of the Instructions page.

Participants then click on “Show me the sentences” and access the Evaluation page (see Figure 15). 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 done 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 other translations seen previously.
- 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 to influence 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 “completed”. 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, she/he will not be allowed to continue collaborating. It is compulsory to successfully answer the first two control sentences. From there onwards, control sentence success has to be kept below 1/3 for the platform to keep the participant in. The recount for success is only be 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 is shown a message "We are sorry to tell you that 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 Ebaluatioa". The evaluations completed by the participants are erased and require a new participant to complete them.

Figure 15: Screenshot of the Evaluation page.

To continue evaluating, participants click on the "Next" button. This action reloads the page and shows a new evaluation unit.

Participants 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 (see Figure 16). This page summarizes the participant's contribution and reminds her/him that she/he can return to the site and keep contributing any time.

Figure 16: Screenshot of the Logout page.

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. Next, we study the results per test set to see if performance differences exist among the systems depending on subset. Finally, we provide an initial error analysis to identify frequent errors.

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

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

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 6 below 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. We see that the kappa scores for all the system pairs range between 0.49 and 0.53, within the moderate agreement range.

System pair	Kappa score
SMTb VS SMTs	0.52
SMTb VS Google	0.50
SMTb VS Matxin	0.52
SMTb VS Hybrid	0.50
SMTs VS Google	0.51
SMTs VS Matxin	0.51
SMTs VS Hybrid	0.53
Google VS Matxin	0.49
Google VS Hybrid	0.51
Matxin VS Hybrid	0.51

Table 6: Inter-annotator kappa scores for the comparison results per system-pair.

These scores are solid compared to the kappa scores obtained in the global MT evaluation campaigns. During the annual machine translation evaluation shared-tasks, researchers (and crowd participants in the latest edition) rank the output of five MT systems. Their kappa scores, as shown in Table 7, range between 0.168 and 0.494. The 5-output ranking method 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. For the WMT11, WMT12 and WMT14 campaigns, it was shared-task participants who performed the evaluations, i.e. experts, to a higher or lower extent. WMT13 collected judgements from both shared-task participants and non-experts hired through Amazon's Mechanical Turk, an online marketplace for work.[17] As expected, experts obtained higher kappa scores than Turkers. 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 7: 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.

Kappa scores provide an objective measure of the occasions in which participants agree on a specific question considering chance agreement. However, many reasons can make participants agree or disagree. Very high scores might mean that the task is easy because the quality of the systems is very different, but it might also be the case that participants

[17] https://www.mturk.com/mturk/welcome

all misunderstood the task and have evaluated something different from what you set out for. Low scores mean that the task was difficult. It might be inherently difficult because the systems perform similarly, the evaluation method might not be appropriate, or the instructions were not clear enough and as a result participants are interpreting them as they see fit. 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 allows us to pursue the analysis of results confidently.

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 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 (see Table 8). Because these are all valid answers, we will consider all evaluations when reporting the results.

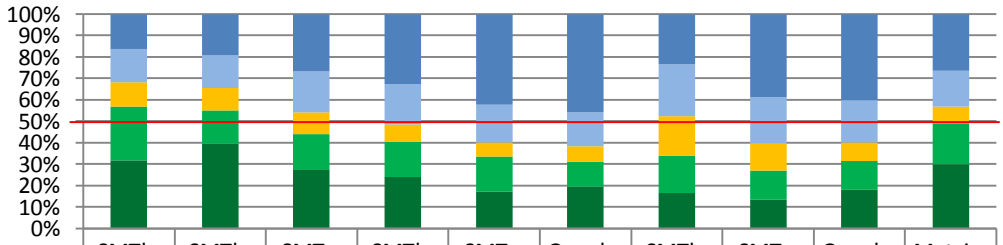
	SMTb-SMTs	SMTb-Google	SMTb-Matxin	SMTb-Hybrid	SMTs-Google	SMTs-Matxin	SMTs-Hybrid	Google-Matxin	Google-Hybrid	Matxin-Hybrid
Total evaluations	2635	2632	2660	2653	2600	2630	2623	2616	2618	2616

Table 8: 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 obtained by 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").

From the evaluations collected during Ebaluatoia (see Table 9), 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 significant, with only

8 additional sentences allocated to SMTs (229 sentences for SMTs and 221 for Google, 50 equal).[18]



	SMTb-SMTs	SMTb-Google	SMTs-Google	SMTb-Matxin	SMTs-Matxin	Google-Matxin	SMTb-Hybrid	SMTs-Hybrid	Google-Hybrid	Matxin-Hybrid
System 1++	80	95	133	162	210	229	117	192	200	132
System 1+	78	76	96	95	89	78	121	108	98	83
equal	57	53	50	40	32	36	92	64	44	38
System 2+	126	77	84	82	82	59	87	68	67	96
System 2++	159	199	137	121	87	98	83	68	91	151

Table 9: Number 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 (257 and 203 sentences, respectively, 40 equal) and SMTh (238 and 170 sentences, respectively, 92 equal). The proportion of translations rated as equal for the SMTb-SMTh 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 is never the preferred system of participants. This is not surprising, as Matxin, 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 times. 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.

SMTh is the preferred system only when paired against Matxin (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. We said that it is Matxin that guides the hybrid translation process. Because this is an early prototype with considerable coverage constrains, we can assume that the RBMT foundation of SMTh will

[18] The difference in sentences in all system-pairs is statistically significant at p>0.05 except for the SMTs-Google pair (p=0.59612) based on a Z-test. 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).

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's output, the final translation is enhanced with respect to the pure Matxin translation.

System performance is compared for several sets along the evaluation. In order to easily compare the overall results, we will add a summary box at the end of each section. For the Ebaluatoia results, the ranking of the systems can be summarised as follows, from better to worse:

SMTs ≈ Google > SMTb > SMTh > Matxin

4.3 Overall automatic scores

We have separately calculated system performance using automatic metrics (see Table Table 10). This is not directly comparable to the results of Ebaluatoia because the evaluation set is different. For automatic metrics to be calculated, a reference translation of the source sentences is necessary, as well as the machine translation output. The evaluation set used in Ebaluatoia was a purposely-built set for which we lack reference translations. The automatic scores were calculated using two evaluation sets of 1,500 sentences previously extracted from the SMT training corpus, one from the Elhuyar evaluation set and one from the Paco evaluation set. Remember that the Elhuyar subcorpus (85% of the training corpus) was used to train and optimize the statistical systems and that the Paco subcorpus (15% of the training corpus) was used to train the systems, but it was not used during optimization.

Elhuyar evaluation set				**Paco evaluation set**			
	BLEU	**NIST**	**TER**[19]		**BLEU**	**NIST**	**TER**
SMTb	36.75	7.69	50.63	SMTb	24.07	5.71	69.78
SMTs	36.01	7.69	50.23	SMTs	24.09	5.72	68.56
Matxin	04.06	3.21	86.70	Matxin	03.92	3.06	89.32
SMTh	27.21	6.81	60.14	SMTh	13.63	4.82	79.03
Google	14.93	4.96	71.33	Google	22.50	5.82	69.19

Table 10: Automatic scores for the MT systems under evaluation for the Elhuyar and Paco subcorpora.

According to automatic scores, SMTb is the best-scoring system, at par with SMTs in the Paco corpus. This is in disagreement with the human evaluation. Participants clearly preferred SMTs over SMTb. This discrepancy between automatic metrics and human evaluation with regards to the value of addressing morphological features for agglutinative languages in SMT should be taken into account, particularly during development. Automatic scores do not always reflect the contribution of segmentation. The decrease in BLEU points should not stop this strand of research which real users clearly state is worth the effort.

[19] Note that BLEU scores quality whereas TER reports errors. Therefore, the higher the BLEU score the better the translation is. For TER, the lower the error rate, the better the translation is.

In line with the results from Ebaluatoia, SMTh lags behind SMTb and SMTs. Scores predict a very large difference in quality between the systems, with a difference of over 10 BLEU points.

As expected, we observe the great impact the optimization set has on the scores of SMT systems. All the statistical systems evaluated score very differently in the two sets. Both SMTb and SMTs suffer a drop of about 12 points when evaluating them on the Paco set. The drop for SMTh is even harsher, with 14 points. Interestingly, Google obtains a great BLEU increase. Whereas the difference between Google and SMTb/SMTs is of 21 BLEU points in the Elhuyar set, the difference is dramatically reduced to 1.5 points when evaluated in the Paco set. The texts included in the Elhuyar subcorpus are IT documentation and academic textbooks that are proprietary and probably not available online, whereas the Paco subcorpus includes entertainment data crawled from the Web. Clearly, the first type of texts is more difficult for Google to obtain and their system is probably not tuned to work on this type of data. However, the data in the Paco set is available online and might even be part of the training data of Google's system. The results from Ebaluatoia put SMTs and Google at the same level, whereas automatic metrics still favour SMTs. This is most probably because even if not used for optimization, SMTs has been trained on the Paco subcorpus. The difference in BLUE score might disappear if we assess the system in a different out-of-domain evaluation set. Unfortunately, we currently do not have additional out-of-domain parallel data to test this.

Matxin scores lowest, by far. The difference in BLEU points as compared to 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. Automatic scores are calculated against a reference translation. They do not consider the correctness of the machine translation but rather compare the difference between the MT output and the reference translation. The further the MT output is from the reference, the lower the automatic score will be. This type of measurements tends to be very harsh on rule-based systems, which tend to output grammatically correct output that might lack fluency. Moreover, the SMT systems we have developed have been trained on similar corpus data, and therefore, are trained to output reference-style text.

Matxin has not been specifically trained to translate text on the Elhuyar and Paco subcorpora. BLEU scores for the RBMT system are similar across evaluation sets (4.06 and 3.92), meaning that the system is robust and deterministic. Matxin has proven to be a consistent system that can deal with a set of grammatical structures across domains.

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

Elhuyar evaluation set

SMTb > SMTs > SMTh > Google > Matxin

Paco evaluation set

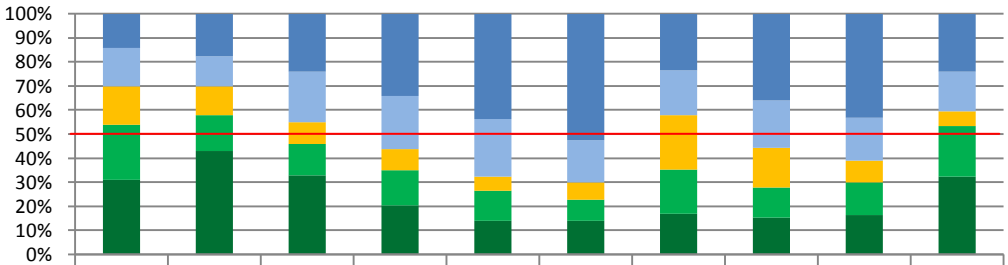
SMTs ≈ SMTb > Google > SMTh > Matxin

4.4 Analysis of results per test subset

In this section we analyse the participants' responses for each of the evaluation subsets listed in section 3.3. Because the subsets were collected from different sources, they may display different textual features, and this analysis might help us study whether systems perform uniformly across subsets or differences exist. Should a system be particularly successful in a subset, we could further study the linguistic characteristics of the set to try to specialise our systems.

Paco set

The results from the Paco set (200 sentences) follow the trend of the general results (see Table 11). We find two noticeable differences in Google's performance. We see that although the overall results remain the same, the number of sentences that score for Google undisputedly (System X++) has increased substantially (over 10%) when compared against SMTs. Similarly, Google increases its superiority against Matxin for the Paco set with an overall increase of almost 10%.



	SMTb-SMTs	SMTb-Google	SMTs-Google	SMTb-Matxin	SMTs-Matxin	Google-Matxin	SMTb-SMTh	SMTs-SMTh	Google-SMTh	Matxin-SMTh
System 1++	14	17,5	24	34	43,5	52,5	23,5	36	43	24
System 1+	16	12,5	21	22	24	17,5	18,5	19,5	18	16,5
equal	16	12	9	9	6	7	22,5	16,5	9	6
System 2+	23	15	13	14,5	12,5	9	18,5	12,5	13,5	21
System 2++	31	43	33	20,5	14	14	17	15,5	16,5	32,5

Table 11: Ebaluatoia results for the Paco evaluation set (%).

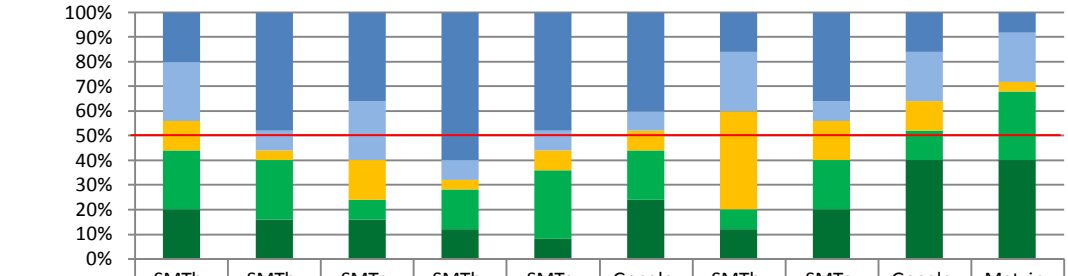
The overall system ranking for the Paco set is as follows:

Google ≈ SMTs > SMTb > SMTh > Matxin

Elhuyar set

The results from the Elhuyar set (25 sentences) show differing trends. Firstly, we see that SMTb has increased its scorings against all competitors. It performs similarly to SMTs, which has lost its advantage against SMTb. It has increased the difference against Matxin and Google, and although it has remained constant in its overall wins against SMTh, the number of sentences in the "equal" group has increased considerably. In fact, SMTh has

improved its scorings against all systems except SMTb. Matxin remains behind its competitors but we observe an important increase in the proportion of outputs preferred by the participants when paired against SMTs and Google. This is probably because our systems were optimized on this subset.



	SMTb-SMTs	SMTb-Google	SMTs-Google	SMTb-Matxin	SMTs-Matxin	Google-Matxin	SMTb-SMTh	SMTs-SMTh	Google-SMTh	Matxin-SMTh
System 1++	20	48	36	60	48	40	16	36	16	8
System 1+	24	8	24	8	8	8	24	8	20	20
equal	12	4	16	4	8	8	40	16	12	4
System 2+	24	24	8	16	28	20	8	20	12	28
System 2++	20	16	16	12	8	24	12	20	40	40

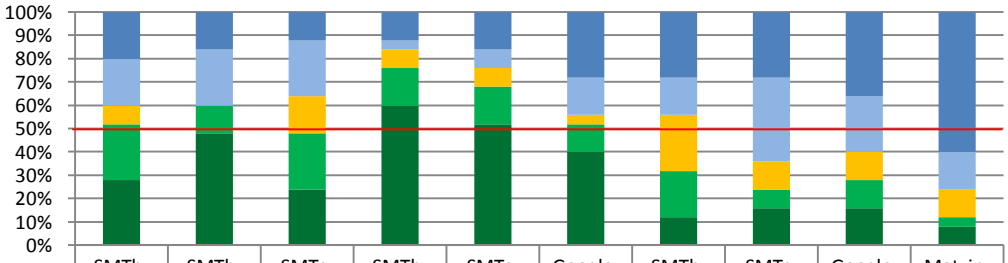
Table 12: Ebaluatoia results for the Elhuyar evaluation set (%).

The overall system ranking for the Elhuyar set is as follows:

SMTb ≈ SMTs > SMTh > Google > Matxin

Hello set

The Hello set (25 sentences) displays an interesting divergence from the overall Ebaluatoia results. Matxin performs particularly well for this set and surpasses all four competitors. The remaining pairs perform similarly to the overall results. Google reverts to the general proportions against the remaining three systems and maintains its superiority against SMTb, SMTs and SMth. SMTs recovers the advantage against SMTb and SMTh.



	SMTb-SMTs	SMTb-Google	SMTs-Google	SMTb-Matxin	SMTs-Matxin	Google-Matxin	SMTb-SMTh	SMTs-SMTh	Google-SMTh	Matxin-SMTh
System 1++	20	16	12	12	16	28	28	28	36	60
System 1+	20	24	24	4	8	16	16	36	24	16
equal	8	0	16	8	8	4	24	12	12	12
System 2+	24	12	24	16	16	12	20	8	12	4
System 2++	28	48	24	60	52	40	12	16	16	8

Table 13: Ebaluatoia results for the Hello evaluation set (%).

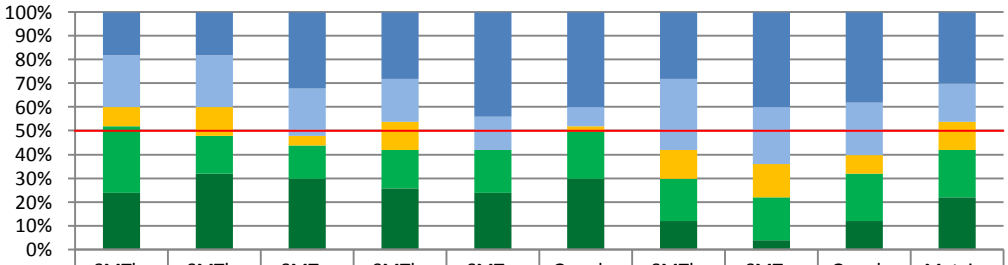
The overall system ranking for the Hello set is as follows:

Matxin > Google > SMTb > SMTs > SMTh

BBC1 set

In the BBC1 set (50 sentences), although Matxin is not the preferred system any more, its scores are considerably higher than in the overall results against all systems. The remaining scores are, once again very similar to the overall scores. The difference between SMTb and SMTs is slightly smaller but the latter still outperforms SMTb. And both SMTb and SMTs score better than Google. A detail to mention is that when paired against SMTs, although the overall numbers in favour of SMTh remain the same, the number of sentences for "System X+" is higher than in the overall results.

An inconsistency appears with the preference for SMTb, Matxin and Google. Participants prefer Google over SMTb and SMTb over Matxin, but then they seem to prefer Matxin over Google. We believe that this shows that the quality of the three systems is very similar.



	SMTb-SMTs	SMTb-Google	SMTs-Google	SMTb-Matxin	SMTs-Matxin	Google-Matxin	SMTb-SMTh	SMTs-SMTh	Google-SMTh	Matxin-SMTh
System 1++	18	18	32	28	44	40	28	40	38	30
System 1+	22	22	20	18	14	8	30	24	22	16
equal	8	12	4	12	0	2	12	14	8	12
System 2+	28	16	14	16	18	20	18	18	20	20
System 2++	24	32	30	26	24	30	12	4	12	22

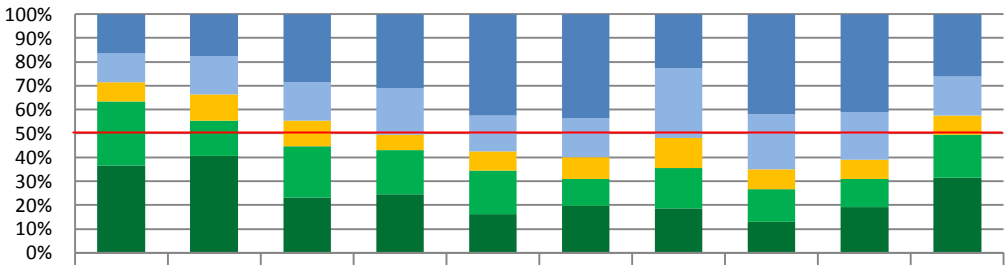
Table 14: Ebaluatoia results for the BBC1 evaluation set (%).

The overall system ranking for the BBC1 set is as follows:

SMTs > Google ≈ Matxin SMTb ≈ > SMTh

BBC2 set

The BBC2 set displays, once again, very similar results to those of the overall set. The only main divergence worth mentioning is the improvement of the SMTs system over the SMTb, by obtaining over 60% of the sentences.



	SMTb-SMTs	SMTb-Google	SMTs-Google	SMTb-Matxin	SMTs-Matxin	Google-Matxin	SMTb-SMTh	SMTs-SMTh	Google-SMTh	Matxin-SMTh
System 1++	16,5	17,5	28,5	31	42,5	43,5	22,5	42	41	26
System 1+	12	16	16	19,5	15	16,5	29,5	23	20	16,5
equal	8	11	11	6,5	8	9	12,5	8,5	8	8
System 2+	27	15	21,5	18,5	18,5	11,5	17	13,5	12	18
System 2++	36,5	40,5	23	24,5	16	19,5	18,5	13	19	31,5

Table 15: Ebaluatoia results for the BBC2 evaluation set (%).

The overall system ranking for the BBC2 test is as follows:

SMTs ≈ Google> SMTb > SMTh > Matxin

4.4.1 Summary of results per test subset

As expected, we see that the two evaluation subsets with the highest number of sentences set the trend for the overall results. Both the Paco and the BBC2 sets display very similar trends. Matxin performs slightly better in the BBC2 set. This might be because SMTb and SMTs are trained with the Paco corpus and therefore provide better quality output, whereas Matxin remains constant. However, we see that the same trend applies when paired against Google. We might therefore also conclude that the BBC2 set sentences are more suitable for Matxin than those in Paco. It might be the case that the BBC2 sentences are more carefully written and well-formed that web data from Paco.

It is also interesting to see that the difference in performance between the SMTb and SMTs is clear in all sets except in the Elhuyar set. Two reasons might be behind this result. Firstly, the Elhuyar training corpus is much larger than the Paco corpus – although it is also very diverse –, and therefore, SMTb had sufficient data to learn morphology-related information without the need for segmentation. Secondly, both statistical systems were trained and optimized on the Elhuyar corpus, and therefore, they are tuned to translate similar text. The SMTb then has more difficulty than SMTs to cope with dissimilar data.

What has emerged from this evaluation subset analysis is that Matxin outperforms the statistical systems in a number of specific contexts. A closer analysis of the Hello set in particular, could help us pinpoint the sentence type in which Matxin succeeds well over the statistical systems. We will briefly address this in the next section.

4.5 Structural analysis of subset source sentences

As a first attempt to do this, we have compared the dependency structures in the Hello subset (25 sentences) and the remaining evaluation set (475 sentences). We have analysed the source sentences with the Stanford parser (the same parser used by Matxin) and compared the proportion of dependency pairs. When describing Matxin in Section 3.1.3, we said that even if the analyser provides named dependencies for each element in the sentence, we then gather elements into larger chunks or phrases. We have performed dependency calculations based on this unit.

Due to the limited number of sentences in the Hello subset, high-level sentence structures, that is, the phrases that are depend directly on it, do not reveal meaningful information. The Hello subset has 21 different combinations out of 25 sentences, and the remaining subset has 208 combinations (see Table 16).

Hello evaluation subset				Remaining evaluation set		
Proportion of phrase structure	Number of phrase structure	Phrase structure[20]		Proportion of phrase structure	Number of phrase structure	Phrase structure
0.12	3	nsubj-ccomp		0.0842105	40	nsubj-xcomp
0.08	2	prep-nsubj-dobj		0.0778947	37	nsubj-dobj
0.08	2	nsubj-xcomp		0.0505263	24	nsubj-dobj-prep
0.04	1	nsubj-prep		0.0463158	22	nsubj-ccomp
0.04	1	nsubj-rcmod-cc-conj		0.0442105	21	nsubjpass-prep

Table 16: Top 5 high-level sentence structures in the Hello set and the remaining set.

With the aim of collecting a more considerable number of examples, we have extracted the dependency pairs of the main verb and its direct chunks. The analyser takes the main verb phrase as the central (root) element of a sentence to construct the dependency tree. Table 17 shows the phrase types that are dependent of the root for the Hello subset and the remaining set.

Overall, there are 71 direct dependencies from the root in the Hello set covering 14 types. We observe that the most frequent phrases in the Hello subset are nominal subjects (nsubj), prepositional phrases (prep), direct objects (dobj) and clausal complements (ccomp). It seems that Matxin can handle these structures better or at least at par with the other systems.

Moreover, if we compare the proportions of the dependency types across sets, we see that the proportion of prepositional phrases and clausal complements is higher in the Hello subset than in the remaining set, which suggests that these phrases might be better handled by Matxin.

If we consider the phrase types in the remaining set, we see 24 different types that account for 1134 direct dependencies from the root. We notice that there are a good number of phrase types that are not present in the Hello subset. We cannot claim that it is those that are particularly difficult for Matxin but clearly the system did not have to address them in the Hello corpus. Also, we see that the proportion of adverbial modifiers (advmod) open clausal complements (xcomp), that is, clause complements without their own subject, and passive subjects (nsubjpass) is higher in the remaining set.

[20] For the extended name of the dependency abbreviations see de Marneffe, C. and Manning, C. 2008. Stanford Dependencies manual at http://nlp.stanford.edu/software/dependencies_manual.pdf

Hello evaluation subset				Remaining evaluation set		
Proportion of root-phrase pairs	Number of root-phrase pairs	Phrase types		Proportion root-phrase pairs	Number of root-phrase pairs	Phrase types
0.28169014	20	nsubj		0.26761619	357	nsubj
0.25352113	**18**	**prep**		0.17166417	229	prep
0.09859155	7	dobj		0.11244378	150	dobj
0.07042254	**5**	**ccomp**		**0.08995502**	**120**	**advmod**
0.05633803	4	dep		**0.07871064**	**105**	**xcomp**
0.04225352	3	xcomp		**0.06146927**	**82**	**nsubjpass**
0.04225352	3	nsubjpass		0.04122939	55	ccomp
0.04225352	3	advmod		0.03523238	47	cc
0.02816901	2	conj		0.02698651	36	conj
0.02816901	2	cc		0.02698651	36	advcl
0.01408451	1	vmod		0.02623688	35	acomp
0.01408451	1	rcmod		0.02248876	30	dep
0.01408451	1	prt		0.00674663	9	vmod
0.01408451	1	acomp		0.005997	8	tmod
				0.005997	8	prt
				0.005997	8	expl
				0.00524738	7	csubj
				0.00224888	3	iobj
				0.00149925	2	rcmod
				0.00149925	2	parataxis
				0.00149925	2	discourse
				0.00074963	1	pobj
				0.00074963	1	cop
				0.00074963	1	appos
	71				1334	

Table 17: Summary of phrases that depend directly from the verb in the Hello evaluation subset and the remaining set.

Finally, we have extracted all dependency pairs in both subsets. We see that the proportions within the sets remain similar (Table 18). We see a slight increase in the proportion of adverbial modifiers (advmod) in the remaining set but most phrase types do not vary in more than about 2%.

Hello evaluation subset				Remaining evaluation set		
Proportion of phrases in Hello	Number of phrases in Hello	Phrase types in Hello		Proportion phrases in remaining corpus	Number of phrases in remaining corpus	Phrases in remaining corpus
0.25308642	41	prep		0.22636301	656	prep
0.19135802	31	nsubj		0.18115942	525	nsubj
0.08641975	14	dobj		0.10282954	298	dobj
0.06790123	11	cc		**0.09213251**	**267**	**advmod**
0.0617284	10	dep		0.05555556	161	xcomp
0.05555556	9	conj		0.05486542	159	cc
0.04320988	**7**	**advmod**		0.04175293	121	conj
0.03703704	6	xcomp		0.03657695	106	nsubjpass
0.03703704	6	ccomp		0.02864044	83	dep
0.0308642	5	vmod		0.02657005	77	ccomp
0.0308642	5	mark		0.0220842	64	acomp
0.01851852	3	tmod		0.01897861	55	vmod
0.01851852	3	nsubjpass		0.01863354	54	mark
0.01851852	3	acomp		0.01690821	49	rcmod
0.01234568	2	prt		0.01483782	43	advcl
0.00617284	1	rcmod		0.00793651	23	pcomp
0.00617284	1	pcomp		0.00793651	23	amod
0.00617284	1	csubj		0.00690131	20	prt
0.00617284	1	appos		0.00517598	15	expl
0.00617284	1	amod		0.00483092	14	det
0.00617284	1	advcl		0.00448585	13	tmod
				0.00276052	8	pobj
				0.00276052	8	appos
				0.00241546	7	predet
				0.00241546	7	csubj
				0.00207039	6	number
				0.00207039	6	npadvmod
				0.00172533	5	quantmod
				0.00138026	4	poss
				0.00138026	4	iobj
				0.0010352	3	neg
				0.0010352	3	mwe
				0.0010352	3	aux
				0.00069013	2	parataxis
				0.00069013	2	nn
				0.00069013	2	discourse
				0.00034507	1	preconj
				0.00034507	1	cop
	162				2898	

Table 18: Summary of all dependency pairs for chunks in the Hello evaluation subset and the remaining set.

Although we cannot draw conclusive results due to the limited number of sentences in the Hello subset, the analysis seems to suggest that Matxin handles the most common structures (sentences that combine subject, object and prepositional phrases) better than the remaining systems. When it comes to more complex structures (open clausal complements and adverbial modifiers) the statistical systems seem to outperform it. This is in line with Matxin's stage of development, which cover the most basic structures but more complex and rarer ones are still to be implemented. This, of course, does not necessarily mean that Matxin's output for the structures covered is perfect.

4.6 Error analysis

A qualitative analysis of the translations of each system will allow us to shed light on the type and frequency of the errors systems make. We will use this information to guide future system development.

Error counts allow us to identify the type of errors the systems make and quantify them, and this is key to guide further development and research. However, this is not directly proportional to system quality. Errors differ in severity. For example, the use of a locative genitive postposition instead of a possessive genitive will not impede the comprehensibility of a text as much as a noun phrase that has been split across the sentence. Also, a particular type of error might be more or less severe. An incorrect auxiliary in a sentence where the subject and objects are not made explicit will be a more serious error than where these are explicit. A study of the correlation between the errors and human preferences would allow us to assign severity levels to the different categories and to guide the focus of further research. This study falls out of the scope of this work and will be listed as future work.

As an initial attempt, we have selected 25 random source sentences (307 words) and have performed an error analysis (see the source sentences together with their translations and Ebaluatoia scores in ANNEX I). We have classified the errors found in the translations according to a general linguistic typology.

Lexis

The Lexis category includes incorrect lexical choices as well as incorrect translations of longer set phrases (see Table 19).

incorrect lexical choice	Miranda Kerr is the new face of H&M"s **SS** 14 campaign.	
	Miranda Kerr da **ZERBITZU SEKRETU** kanpaina 14 H&M aurpegi berria.	Matxin
incorrect phrase translation	The way we play **as children** informs the skills we develop.	
	Bide **haurrak bezala** jokatuko dugu jakinarazten gaitasunak garatzen ditugu.	Google

Table 19: Examples of errors in the Lexis category.

Morphosyntax

The Morphosyntax category includes morphological and syntactic errors (see Table 20). We have fused both categories into one as, due to the nature of Basque, these types of errors are often so intertwined that it is difficult to opt for one category over the other. Moreover, this classification is proposed as a tool to easily summarise and assimilate system error information but the exact classification of the items will not have any impact on future research decisions as errors are addressed based on their fixing requirements rather than on their linguistic nature. This category includes issues with postpositions[21] and subordinate markers, as well as issues with various structures such as superlative constructions or coordinate constructions. We also include a subcategory for determiners, which include both suffixes and free-standing elements. We include in this category cases of missing or additional grammatical categories, as well as errors related to elements like coordinators, question words and particles, and negative particles. Finally, a specific subcategory has been added to cover part-of-speech (POS) errors as this ambiguity problem is very frequent in MT.

incorrect postposition	At least this way, they have been able to see the child smiling from time to time.	
	Gutxienez, horrela, ikusi ahal izan duten **haurrak** irribarrez noizean behin.[22]	SMTb
extra subordinate marker	The death penalty constitutes a symptom of a culture of violence, not a solution to it	
	Heriotza-zigorra sintoma bat da indarkeriaren kultura bat ez **ba**da, konponbidea	SMTh
incorrect construction of coordination	You [can always consult your correspondence at Clavenet] and [can receive postal deliveries again whenever you like].	
	Beti kontsultatu ahal izango duzu zure korrespondentzia clavenet berriro jaso ditzake posta-entregak **eta** edozein unetan izanen duzua.	SMTs
determiner error	Nektarios Basdekis is a computer expert and a photographer.	
	Nektarios Basdekis da ordenagailu aditu bat eta photographer bat.	Matxin
extra question particle	What message does that send out?	
	Zer mezu bidaltzen duten egiten **al** du?	SMTh
missing noun	Second **lieutenant** Julio Romero Marcheut, with bullet and bayonet wounds, defends himself against the Carlists.	
	Bigarrena: julio romero marcheut, buleta duten eta bayonet zauriak, defendatu zuen karlisten aurka.	SMTb
POS error	Facebook does not **hand** over full access to a person's account due to privacy concerns	
	Facebook ez du **esku** sarbide osoa pertsona baten kontuak direla - eta pribatutasun-arazoak.	SMTs

Table 20: Examples of errors in the Morphosyntax category.

[21] Postpositions include both grammatical case-markers (suffixes used to identify the subject, direct object and indirect object) and adverbial phrase case-markers.

[22] The plural suffix and the ergative suffix in Basque is the same, -k, and therefore the form of a singular definite noun in the ergative case and the form of a plural definite noun in the absolutive case – which does not have any suffix – are indistinguishable. This property is called syncretism. Because the MT systems' output information is limited to forms, it is not possible to establish which the intended form it is. Given that the systems very rarely have problems with plurality, when faced with an ergative-absolutive error, we have considered the form to be marking case rather than plurality.

Verbs

A separate category has been defined for verb phrases because they differ significantly in English and Basque, complexity being higher for Basque, which makes them a frequent source of errors. English verb phrases consist of a lexical verb, which can stand alone or be preceded by one or more auxiliary verbs which mark meanings associated with aspect, voice or modality. English verbs show distinctions of tense and can include modal auxiliaries. In turn, most Basque verb phrases consist of a participial verb and a conjugated auxiliary. The former carries aspectual and, in part, tense and voice information, and the latter conveys information about argument structure, tense and modality. The variability of conjugated auxiliaries poses great difficulty for statistical systems to learn correct equivalences. We have divided this category into subgroups that represent the different types of errors that can appear in verb phrases, that is, aspect, tense, modality or paradigm, subject person in auxiliary or object number in auxiliary (see Table 21). We have also recorded missing or additional auxiliaries and lexical verbs, as well as complete verb phrases. Note that more than one error can be present in a single verb phrase.

missing verb phrase	Second lieutenant Julio Romero Marcheut, with bullet and bayonet wounds, **defends** himself against the Carlists.	
	Bigarren tenientea julio romero marcheut, bala-zauriak dituzten eta bere buruaren aurka, karlista.	SMTs
missing participial verb	The dresses **were adorned** with thousands of sequins and crystals.	
	Soinekoak; **ziren**, eta milaka sequins eta kristalak.	SMTb
missing auxiliary	Prostitution **crosses** that line for you.	
	Prostituzioa zuretzat lerro hori **zeharkatzen.**	Google
incorrect aspect	So how many people should you date before you **decide** to settle down?	
	Beraz, zenbat pertsona behar eguneratuta duzu behera kitatzeko **erabakitzen duzu** aurretik?	Google
incorrect tense	It was made using only handtools and **required** approximately 360 hours work.	
	Handtools bakarrik erabiliz egin zen eta 360 ordu inguru **behar de**n lana.	SMTb
extra modal auxiliary	Your innate love of animals **brought** you to chimpanzees.	
	Zure animaliak maitasuna berezkoa **ekarri** txinpantzeen **behar duzu.**	Google
incorrect paradigm	**It was made** using only handtools and required approximately 360 hours work.	
	Hari egin zitzaion baina handtoolak erabili eta behar izanda gutxi gorabehera ordu 360 lana.	Matxin
incorrect subject in auxiliary	You can always consult your correspondence at Clavenet and **can receive** postal deliveries again whenever you like.	
	Zuk beti zure korrespondentzia kontsulta dezakezu Clavenetekin eta posta banaketak **har ditzake** berriro zuk gogoko duzunean.	Matxin
incorrect object in auxiliary	At least this way, they have been able to see the child smiling from time to time.	
	Gutxienez horrela ikusi ahal izan dituzte haurrak une batetik bestera, irribarrez.	SMTh

Table 21: Examples of errors in the Verb category.

Order

Order also has a dedicated category due to the impact it has on the overall comprehensibility of the translations and because it is a property that can be addressed specifically in MT training for both rule-based and statistical systems (see Table 22). We have distinguished between higher-order ordering issues and phrase-internal errors.[23] Additionally, specific subcategories were created for the most recurring order issues such as head and relative clause positions, noun phrases and verb phrases. The latter include subgroups for incorrect internal reorderings and cases where elements belonging to a single phrase have been split into non-consecutive phrases.

incorrect sentence-level order	His work has given one of the most powerful of all impulses to the progress of science.	
	Eman du bere obra garrantzitsuenetako bat, bulkada guztien zientziaren aurrerapena.	SMTb
incorrect head-relative clause order	The way we play as children informs **the skills** we develop.	
	Haurrek, era batera edo bestera jokatuei jakinarazten die **gaitasunak** garatuko ditugu.	SMTs
internal order of noun phrase	**The final eight books** span poetry, novels and short stories.	
	Azken liburu zortzik poesia nobelak eta baxu istorioak zeharkatzen dituzte.	Matxin
split noun phrase	**Your innate love of animals** brought you to chimpanzees.	
	Zure sortzetiko maitasuna ekartzen baduzu txinpantzeak **animalia.**	
		SMTh

Table 22: Examples of errors in the Order category.

Punctuation

The category Punctuation includes both punctuation and orthography issues (see Table 23). These include incorrect uses of punctuation marks, capitalization errors and orthotactic constrains (orthographical rules governing the gluing of lemmas and affixes).

incorrect position of comma	The inaugural shortlist of the latest literary award on the block**,** the Folio Prize**,** has been unveiled.	
	Inaugurazio azken literatur blokea**,** Folio Saria da saria laburrean izan**,** ha inauguratu dira.	Google
capitalization error	If you ask it to, **Vini** will reject any attempt at payment made using this card.	
	Eskatu nahi baduzu, **vinik** ahalegin guztiak baztertzen ditu, horren bidez egindako ordainketa txartela.	SMTb
orthotactic error	The introduction of communication technologies and **Internet** in direct marketing supports this idea.	
	Sartzea, komunikazioaren teknologiak eta **internetko** zuzeneko marketina onartzen du ideia hori.	SMTs

Table 23: Examples of errors in the Punctuation category.

[23] Basque is a relatively free-order language with respect to high-order constituents and therefore, almost (if not all) combinations are correct. However, in some cases a particular ordering might sound odd because of focality reasons. The sentence-level ordering errors presented here might be disputable as the translation sentences are out of context and therefore it might be the case that the translation orderings are acceptable in the specific contexts they belong to.

Untranslated

Finally, we have added a new category called Untranslated for the source words that have been left in the original language, in English, rather than translating them (see Table 24).

untranslated	The dresses were adorned with **thousands** of **sequins** and crystals.	
	Soinekoak apainduak ziren beira **sequin**etako eta **thousand**ez.	Matxin

Table 24: Examples of errors in the Untranslated category.

Apart from the error typology described above, following custom SMT evaluation, we have classified each error as incorrect, missing or extra, to have a more comprehensive understanding of the systems' behaviour.

The difficulty of error analysis in MT varies significantly depending on the output quality. High-quality output with few mistakes renders the task simple and effortless. The errors -at least the most glaring ones- will be easy to spot and classify. However, the lowest the quality of the output, the more errors will co-occur and combine in the text. Identifying and classifying errors becomes a complex task. The way in which we have addressed the task is to try to record the lexical, morphological and syntactic changes, including reorderings that would be needed to transform the MT output into a correct translation (see Example 6). For the most complex cases, this strategy does not necessarily result in completely fluent and adequate translations even if we fix the errors recorded, but it should provide well-formed sentences. Also, it should be noted that because many ways to transform the MT output may exist, different modifications of the MT output may be possible, and as a result, different evaluators might classify errors differently. Still, the evaluator tends to minimise the number of errors (changes) and somehow follows the HTER model (see Section 2.2.2).

Miranda Kerr is the new face of H&M's SS 14 campaign.
Miranda **k**err aurpegi berria d**ela h**&**m's ss** 14 kanpaina**n**.

kapainan – incorrect inesive postposition, should be locative genitive
dela – extra subordinate marker
aurpegi berria [...]
...h&m's ss 14 kanpainan – noun phrase construction – split
kerr – capitalization error
h&m – capitalization error
's – untranslated genitive marker
ss – untranslated noun

Example 6: Example of error analysis.

SMTb

We classified 155 errors in the SMTb translations (see Table 25). The Morphosyntax category includes the highest number with 71 errors, out of which 33 belong to the postposition subgroup. 10 postpositions have been judged to be incorrect and up to 22 are missing, which shows the difficulty of the system to learn equivalences between English prepositions and Basque postpositional suffixes (including suffixes for grammatical cases). Subordinate markers also show a considerable number of errors, 11 in total. Again, English subordinate pronouns are isolated words whereas Basque attaches suffixes to the auxiliary or conjugated verbs. This adds to the inherent complexity of Basque verbs and does not help the system learn the equivalences. POS errors are also frequent with 11 cases reported for this set of 25 sentences.

Linguistic category	Total errors	Error type	Incorrect	Missing	Extra	Total
Lexis	8	Lexical choice	8			8
Morphosyntax	71	Postpositions	10	22	1	33
		Determiners	3	1	4	8
		Subordinate marker	3	4	4	11
		Coordinate construction	1			1
		Question word			1	1
		Coordinator			2	2
		Preposition		1		1
		Pronoun		1		1
		Adjective		1		1
		Noun		1		1
		POS – ambiguous source	11			11
Verb	28	Participial form		2		2
		Aspect in participial form	6			6
		Auxiliary		1	2	3
		Tense in auxiliary	3			3
		Modal word or marker in auxiliary	2		1	3
		Paradigm of auxiliary	5			5
		Subject person in auxiliary	4			4
		Object number in auxiliary	2			2
Order	25	Sentence-level	8			8
		Head-relative clause	2			2
		Noun-complement	1			1
		Noun phrase composition - internal	7			7
		Noun phrase composition - split	7			7
Punctuation	16	Capitalization	9			9
		Comma			4	4
		Semicolon			1	1
		Colon			2	2
Untranslated	7					7
	155		**92**	**34**	**22**	**155**

Table 25: Error classification for SMTb.

Order issues also amount to 25 errors. Noun phrase errors account for 14 issues, quite a considerable number given the limited set of sentences.

Punctuation includes quite a high number of errors, 16, but we see that most are capitalization issues, which are not particularly difficult to fix, although they do confuse the reader, who can otherwise use capital letters to identify elements of the sentences while reading.

The Lexis and Untranslated categories include 8 and 7 errors, respectively. Errors in lexical choice are usually the result of polysemy. Therefore, their impact in comprehension will depend on the extent to which the translation distances from the intended sense. The impact of the untranslated words will depend on the level of source language knowledge of the reader, who might be able to understand it or not.

SMTs

For SMTs, 144 errors have been classified (see Table 26). Overall, the proportion of errors for the different categories remains very similar to SMTb. The subgroups with the highest number of errors are again postpositions, POS and capitalization. We see that even if this system was specifically trained to better learn postposition and marker equivalences, the number of postposition errors has increased in 4. There are 3 more incorrect postpositions recorded but 3 fewer missing ones. Interestingly, the system output 5 extra postpositions. We do see an improvement over subordinate markers from SMTb, which recorded 11 errors and SMTs displays 5 incorrect uses, none missing or extra. SMTs also shows 5 errors in coordinate constructions, which were not present in SMTb.

The main difference between the SMTs over SMTb in the Verb category is the errors found for verb phrases, which were not present for SMTb. 3 additional phrases were output, 2 were missing, and 1 was incorrect.

Finally, we see a slight improvement in the construction of noun phrases, with 5 errors below SMTb.

Linguistic category	Total errors	Error type	Incorrect	Missing	Extra	Total
Lexis	8	Lexical choice	6			6
		Phrase translation	2			2
Morphosyntax	72	Postpositions	13	19	5	37
		Determiners		4	5	9
		Number in noun	1			1
		Subordinate marker	1	1	4	6
		Coordinate construction	5			5
		Noun		1		1
		Negative particle			1	1
		POS	12			12
Verb	25	Verb phrase	1	2	3	6
		Aspect in participial form	5			5
		Auxiliary			3	3
		Tense in auxiliary	2			2
		Modal word or marker in auxiliary	1		1	1
		Paradigm of auxiliary	6			6
		Subject person in auxiliary	2			2
Order	21	Sentence-level	7			7
		Phrase-internal	3			3
		Head-relative clause	2			2
		Noun phrase composition - internal	5			5
		Noun phrase composition - split	4			4
Punctuation	14	Capitalization	10			10
		Comma	2		1	3
		Hyphen			1	1
Untranslated	4					4
	144		**90**	**27**	**24**	**144**

Table 26: Error classification for SMTs.

SMTh

A total of 132 errors have been recorded for SMTh (see Table 27). Overall the proportions remain constant but we see a drop in the Morphosyntax and Order categories. Even if it is still the subgroup with the highest number of errors, postposition errors have lowered from 33 and 37 for SMTb and SMTs to 24, whereas subordinate markers stay at 5. Determiner-related errors have also been reduced from 8 and 9 to 3.

Order-related errors have decreased from 25 and 21 for SMTb and STMs to 14. Sentence-level errors have been reduced to 2 and head and relative clause position errors have disappeared. Noun phrase composition errors remain at 8.

Linguistic category	Total errors	Error type	Incorrect	Missing	Extra	Total
Lexis	10	Lexical choice	8			8
		Phrase translation	2			2
Morphosyntax	61	Postpositions	9	15		24
		Determiners	1		2	3
		Number in noun	3			3
		Subordinate marker	1		4	5
		Coordinate construction	4			4
		Superlative construction	1			1
		Question word or particle			1	1
		Coordinator			1	1
		Pronoun (demonstrative)			2	2
		Adjective		2	1	3
		Noun		1		1
		Adverb		1	2	3
		Negative particle			1	1
		POS	9			9
Verb	23	Verb phrase		1	2	3
		Aspect in participial form	3			3
		Auxiliary		2	3	5
		Tense in auxiliary	4			4
		Modal word or marker in auxiliary		1		1
		Paradigm of auxiliary	3			3
		Subject person in auxiliary	2			2
		Object number in auxiliary	2			2
Order	14	Sentence-level	2			2
		Phrase-internal	4			4
		Noun phrase composition - internal	3			3
		Noun phrase composition - split	5			5
Punctuation	17	Capitalization	9			9
		Comma	1	2	3	6
		Colon			1	1
		Full stop			1	1
Untranslated	7		7			7
	132		**83**	**24**	**25**	**132**

Table 27: Error classification for SMTh.

Matxin

A total of 112 errors have been classified for Matxin, 27.74% less than SMTb (see Table 28). Error proportions across categories have changed considerably. Error in the Lexis category have increased to 22 from 8-14 in the statistical systems.

The Morphosyntax category still shows a high number of errors in the postpositions subgroup but the errors in general are spread across the subgroups, with a noticeable drop in POS errors.

Linguistic category	Total errors	Error type	Incorrect	Missing	Extra	Total
Lexis	22	Lexical choice	19			19
		Phrase translation	3			3
Morphosyntax	54	Postpositions	17	11	2	30
		Determiners	3		1	4
		Number in adjective	1			1
		Subordinate marker		3		3
		Coordinate construction	3			3
		Superlative construction	1			1
		Preposition		1	2	3
		Pronoun			1	1
		Adjective			1	1
		Negation	1			1
		POS – ambiguous source	4			4
		POS – unambiguous	2			2
Verb	11	Aspect in participial form	1			1
		Auxiliary	1		1	2
		Paradigm of auxiliary	5			5
		Subject person in auxiliary	1			1
		Object number in auxiliary	2			2
Order	15	Sentence-level	1			1
		Head-relative clause	1			1
		Noun phrase composition - internal	5			5
		Noun phrase composition - split	6			6
		Verb chain composition	2			2
Punctuation	3	Capitalization	1			1
		Comma		1		1
		Orthotactics	1			1
Untranslated	7					7
	112		**81**	**16**	**8**	**112**

Table 28: Error classification for Matxin.

We see a decrease in the number of Verb category errors, a clear result of the MT approach which has well-establish equivalence rules. As expected, the most frequent error within this group is the choice of paradigm, a difficult disambiguation task given the ergative-absolutive nature of Basque, that is, syncretism occurs between plural absolutive and singular ergative and so the subject of an intransitive verb carries the same marker as the direct object of a transitive one.

Errors in the Order category remain similar to SMTh's, considerably lower than SMTb and SMTs. The errors in the Punctuation category also decrease. This is mainly due to a reduced number of capitalization errors. The number of errors in the Untranslated category has not been reduced.

It is worth noting that Matxin has a considerably lower number of extra elements, as low as 8 compared to the 24-25 of the other systems, and also a lower number of missing elements, 16, compared to the 24-30 of the statistical systems.

Google

We see that the Verb class has quite a considerable number of errors, 28, even if it addresses a limited set of cases. Most errors fall in the incorrect category with only two participial forms and one auxiliary missing. We see that errors are quite spread out but aspect is the subgroup with more errors (6).

Google's error proportions across categories are closer to those of the other statistical systems although differences arise (see Table 29). Lexical choice errors, within the Lexis category, lie somewhere between the SMT systems, who perform better, and Matxin, with a relatively high number of errors. As is the trend across systems, postpositions, in the Morphosyntax category, are the most prominent with 32 errors. Interestingly, Google shows no errors in coordinate constructions.

Linguistic category	Total errors	Error type	Incorrect	Missing	Extra	Total
Lexis	14	Lexical choice	12			12
		Phrase translation	2			2
Morphosyntax	58	Postpositions	18	12	2	32
		Determiners		2	6	8
		Number in noun	2			2
		Subordinate marker		2	3	5
		Sperlative construction	1			1
		Adverb		1	4	5
		POS – ambiguous source	2			2
		POS - unambiguous	3			3
Verb	33	Full verb	1			1
		Aspect in participial form	5			5
		Auxiliary	1	10	6	17
		Time in auxiliary	3			3
		Modal word or marker in auxiliary	1	2		3
		Paradigm of auxiliary	2			2
		Object number in auxiliary	2			2
Order	13	Sentence-level	2			2
		Phrase internal	4			4
		Noun phrase composition - internal	4			4
		Noun phrase composition - split	1			1
		Verb chain composition	2			2
Punctuation	7	Capitalization	2			2
		Comma	1	1		2
		Hyphen			1	1
		White space			2	2
Untranslated	7					7
	132		**71**	**30**	**24**	**132**

Table 29: Error classification for Google.

The number of errors recorded in the Verb category for Google is the highest across systems. In particular, Google seems to have difficulty with auxiliaries, where 10 are reported to be missing, 6 extra and 1 incorrect. On the other hand, it has the lowest number of issues classified in Order. Noun phrase composition only has 5 errors compared to 8-11 in the other systems, no relative clause and head positioning errors are present and sentence-level issues are only 2. Phrase internal ordering has appeared in 4 occasions.

Punctuation errors are low, with only two capitalization issues and untranslated words are 7, similar to the remaining systems.

4.6.1 Summary of error analysis

Overall, we see that the number of errors recorded for the 25 sentences, 112-155 across systems, is considerably high, with an average of 4.48-6.2 errors per sentence. The most frequent errors are those related to postpositions and verbs, two categories that show high complexity in Basque and differ greatly from the nature of their English counterparts.

An interesting finding of the analysis has been the behaviour of SMTs compared to SMTb. SMTs is an enhanced system trained to address the difficulty of learning suffixes. Surprisingly, the error analysis does not show an improvement over the translation of postpositions and only a slight improvement in subordinate markers has been observed. All in all, however, Ebaluatoia participants clearly prefer SMTs over SMTb. It seems to be the case that even if the translation of suffixes in particular is not improved, segmentation might help the aligner learn equivalences in general better and, as a result, the overall translation is better. A deeper study of the postposition errors might also show that the errors themselves are less serious.

The comparison of the errors recorded for SMTh and the errors of the systems it combines hints at the type of knowledge SMTh exploits from each of them. We see that the number of order errors is similar to Matxin's. Therefore, we learn that the hybrid system is indeed benefiting from a RMBT-guided structure. We see that lexical choice errors remain close to those of the SMT systems, and therefore we argue that SMT phrase candidates seem to improve Matxin's lexical choice. However, we see that Matxin's postposition and verb phrase choices, better than those of the SMT systems, could be further exploited to improve translation.

5 Conclusions and future work

This work set to compare the translation quality of four MT systems developed during the ENEUS project and Google Translate. To do so, we ran a large-scale crowd-based human evaluation campaign called Ebaluatoia February 14-25, 2014, which collected the opinions of regular users. The results from the campaign were then analysed to guide further research. We carried out several initial qualitative analyses to help identify in which direction we should improve the quality of the Basque to English translations. In particular, we analysed the Ebaluatoia results per evaluation subset, we performed a basic structural analysis to account for differences in system performance across evaluation subsets, as well as an error analysis to identify and quantify the errors made by each MT system.

The Ebaluatoia campaign achieved the set goals. We were surprised at the phenomenal response from the community, which exceeded all 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. The unprecedented participation of the Basque community is on its own an outcome of the work. Society has shown that they are interested and respond positively to research initiatives by voluntarily engaging in research-related activities and supporting the work conducted.

From the Ebaluatoia results, we have completed the ranking of the systems under evaluation. According to participant's preferences, Google Translate and the SMT system that uses segmentation are of similar quality. The third preferred system is the SMT baseline, followed by the hybrid system, with Matxin scoring the lowest. Still, Matxin wins in 31-43% of the sentences, showing that it can contribute to better translation quality.

When compared against the common string-based automatic metrics such as BLEU, NIST and TER, we saw that the ranking proposed by automatic metrics and the human evaluation differed significantly. The automatic metrics ranked SMTb and STMs as best systems, almost at par. Google and SMTh were ranked in the 3rd and 4th positions, with SMTh outperforming Google on the Elhuyar subset and Google outperforming SMTh on the Paco subset. Matxin lagged behind with surprisingly low scores. Overall these results contribute to the body of research that cautions against the use of automatic metrics as replacement for human evaluations.

The several analyses carried out as part of this work have allowed us to shed some light into the four specific questions we set to investigate in Section 3.1.6. Mainly through the analysis of results per evaluation subset, we have seen that the MT systems perform similarly across sets. An interesting exception to this was the case of Matxin in the Hello subset. Usually the least preferred of the systems, Matxin outperformed all the other systems on this subset. We carried out an initial comparison of the dependency structures present in the Hello subset and the remaining set, and identified the structures in which Matxin performed better as well as structures that it did not have to address in the Hello subset.

With regards to the difference between the SMT baseline and the SMT with segmentation, we have learnt that human evaluators clearly opt for the latter, which shows that the effort put into addressing morphology in SMT is noticed and welcome by users. A higher number of winning sentences has been allocated to SMTs in all evaluation subsets except for one, where both scored the same. Automatic metrics, in contrast, do not reflect this. The error analysis of SMTs has not shown considerable improvement in the translation of segmentation-related linguistic features with respect to SMTb. This suggests that segmentation does not specifically correct postposition and marker translation, for example, but rather it has an impact on the overall alignment quality, improving quality in general.

Hybrid systems are built with the aim to exploit the advantages of the different systems it combines. SMTh combines two statistical systems SMTb and SMTs, and the rule-based Matxin. It is designed to start from Matxin's dependency structure and select the most appropriate translation fragments enriched with the statistical systems' options. The error analysis has shown that the ordering errors made by the SMTh are fewer than those of the statistical systems, which proves that it is exploiting Matxin's structure. Also, it is clear that the system is rejecting some of Matxin's lexical selections in favour of the candidates of the statistical systems. However, we see that Matxin's knowledge regarding morphosyntax, and in particular, postpositions and verbs could be further exploited to enhance the system.

We included Google Translate in the evaluation campaign to check how the research prototypes performed compared to it. Considered a strong contender, Google has been the winner of Ebaluatoia, but at par with SMTs. This shows, therefore, that a morphologically informed statistical system can reach the same quality as a statistical system trained with supposedly much larger parallel corpora and which might be informed by other pivot languages.

Given the Ebaluatoia results, we can now guide future research and development for each of the systems. We concluded that the baseline SMTb should be abandoned in favour of the morphologically informed SMTs system. Given the progress done thanks to morphology-related information, we aim to find additional ways to introduce this type of information within SMTs. Also, guided by the error analysis of SMTs, we aim to build a separate rule-based post-processing module in the style of DepFix (Rosa et al., 2012) to directly address the most frequent errors.

We will continue developing Matxin to increase its structural coverage. In particular, we saw the need to improve postpositional selection. Moreover, we intend to address lexical disambiguation, one of the categories in which Matxin lagged behind the statistical systems.

SMTh has shown potential for improvement and, thanks to the error analysis, we can now pinpoint some of the specific features in which we can work on, namely, morphosyntax and verb-related features. We aim to test different selection methods to help the decoder exploit Matxin's knowledge further. Also, new hybridization attempts in the form of system selection will also be worked on after further analysis is done in structural analysis.

6 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.

Albrecht, J. and Hwa, R. 2007. A re-examination of machine learning approaches for sentence-level MT evaluation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007),* Prague, Czech Republic.

Arnold, D. 2003. Why translation is difficult for computers. In: H. Somers (ed.) *Computers and Translation: A Translator's Guide*. Amsterdam: John Benjamins Publishing Company. pp.119-142.

Banerjee, S. and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization* at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan. pp.65-72.

Béchar, H., Rubino, R., He, Y., Ma, Y. and van Genabith, J. 2012. An evaluation of statistical post-editing systems applied to RBMT and SMT systems. In *Proceedings of COLING'12*, Mumbai, pages 215–230.

Bojar, O., Buck, C. Federmann, C., Haddow, B., Koehn, P., Macháček, 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. 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*, pages 1–44, Sofia, Bulgaria.pages 1–44, Sofia, Bulgaria.

Brown, P., Cocke, J., Della Pietra, S., Jelinek, F., Della Pietra, V., Lafferty, J., Mercer, R. & Rossin, P. 1990. A statistical approach to machine translation. *Computational Linguistics* 16, pp.79-85.

Brown, P., Della Pietra, S. Pietra, V. and Mercer, R. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2): 263-311.

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*, pages 10–51, Montréal, Canada, June. Association for Computational Linguistics.

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

Callison-Burch, C., Osborne, M. and Koehn, P. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy. pp.249-256.

Carroll, J. 2004. Parsing. In: (ed.) Ruslan Mitkov: The Oxford Handbook of Computational Linguistics. Oxford and New York: Oxford University Press. pp. 233-248.

Chen, Y., Eisele, A., Federmann, C., Hasler, E., Jellinghaus, M. and Theison, S. 2007. Multi-engine machine translation with an open-source SMT decoder. In *Proceedings of WMT07*, pages 193–196, Prague, Czech Republic, June. Association for Computational Linguistics.

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

Collins, M. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, California, pp. 184-191.

de Marneffe, M.C., MacCartney, B. and Manning, C.D. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2006)*.

Dugast, L., Senellart, J. and Koehn, P. 2007. Statistical post-editing on SYSTRAN's rule-based translation system. In *Proceedings of the 2nd Workshop on Statistical Machine Translation (WSMT 2007),* Prague, Czech Republic. pp.220-223.

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.

Font Llitjós, A., Carbonell, J. and Lavie, A. 2005. A framework for interactive and automatic refinement of transfer-based machine translation. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT 2005),* Budapest, Hungary. pp.87-96.

Giménez, J. and Marquèz, L. 2007. Linguistic features for automatic evaluation of heterogenous MT systems. In *Proceedings of the ACL 2007 2nd Workshop on Statistical Machine Translation*, Prague, Czech Republic. pp.159-166.

Hildebrand, A. S. and Vogel, S. 2008. Combination of machine translation systems via hypothesis selection from combined n-best lists. In *Proceedings of Association for Machine Translation in the Americas (AMTA2008)*.

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.

Klein, D. eta Manning, C.D. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*.

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. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Demonstration session.

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.

Krings, H. P. 2001. Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes. Kent, Ohio: The Kent State University Press. Edited/translated by G.S. Koby.

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., Diaz de Ilarraza, A., España-Bonet, C., Sarasola, K. and Màrquez, L. 2011. Deep evaluation of hybrid architectures: simple metrics correlated with human judgments. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT)*, Universitat Politècnica de Catalunya, Barcelona, pages 50-57.

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

Leusch, G., Ueffing, N. and Ney, H. 2003. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of the 9th Machine Translation Summit of the International Association for Machine Translation (MT Summit IX),* New Orleans. pp.240-247.

Lin, C.Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004),* Barcelona, Spain.

Linguistic Data Consortium. 2003. Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translation. Project LDC2003T17.

Liu, D. and Gildea, C. 2005. Syntactic features for evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, (ACL 2005),* Ann Arbor, MI. pp.25-32.

Manning, C. and Schütze, H. 2000. Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts: MIT Press.

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

Mikheev, A. 2004. Text Segmentation. In: (ed.) Ruslan Mitkov: The Oxford Handbook of Computational Linguistics. Oxford and New York: Oxford University Press. pp. 201-218.

Nießen, S., Och, F.J., Leusch, G. and Ney, H. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000),* Athens, Greece. pp.39-45.

NIST Report 2002. Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. On line at http://www.nist.gov/speech/tests/mt/2008/doc/ngram-study.pdf [Last accessed on 08.09.2014].

O'Brien, S. 2006. Machine Translatability and Post-Editing Effort: An Empirical Study using Translog and Choice Network Analysis. PhD Thesis. Dublin City University.

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.

Och, F.J. & Ney, H. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), pp.19-51.

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.

Owczarzak, K. van Genabith, J. and Way, A. 2007b. Dependency-based automatic evaluation for machine translation. In *Proceedings of the Workshop on Syntax and Structure in Statistical Machine Translation (HLT-NAACL 2007),* Rochester, NY. pp.86-93.

Owczarzak, K., Graham, Y. and van Genabith, J. 2007a. Using f-structures in machine translation evaluation. In *Proceedings of the LFG07 Conference*, Standford, CA. pp.383-396.

Papineni, K., Roukos, S., Ward, T. and Zhu, W. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL 2002),* Philadelphia, Pennsylvania. pp.311-318.

Przybocki, M., Sanders, G. and Le, A. 2006. Edit Distance: A metric for machine translation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006),* Genoa, Italy. pp.2038-2043.

Quah, C.K. (2006) Translation and Technology. New York: Palgrave Macmillan.

Rajman, M. and Hartley, A. 2001. Automatically predicting MT systems rankings compatible with fluency, adequacy or informativeness scores. In *Proceedings of the 4th ISLE Workshop on MT Evaluation, MT Summit VIII*, Santiago de Compostela, Spain. pp.29-34.

Rosa, R., Mareček, D., and Dušek, O. 2012. DEPFIX: A System for Automatic Correction of Czech MT Outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics. Submitted.

Russo-Lassner, G. Lin, J. and Resnik, P. 2005. A paraphrase-based approach to machine translation evaluation. Technical Report LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57, University of Maryland, College Park, MD.

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.

Senellart, J. 2007. SYSTRAN MT/TM Integration. *ClientSide News Magazine*, June 2007, Feature, pp.22-25.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. 2006b. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA2006)*.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. 2006b. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 2006),* Cambridge, Massachusetts.

Somers, H. 2003. An overview of EBMT. In: Ed: Carl, M & Way, A. *Recent Advances in Example-based Machine Translation*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Surcin, S., Lange, E. & Senellart 2007. Rapid development of new language pairs at SYSTRAN. In *Proceedings of the 11th Machine Translation Summit of the International Association for Machine Translation (MT Summit XI)*, Copenhagen, Denmark.

Tatsumi, M. & Sun, Y. 2008. A comparison of statistical post-editing on Chinese and Japanese. *Localisation Focus – The International Journal of Localisation*, 7(1), pp. 22-33.

Turian, J., Shen, L. and Melamed, I.D. 2003. Evaluation of machine translation and its evaluation. In *Proceedings of the 9th Machine Translation Summit of the International Association for Machine Translation (MT Summit IX),* New Orleans. pp.386-393.

Vilar, D., Xu, J., D'Haro, L.F. and Ney, H. 2006. Error analysis of statistical machine translation output. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006),* Genoa, Italy.

Whaley, L.J. 1997. *Introduction to typology: the unity and diversity of language*. Thousand Oaks – London – New Delhi: Sage publications.

Zens, R., Och, F.J., and Ney, H. 2002. Phrase-Based Statistical Machine Translation. In M. Jarke, J. Koehler, G. Lakemeyer (Eds.): Advances in Artificial Intelligence. *25. Annual German Conference on AI, KI 2002*, Vol. LNAI 2479, pp. 18-32, Springer Verlag, September 2002.

Appendix I:

Source sentences and MT translations for the set of 25 sentences used for error analysis.

sentID	susbset	Source sentence	SMTb	SMTs	SMTh	Matxin	Google
255	Hello	Miranda Kerr is the new face of H&M's SS 14 campaign.	Miranda kerr aurpegi berria dela h&m's ss 14 kanpainan.	Miranda kerrraren aurpegi berria da h&m ss 14 kanpaina.	Miranda kerr da kanpaina horren 14. h&m ' ss aurpegi berria.	Miranda Kerr da ZERBITZU SEKRETU kanpaina 14 H&M aurpegi berria.	Miranda Kerr H & M - ren SS 14 kanpainaren aurpegi berria da.
289	Elhuyar	Nektarios Basdekis is a computer expert and a photographer.	Nektarios basdekis ordenagailuetan aditua da eta argazkilari bat.	Nektarios basdekis ordenagailu bat da aditua eta argazkilaria.	Nektarios basdekis ordenagailuetan aditua da eta argazkilari bat.	Nektarios Basdekis da ordenagailu aditu bat eta photographer bat.	Nektarios Basdekis ordenagailu aditua eta argazkilaria da.
315	BBC2	It was made using only handtools and required approximately 360 hours work.	Handtools bakarrik erabiliz egin zen eta 360 ordu inguru behar den lana.	Bakarrik erabilita egin zen, gutxi gorabehera, 360 ordu handtools eta eskatzen den lana.	It handtools bakarrik erabili egin da eta beharrezkoa da gutxi gorabehera, 360 ordu lan egin.	Hari egin zitzaion baina handtoolak erabili eta behar izanda gutxi gorabehera ordu 360 lana.	It HANDTOOLS bakarrik erabiliz egin zen, eta 360 lanordu inguru lana eskatzen.
293	Elhuyar	His work has given one of the most powerful of all impulses to the progress of science.	Eman du bere obra garrantzitsuenetako bat, bulkada guztien zientziaren aurrerapena.	Ematen du bere obra garrantzitsuenetako bat bulkada guztien zientziaren aurrerapena.	Bere lana guztietan gehien indartsuak bulkada bat eman du zientzia aurrera egiteko.	Haren lanak gehien bulkada guztietako boteretsuko bat eman du zientziako aurrerapenari.	Bere lana gehienetan bulkada guztien indartsu bat eman zientziaren aurrerapena da.
353	BBC2	So how many people should you date before you decide to settle down?	Beraz, zenbat pertsona behar duzu data erabaki aurretik bizitzen beherantz?	Beraz, jende asko behar duzu nola nahi duzun erabaki behar duzu aurreko egun egonkortzeko?	Beraz, zenbat pertsona duzun data jarri behar izango zenituzke duzun leku batean geratzeko behera erabakitzen edun baino lehen?	Beraz zuk zenbat jendeak data jarri behar izango zenituzke zuk erabakitzen edun baino lehen kokatzea down?	Beraz, zenbat pertsona behar eguneratuta duzu behera kitatzeko erabakitzen duzu aurretik ?
421	BBC2	Second lieutenant Julio Romero Marcheut, with bullet and bayonet wounds, defends himself against the Carlists.	Bigarrena: julio romero marcheut, buleta duten eta bayonet zauriak, defendatu zuen karlisten aurka.	Bigarren tenientea julio romero marcheut, bala-zauriak dituzten eta bere buruaren aurka, karlista.	Buleta duten zauriak eta bayonet bigarren: julio romero marcheut defendatu zuen karlisten aurka.	Baioneta bala zauriekin eta bigarren teniente Julio Romero Marcheutek bere burua defendatzen du Carlistsen kontra.	Bigarren teniente Julio Romero Marcheut bala eta baioneta zauriak, karlisten kontra bere burua defendatzen.

sentID	susbset	Source sentence	SMTb	SMTs	SMTh	Matxin	Google
454	BBC2	The introduction of communication technologies and Internet in direct marketing supports this idea.	Sartzeak komunikazio teknologiak eta interneteko zuzeneko marketina onartzen du ideia hori.	Sartzea, komunikazioaren teknologiak eta internetko zuzeneko marketina onartzen du ideia hori.	Komunikazioaren teknologiek interneten sartu eta zuzeneko marketina onartzen du ideia hori.	Marketin zuzeneko Ziber-espazio komunikazio teknologietako eta aurkezpenak ideia hau sostengatzen du.	Komunikazioaren teknologiak eta Internet marketing zuzenean sartzeak ideia hau onartzen.
432	BBC2	If you ask it to, Vini will reject any attempt at payment made using this card.	Eskatu nahi baduzu, vinik ahalegin guztiak baztertzen ditu, horren bidez egindako ordainketa txartela.	Galdetzen badiozu, vinik hura ezesteko saio bat egingo da ordainketa txartel hau erabiliz egin diren.	Galdetzen badiozu, vinik ahalegin guztiak baztertu egingo da, txartel horren bidez egindako ordainketa.	Vinik saiorik ukatuko dio zuk hark galdetzen badu eginda karta hau erabili ordainketari.	Duen galdetu gero, Vini -txartel hau erabiliz egindako ordainketa edozein saiakera baztertzeko egingo.
175	PACO	The death penalty constitutes a symptom of a culture of violence, not a solution to it.	Heriotza-zigorra sintoma bat da indarkeriaren kultura bat, ez da konponbidea.	Heriotza-zigorraren sintoma da indarkeriaren kultura baten konponbidea, ez baita.	Heriotza-zigorra sintoma bat da indarkeriaren kultura bat ez bada, konponbidea.	Heriotza-zigorrak indarkeriako soluzio bat hura ez kultura bateko sintoma bat osatzen du.	Heriotza zigorra indarkeriaren kultura baten sintoma bat, ez da irtenbide bat osatzen du.
200	PACO	The way we play as children informs the skills we develop.	Nola jokatu azaltzen dugu haurrek gaitasunak garatzen ditugu.	Haurrek, era batera edo bestera jokatuei jakinarazten die gaitasunak garatuko ditugu.	Nola jokatu azaltzen dugu haurrek garatzen ditugun gaitasunak.	Bideak guk jokatzen dugu umeek guk garatzen ditugun trebetasunak jakinarazitzen dituzte.	Bide haurrak bezala jokatuko dugu jakinarazten gaitasunak garatzen ditugu.
144	PACO	The inaugural shortlist of the latest literary award on the block, the Folio Prize, has been unveiled.	Hautagaiak zein zen azken sari literario on the block, folioa saria, agerian jarri da.	Aukeraketa hau inauguratzea literatura sariaren azken blokean, folioa saria, aurkeztu da.	Hautagaiak zein zen azken literatur saria on the block, folioa saria, agerian jarri da.	Bloke Prize Folioko literaturako azken sariko hautagai-zerrenda inauguraziokoa ezagutarazitua izan da.	Inaugurazio azken literatur blokea, Folio Saria da saria laburrean izan, ha inauguratu dira.
206	BBC1	The trailer for the highly-anticipated and controversial film inspired by the events of the Amanda Knox trial is here.	Aurrerakin hori dagokion highly-anticipated eta polemikoa filma inspiratu gertakari garrantzitsuenak knox amanda da saiakuntza hemen.	Aurrerakinak eta eztabaidagarri highly-anticipatedrako filmek inspiratu ekitaldi amanda knoxen probako bertsioa da hau.	Film polemikoa highly-anticipated egiteko eta eragindako gertaerak, amanda knox saiakuntza aurrerakin hori da hemen.	Adoretuta epaiketa Amanda Knoxeko gertaerek highly-anticipated geruzarentzat eta eztabaidagarriarentzat karabana da hemen.	Film handia espero eta polemikoa Amanda Knox epaiketaren ekitaldiak inspiratutako trailerra hemen.

sentID	susbset	Source sentence	SMTb	SMTs	SMTh	Matxin	Google
7	PACO	Facebook does not hand over full access to a person"s account due to privacy concerns.	Facebook esku ez sarbide osoa pertsona bat kontua pribatutasuna dela eta lotuta.	Facebook ez du esku sarbide osoa pertsona baten kontuak direla - eta pribatutasun-arazoak.	Facebook esku ez sarbide osoa pertsona baten kontua dela eta pribatutasuna lotuta.	Facebookek egiten du ez over sarbide betea pertsona baten intimitate kezketara ordaintzeko kontura eskua.	Facebook ez du gehiago entregatu pertsona baten kontura sartzea osoa dela eta pribatutasuna kezka.
462	BBC2	Each bet costs 1 Euro.	Apustu bakoitzak 1 euro kostatzen da.	Apustu bakoitzaren kostua 1 euro.	Apustu bakoitzean 1 euro kostatzen da.	Apustu bakoitza Euro kostatzen da 1.	Apustu bakoitzaren prezioa 1 euro.
154	PACO	The rainy season in Bolivia usually lasts until March.	Bolivian euri-sasoian normalean martxora arte irauten du.	Bolivian, oro har, euri-sasoi martxora arte irauten du.	Bolivian euri-sasoian martxora arte irauten du normalean.	Boliviako urtaro euritsua normalean irauten da Martxo arte.	Bolivian denboraldian euritsua normalean martxora arte irauten du.
174	PACO	At least this way, they have been able to see the child smiling from time to time.	Gutxienez, horrela, ikusi ahal izan duten haurrak irribarrez noizean behin.	Gutxienez, horrela, gai izan dira umea ikusteko, irribarre noizean behin.	Gutxienez horrela ikusi ahal izan dituzte haurrak une batetik bestera, irribarrez.	Haiek izan dira bide gutxienez hau irribarre egitean denboratik umea ikuste gai den denborara.	Gutxienez modu horretan, haurraren noizean irribarrez ikusi ahal izan dute.
253	Hello	Your innate love of animals brought you to chimpanzees.	Zure maite duen berezko ekarritako animalia txinpantzeak ditzakezu.	Zure sortu maitasuna ekartzen baduzu, txinpantzeak animalia.	Zure sortzetiko maitasuna ekartzen baduzu txinpantzeak animalia.	Animalien zure sortzetiko maitasunak zu ekarrita txinpantzeetara.	Zure animaliak maitasuna berezkoa ekarri txinpantzeen behar duzu.
19	PACO	Retrieve the names of all senior students majoring in computer science.	Senior guztien izenak eskuratu, informatika-ikasketak egiten ari diren azken mailako ikasleak.	Guztien izenak eskuratu informatika-ikasketak egiten ari diren azken mailako ikasleak.	Guztien izenak eskuratu ondoren, informatika-ikasketak egiten ari diren azken mailako ikasle.	Garrantzitsutzean informatikan ikasle zaharren guztien izenak berreskuratu itzazu.	Senior ikasle guztiak informatika majoring izenak berreskuratzeko.
172	PACO	The final eight books span poetry, novels and short stories.	Zortzi hutsarte azken liburuak poesia, nobela eta ipuin.	Azken zortzi liburu-hedadura poesia, nobela eta kontakizun laburrak.	Azken zortzi hutsarte poesia liburu nobelak eta ipuinak.	Azken liburu zortzik poesia nobelak eta baxu istorioak zeharkatzen dituzte.	Final zortzi liburuak hartzen poesia, nobela eta kontakizun laburrak.
299	Elhuyar	So the ships were forced to shelter in Santurce until conditions were viable to entry.	Beraz, itsasontziak babesa behar izan zituen, harik eta baldintzarik santurce bideragarriak eta sarrera.	Beraz, ontziek behar izan zituen aterpe baldintzak ez zirela bideragarriak santurtziko sarrerara arte.	Beraz, ontziak izan zen santurceko babesteko. baldintzak ez zirela bideragarriak sarrera egin arte.	So ontziak behartuak ziren Santurceko aterpera baldintzak ziren sarrerara bideragarria.	Beraz ontziak ziren Santurtzin aterpe behartu baldintza sarrera bideragarriak ziren arte.

sentID	susbset	Source sentence	SMTb	SMTs	SMTh	Matxin	Google
3	PACO	An exemption on biodegradable bags would be disastrous.	Salbuespen bat: poltsa biodegradagarriak hondamena litzateke.	Poltsa biodegradagarrian salbuespen bat oso arriskutsua izango litzateke.	Salbuespen bat: poltsa biodegradagarriak hondamena izango litzateke.	Poltsa biodegradagarrietako salbuespen bat izango litzateke zorigaiztokoa.	Poltsa biodegradagarriak on salbuespena negargarria izango litzateke.
262	Hello	What message does that send out?	Mezua bidaltzen duten, zer egiten du?	Zer kanpora bidaltzen du mezua?	Zer mezu bidaltzen duten egiten al du?	Mezu what egiten du bidaltzen duela out?	Zer mezua bat bidaltzen du kanpora ?
335	BBC2	You can always consult your correspondence at Clavenet and can receive postal deliveries again whenever you like.	Beti izango duzu zure korrespondentzia clavenet kontsultatu, eta berriro jaso ditzake posta entregak, nahi duzun bakoitzean.	Beti kontsultatu ahal izango duzu zure korrespondentzia clavenet berriro jaso ditzake posta-entregak eta edozein unetan izanen duzu.	Kontsultatu ahal izango dituzu, eta zure korrespondentzia clavenet jaso ditzake posta entregak berriro nahi duzun bakoitzean.	Zuk beti zure korrespondentzia kontsulta dezakezu Clavenetekin eta posta banaketak har ditzake berriro zuk gogoko duzunean.	Daiteke beti kontsultatzen baduzu zure korrespondentzia Clavenet at eta posta bidalketak berriro jaso ahal Nahi duzunean.
64	PACO	Prostitution crosses that line for you.	Prostituzioa gurutzeak lerro hori zuretzat.	Prostituzioa zeharkatzen duen lerroa zuretzat.	Prostituzioa gurutzeak lerro hori zuretzat.	Prostituzioak gurutzatzen du zuretzat lerroa.	Prostituzioa zuretzat lerro hori zeharkatzen.
328	BBC2	The dresses were adorned with thousands of sequins and crystals.	Soinekoak; ziren, eta milaka sequins eta kristalak.	Jantzi apaindu zituzten milaka sequins eta kristalak.	Jantzi zituzten apainduta dago, eta milaka sequins eta kristalak.	Soinekoak apainduak ziren beira sequinetako eta thousandez.	Soinekoak sequins eta kristalak milaka apaindutako ziren.