Labaka, G., España-Bonet, C., Màrquez, L. et al. *A hybrid machine translation architecture guided by syntax.* **Machine Translation** 28, 91–125 (2014). This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1007/s10590-014-9153-0

A Hybrid Machine Translation Architecture Guided by Syntax*

Gorka Labaka · **Cristina España-Bonet** · **Lluís Màrquez** · **Kepa Sarasola**

Received: 17 December 2013 Accepted: 17 August 2014

Abstract This article presents a hybrid architecture which combines rule-based machine translation (RBMT) with phrase-based statistical machine translation (SMT). The hybrid translation system is guided by the rule-based engine. Before the transfer step, a varied set of partial candidate translations is calculated with the SMT system and used to enrich the tree-based representation with more translation alternatives. The final translation is constructed by choosing the most probable combination among the available fragments with a monotone statistical decoding following the order provided by the rule-based system. We apply the hybrid model to a pair of distant languages, Spanish and Basque, and perform extensive experimentation on two different corpora. According to our empirical evaluation, the hybrid approach outperforms the best individual system across a varied set of automatic translation evaluation metrics. Following some output analysis to better understand the behaviour of the hybrid system, we explore the possibility of adding alternative parse trees and extra features to the hybrid decoder. Finally, we present a twofold manual evaluation of the translation systems studied in this paper, consisting of (*i*) a pairwise output comparison and (*ii*) a individual task-oriented evaluation using HTER. Interestingly, the manual evaluation shows some contradictory results with respect to the automatic evaluation: humans tend to prefer the translations from the RBMT system over the statistical and hybrid translations.

* The final publication is available at Springer via http://dx.doi.org/10.1007/s10590-014-9153-0

G. Labaka
IXA Research Group, Department of Computer Languages and Systems, University of the Basque Country (UPV/EHU), Manuel de Lardizabal 1, 20018 Donostia, Spain. E-mail: gorka.labaka@ehu.es

C. España-Bonet
TALP Research Center, Department of Computer Science, Technical University of Catalonia – Barcelona Tech. Jordi Girona 1-3, 08034 Barcelona, Spain. E-mail: cristinae@lsi.upc.edu

L. Màrquez
Qatar Computing Research Institute, Qatar Foundation, Tornado Tower, Floor 10, P.O. Box 5825, Doha, Qatar. E-mail: lmarquez@qf.org.qa. During the research period of this article, he was a member of the TALP Research Center, Department of Computer Science, Technical University of Catalonia – Barcelona Tech.

K. Sarasola
IXA Research Group, Department of Computer Languages and Systems, University of the Basque Country (UPV/EHU), Manuel de Lardizabal 1, 20018 Donostia, Spain. E-mail: kepa.sarasola@ehu.es

Keywords Hybrid machine translation · rule-based MT · phrase-based statistical MT · Spanish–Basque MT

1 Introduction

Different machine translation (MT) paradigms have different advantages and disadvantages. When comparing systems developed within the two main MT paradigms, namely *rule-based* and *phrase-based statistical* machine translation (RBMT and SMT, respectively), one can observe some complementary properties —see [11] for a comparative introduction to both approaches. RBMT systems tend to produce syntactically better translations and deal with long distance dependencies, agreement and constituent reordering in a more principled way, since they perform the analysis, transfer and generation steps based on morphosyntactic knowledge. On the downside, they usually have problems with lexical selection due to a poor modeling of word-level translation preferences. Also, if the input sentence is unparsable due to the limitations of the parser or because the sentence is ungrammatical, the translation process may fail and produce very low quality results. Phrase-based SMT models are usually better with lexical selection and fluency, since they model lexical choice with distributional principles and explicit probabilistic language models trained on very large corpora. However, SMT systems may generate structurally worse translations and experience problems with long distance reordering, since they model translation information more locally. Sometimes, they can produce very obvious errors, which are annoying for casual users, e.g., lack of local gender and number agreement, bad punctuation, etc. Moreover, SMT systems can experience a severe performance degradation when applied to domains different from those used for training (*out-of-domain* evaluation).

Given the complementarity of the pros and cons from both approaches, several proposals of *combined* and *hybrid* MT models have emerged in the recent literature, with the aim of getting the best of both worlds (see Section 2 for a discussion on the related work). In this article we present a hybrid architecture, which tries to exploit the advantages of RBMT and phrase-based SMT paradigms. Differently to the dominant trend in the previous years, we do not make a posterior combination of the output of several translators, nor enrich a statistical translation system with translation pairs extracted from the rule-based system. Instead, we approach hybridisation by using a RBMT system (Matxin [34]), to guide the main translation steps and a SMT system to provide more alternative translations at different levels in the transferred and reordered syntactic tree before generation. The generation of the final translation implies the selection of a subset of these pre-translated units from either systems. This is performed by a monotone phrase-based statistical decoding, following the reordering given by Matxin and using a very simple set of features. We will refer to our hybrid system as 'Statistical Matxin Translator', or SMatxinT in short.

The rationale behind this architecture is that the RBMT component should perform parsing, and rule-based transfer and reordering to produce a good structure in the output, while SMT helps in the lexical selection by providing multiple translation suggestions for the pieces of the source language corresponding to the tree constituents. In practice, the SMT subsystem may also work as a back off. If the RBMT component does a very bad analysis, transfer, or reordering, the hybrid decoder will still be able to pick translations produced by the SMT system alone, which may correspond to very long segments or even the complete source sentence.[1] The decoder accounts also for fluency by using language models. Since

[1] A complete SMT-translation of the source sentence is always available in the hybrid decoder.

the structure of the translation is already decided by the RBMT subsystem, this decoding can be monotone and therefore efficient.

Note that there are several other works in the literature that extended some components of RBMT systems with statistical information (e.g., corpus-based lexical selection [44] and transfer rules learnt from bilingual corpora [50]). Our approach is conceptually related to the former, but it extends the lexical selection concept to the selection of arbitrarily long partial translations and, moreover incorporates statistical features into the decoding accounting for fluency and source-system selection. Another important remark is that our system also differs from using a regular phrase-based SMT system after rule-based reordering. First of all, the partial translations provided by the RBMT system are also included in the decoding. Secondly, the translation pairs incorporated from the SMT system are produced by translating source fragments corresponding to syntactic nodes in the RBMT tree, thus being different from a traditional SMT translation table. The dependencies between RBMT and SMT in our hybrid system are in both directions.

We have instantiated and applied the hybrid architecture to a pair of structurally and morphologically distant languages, Spanish and Basque, and performed experimental evaluation on two different corpora. First of all, these experiments allowed us to validate the assumptions on which the hybrid architecture rely and to better understand the behaviour of the system. Second, the obtained results showed that SMatxinT outperforms the individual RBMT and SMT systems, with consistent improvements across a varied set of automatic evaluation measures and two different corpora. One issue also analysed in this article is the strong dependence of the hybrid system on the syntactic analysis of the source sentence. In an attempt to increase robustness against parsing errors, we incorporated an alternative parse tree in the mix as a way to increase parsing diversity. Also, some linguistically-motivated features for the hybrid decoder were proposed to compensate for the strong preference of the hybrid system towards using more SMT-originated partial translations.

Finally, we also conducted two types of manual evaluation on a small sample of the datasets in order to explore qualitative differences among models. Interestingly enough, the manual evaluation reflected some contradictory results compared to the automatic evaluation. Humans tended to prefer the translations from the RBMT system over the statistical and hybrid translations. The optimisation of our hybrid system was done against the automatic metrics, which we managed to improve on the evaluation benchmarks. Unfortunately, this was not appropriately modelling the human perception of translation quality.

An initial version of the hybrid system presented in this work was introduced in a previous article by the same authors [16]. The current work builds on the previous one and provides extensions along three different lines: (*i*) the hybrid system is evaluated in more depth, including a study of its output; (*ii*) the incorporation of alternative parse trees and extra features is explored; (*iii*) a thorough manual evaluation of translation quality is performed, providing quantitative and qualitative comparative results.

The rest of the article is organised as follows. Section 2 overviews the related literature on MT system combination and hybridisation. Section 3 presents our hybrid architecture describing in detail all its components, including the individual MT systems used. Section 4 describes the experimental work carried out with the hybrid architecture and discusses the results obtained. Section 5 covers the human evaluation of the proposed systems. Finally, Section 6 concludes and highlights open issues for future research.

2 Related Work

This section is structured into five subsections. The first three overview the main approaches to compound systems for machine translation: (*i*) system output combination, (*ii*) hybridisation with SMT leading, and (*iii*) hybridisation with RBMT leading. The fourth subsection explains the differences of our approach from the previously reviewed approaches. Finally, the last subsection discusses some work on the differences between automatic and manual evaluations and the lack of consistency between their result, as this topic is also a relevant part of this article.

2.1 System combination

System combination, either serial or by a posterior combination of systems' outputs, is a first step towards hybridisation. Xu et al. [51] presented a system that combines three SMT and two RBMT systems at the level of the final translation output. They applied the CMU open toolkit MEMT [27] to combine the translation hypotheses. Translation results showed that their combined system significantly outperforms individual systems by exploiting strengths of both rule-based and statistical translation approaches. Although it has shown to help in improving translation quality, system combination does not represent a genuine hybridisation since systems do not interact among them (see [48] for a classification of HMT architectures).

Combination strategies usually make use of confusion networks in order to combine fragments from a number of different systems [33]. The standard process to build such confusion networks consists of two steps: (*i*) Selection of a backbone. After predicting quality scores for the translations produced by the systems participating in the combination, the one with the best score is picked. (*ii*) Monolingual word alignment between the backbone and other hypotheses in a pairwise manner. Once such a confusion network is built, one can search for the best path using a monotone consensus network decoding. One crucial factor in the overall performance of this type of system combination resides in the selection of a backbone. For example, Okita et al. [41] improve system combination results by means of a new backbone selection method based on Quality Estimation techniques.

2.2 Hybridisation with SMT leading

Much work has been done on building hybrid systems where the statistical component is in charge of the translation and the companion system provides complementary information. For instance, both Eisele et al. [14] and Chen and Eisele [9] introduced lexical information coming from a rule-based translator into a SMT system in the form of new phrase pairs for the translation table. In both cases, translation quality is improved on out-of-domain tests according to automatic measures ($\sim$1.5 points of BLEU increase). Unfortunately, they did not conduct any manual evaluation.

Sanchez-Cartagena et al. [43] built a system consisting of enriching a SMT system with bilingual phrase pairs matching transfer rules and dictionaries from a shallow-transfer RBMT system. Automatic evaluation showed a clear improvement of translation quality (*i*) when the SMT system was trained on a small parallel corpus, and (*ii*) when it was trained on larger parallel corpora and the texts to translate came from a general (news) domain that

was well covered by the RBMT system. Human evaluation was performed only in the latter scenario and it confirmed the improvements already measured automatically.

2.3 Hybridisation with RBMT leading

The opposite direction, that is, where the RBMT system leads the translation and the SMT system provides complementary information, has been less explored. Habash et al. [26] enriched the dictionary of a RBMT system with phrases from an SMT system. They created a new variant of GHMT (Generation-Heavy Machine Translation), a primarily symbolic system, extended with monolingual and bilingual statistical components that had a higher degree of grammaticality than a phrase-based statistical MT system. Grammaticality was measured in terms of correctness of verb–argument realisations and long-distance dependency translation. They conducted four sets of experimental evaluations to explore different aspects of the data sets and system variants: (*i*) automatic full system evaluation; (*ii*) automatic genre-specific evaluation; (*iii*) qualitative evaluation of some concrete linguistic phenomena; and (*iv*) automatic evaluation with rich linguistic information (English parses).

Enache et al. [15] presented a system for English–to–French patent translation. The language of patents follows a formal style appropriate for being analysed with a grammar, but, at the same time, it uses a rich and particular vocabulary which is better dealt with statistical methods and existing corpora. Following this motivation, the hybrid system translated recurrent structures with a grammar and used SMT translation tables and lexicons to complete the translation of the sentences. Both manual and automatic evaluations showed a slight preference for the hybrid system in front of the two individual translators.

Dove et al. [13] used RBMT output as a baseline, and then refined it through comparison against a language model created with SMT techniques. Similarly, Federmann et al. [20] used the translations obtained with a RBMT system and substituted selected noun phrases by their SMT counterparts. Globally, their hybrid system improved over the individuals when translating into languages with a richer morphology than the source. In a later work, Federmann et al. [19] based the substitution on several decision factors, such as part-of-speech, local left and right contexts, and language model probabilities. For the Spanish–English language pair each configuration performed better than the baseline, but the improvements in terms of BLEU score were small and not conclusive. In a similar experiment, Sánchez-Martínez et al. [45] reported small improvements in English–to–Spanish translation and vice versa, when using marker-based bilingual chunks automatically obtained from parallel corpora.

Federmann et al. also presented a hybrid English–to–German MT system to the WMT11 shared translation task [21]. Their system was able to outperform the RBMT baseline and turned out to be the best-scored participating system in the manual evaluation. To achieve this, they extended a rule-based MT system to deal with parsing errors during the analysis phase. A module was devised to specifically compare and decide between the tree output by a robust probabilistic parser and the multiples trees from the RBMT analysis forest. The probabilistic parse complemented well the system in the difficult cases for the rule-based parser (ungrammatical input or unknown lexica). Their MT system was able to preserve the benefits of a rule-based translation system, such as a better generation of the target language text. Additionally, they used a statistical tool for terminology extraction to improve the lexicon of the RBMT system. They reported results from both automatic evaluation metrics and human evaluation exercises, including examples showing how the proposed approach improved machine translation quality.

There are several machine-learning-based frameworks for hybrid MT in the literature as well. Federmann [18] showed how a total order can be defined on translation output and used for feature vector generation. His method differs from the previous work as he considers joint binarised feature vectors instead of separate ones for each of the available source systems. He proposed an algorithm to use a classification model trained on these feature vectors to create hybrid translations. Hunsicker et al. [28] described a substitution-based hybrid MT system extended with machine learning components to control phrase selection. Their approach is guided by a RBMT system, which creates template translations. The substitution process was either controlled by the output of a binary classifier trained on feature vectors from the different MT engines, or dependent on weights for the decision factors, which were tuned using MERT. As for evaluation, they observed improvements in terms of BLEU scores over a baseline version of the hybrid system.

2.4 Our approach

A previous manual study on the quality of our in-house individual systems for Spanish–to–Basque translation, revealed that the best performing system was the RBMT system, especially on out-of-domain examples [31]. This is what motivated the proposal of a hybrid system where the RBMT system leads the translation and the SMT system provides complementary information. The strategy of our system does not involve a confusion network, like in the system combination approach. Instead, a monotone statistical decoding is applied to combine the alternative translation pairs following the backbone and the order given by the RBMT generation parse tree.

Similar in spirit to the systems described in subsection 2.3, translations produced by our system (SMatxinT) are guided by the RBMT system in a way that will be clarified in the following sections. In contrast to these others systems, SMatxinT is enriched with a wider variety of SMT translation options, including not only short local SMT phrases, but also (*i*) translations of longer fragments up to the complete sentence (potentially useful when the RBMT system fails at producing good translations due to parsing errors), and (*ii*) SMT translation(s) of each node in the tree extracted from a broader context, that is, extracted via alignments from translations of higher level nodes. Finally, note that the alternative translations coming from the SMT system are guided by the structure of the RBMT parse tree, and thus, are not simply a copy of the translation table.

2.5 Automatic and manual evaluations

Automatically comparing the performance of MT systems from different paradigms is a real open problem for the MT research community. For example, the organisers of the Eighth Workshop on Statistical Machine Translation (WMT13, [6]) recognised recently that system results are very difficult to compare to each other, mainly because the automatic metrics used are often not adequate, as they do not treat systems of different types fairly. In the context of this article, where several types of systems need to be compared, this is a serious problem. In the last years, a myriad of metrics for automatic evaluation have been proposed, some of them including high level linguistic analysis (e.g., using syntactic and semantic structures). But WMT13 organisers claim that there is no consolidated picture, and that different metrics seem to perform best for different language aspects and system types.

Besides, the use of human evaluation is not widely extended (note that all but three of the papers mentioned in this section just ignore it). It is a fact that the results of manual and automatic evaluations do not always agree. The organisers of the Workshop on Machine Learning for Hybrid Machine Translation (ML4HMT-2011) concluded that a more systematic comparison of hybrid approaches needed to be undertaken, both at a system level and with respect to the evaluation of such systems' output [17]. Remarkably, they developed the *Annotated Hybrid Sample MT Corpus*, which is a set of 2,051 sentences translated by five MT systems of different nature (Joshua, Lucy, Metis, Apertium, and MaTrEx), and organised a shared task on applying machine learning techniques to optimise the division of labour in Hybrid MT. The evaluation results of the four participants were clearly controversial. The best system according to nearly all the automatic evaluation measures only reached a third place in the manual evaluation. And vice versa, the best system according to the manual assessments ranked last in the automatic evaluation. In the following edition (ML4HMT-2012 workshop) they obtained contradictory evaluation results as well. The DFKI system performed best in terms of METEOR while the DCU-DA system achieved the best performance for NIST and BLEU scores [22]. Unfortunately, a manual evaluation of the participants in the 2012 workshop was not carried out.

In this article, we fundamentally use automatic evaluation with a variety of available metrics in the development phase of our hybrid MT system. When testing and comparing the hybrid system against the individual ones, we use again the same set of automatic metrics, but also perform a thorough manual evaluation on a small subset of the data in order to gain more insight on the differences observed across systems.

Among the above described systems for hybridisation with RBMT leading, only Enache et al. [15] performed both manual and automatic evaluations. In both cases they showed some advantage of the hybrid system compared to the two individual translators. But it has to be noted that their system is not a translation system for open text as SMatxinT. Instead, it is a specialised translator for the restricted domain of patent translation.

3 System Architecture

Our hybrid model is built on a rule-based Spanish–Basque machine translation system (Matxin) and the best phrase-based SMT systems available in-house; an standard phrase-based statistical MT system developed with Moses, and a modification of it to specifically deal with Basque morphology. The following subsection describes the individual systems and variants, whereas Section 3.2 presents the full hybrid architecture.

3.1 Individual systems

Matxin. Matxin is an open-source RBMT engine for Spanish–to–Basque translation [4]. The engine is based on the traditional transfer model which is divided in three steps: (*i*) analysis of the source language into a dependency tree structure, (*ii*) transfer from the source language dependency tree to a target language dependency structure, and (*iii*) generation of the output translation from the target dependency structure. In the following we briefly describe each of the stages. For more details, please refer to the Matxin publication [4]. Note that due to authorship issues the bilingual dictionary openly distributed is a reduce version of the one we used for research. Nevertheless, the rest of the modules used here are part of the open-source engine.

The *analysis* step is done with a modified version of FreeLing [8]. The shallow parser was adapted to parse the source Spanish sentences into dependency trees. *Transfer* from Spanish into Basque is done in two parts: lexical and structural. For the *lexical transfer*, Matxin uses a bilingual dictionary based on the Elhuyar wide-coverage dictionary[2] compiled into a finite-state transducer. Parallel corpora were also used to enrich this dictionary with named entities and terms. Verb-argument structure information, automatically extracted from the Basque monolingual corpus, was used to disambiguate among the possible translation of prepositions. For the *structural transfer*, that is, going from the source language tree to the target language tree, a set of manually developed rules was applied. Matxin also contains a specific module for translating verb chains [3]. *Generation*, like transfer, is decomposed into two steps. The first step, *syntactic generation*, consists of deciding in which order to generate the target constituents within the sentence, and the order of the words within the constituents. The second step, *morphological generation*, consists of generating the target surface forms from the lemmas and their associated morphological information.

Baseline SMT system. Our *b*asic statistical MT system, SMT_b, is built using freely available state-of-the-art tools: the GIZA++ toolkit [38] to estimate word alignments, the SRILM toolkit [47] for the language model and the Moses decoder [30]. In the development we used a log-linear [39] combination of several common feature functions: phrase translation probabilities in both directions, word-based translation probabilities also in both directions, a phrase length penalty, a word penalty and the target language model. We also used a lexical reordering model ('msd-bidirectional-fe' training option in Moses scripts). The language model is a simple 3-gram language model using the SRI Language Modelling Toolkit, with modified Kneser-Ney smoothing. The language modelling is limited to 3-grams due to the high sparsity derived from the Basque rich morphology and the limited size of monolingual text (28 million words). Parameter optimisation was done by means of Minimum-Error-Rate Training, MERT [38]. The metric used to carry out this optimisation is BLEU [42].

Morpheme-based SMT. A second variant of the SMT system, SMT_g (*'g'* for generation; see below), is used to address the rich morphology of Basque. In this system, each Basque word is split into several tokens using Eustagger [1], a well known Basque morphological lemmatiser/tagger. We create a different token for each morpheme, where the affixes are replaced by their corresponding morphological tags. By dividing words in this way, one expects to reduce the sparseness produced by the agglutinative nature of Basque and the small size of the parallel training corpus. Adapting the baseline system to work at the morpheme level mainly consists in training Moses on the segmented text, using the same training options as for SMT_b. The translation system trained on segmented words generates sequences of morphemes. So, in order to obtain the final Basque text from the segmented output, a generation post-process is necessary. We also incorporate a word-level language model after generation (note that the language model used for decoding at morpheme level is trained on the segmented text). As in Oflazer et al. [40], we use n-best list reranking to incorporate this word-level language model.

3.2 Hybrid architecture

The design of the SMatxinT architecture is motivated from the previously commented pros and cons of general RBMT and SMT systems. Our aim is threefold. First, the hybrid system

[2] http://www.elhuyar.org

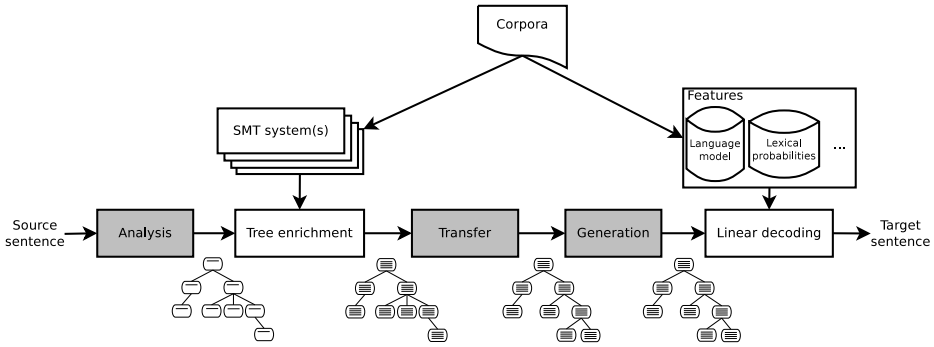

Fig. 1 General architecture of SMatxinT. The Matxin (RBMT) modules which guide the MT process are depicted as grey boxes. The two new processing modules are *tree enrichment* and *linear decoding*.

should delegate most of the syntactic structure and reorder of the translation to the RBMT system. Second, the hybrid system should be able to correct possible mistakes in the syntactic analysis by backing off to SMT-based translations. Third, SMT local translations of short fragments are also considered as they can improve lexical selection. On top of the previous three aspects, we also consider a statistical language model, which may help producing more fluency translations.

The main idea of the hybrid system is to enrich every node of the RBMT translation tree with one or more SMT translation options and then implement a mechanism to choose which translation options are the most adequate ones following the order of the tree. Within our framework, this means that SMatxinT adopts the architecture and data structures from Matxin (see previous section). The traditional transfer model of Matxin is modified with two new steps: (*i*) A *tree enrichment* phase is added after analysis and before transfer, where SMT translation candidates are added to each node of the tree. These translations correspond to the text chunks dominated by each tree node (i.e., the syntactic phrases identified by the parser) and they go from individual lexical tokens to the complete source sentence in the root. (*ii*) After generation, an additional *monotone decoding* step is responsible for generating the final translation by selecting among RBMT and SMT partial translation candidates from the enriched tree. This architecture is depicted in Figure 1 where one can see how the new SMatxinT modules are integrated within the RBMT chain (in grey). The following subsections explain in more detail the two new modules.

3.2.1 Tree enrichment

After syntactic analysis and before transfer the tree enrichment module uses one (or several) SMT systems to assign multiple alternative translations for each source text chunk. This process relies on the phrase segmentation created by Matxin dependency parser and incorporates, for each node in the tree, three types of translations:

1. $local_{SMT}$: SMT translation of the lexical tokens contained in the node
2. $local_{SMT\text{-}C}$: SMT translation(s) of the node in a broader *c*ontext, that is, extracted from translations of higher level nodes in the tree. Only the shorter translation that is consistent with the word alignment is selected. Even so, each node can contain more than one translation, one for each ancestor node
3. $full_{SMT}$: SMT translation corresponding to the entire subtree dominated by the node. The final decoder will have to choose between using this translation or a combination

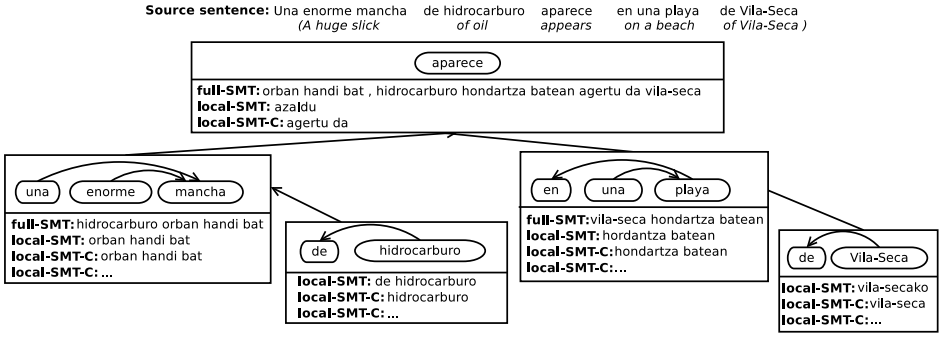

Fig. 2 Example of a dependency tree enriched with SMT translations. To allow for a more simple figure, only one *local-SMT-C* translation has been displayed.

of local translation (SMT or RBMT) of the nodes that compose it. All the proposed translations are extracted according to the parsing, so that their boundaries coincide

These three types of SMT translation candidates intend to satisfy the goals set in the design of the hybrid translator. With $local_{SMT}$ the hybrid translates the source in the order specified by the RBMT system, but choosing for each chunk among RBMT or SMT translation alternatives. $full_{SMT}$ translations allow the hybrid system to use longer pure SMT translations, recovering from potential structural errors in the parse tree. In the limit, the hybrid system would be able to use the SMT translation for the full sentence, contained in the root node. Finally, $local_{SMT\text{-}C}$ translations try to address the potential problem caused by the short length of some chunks and the difficulty of translating them without the context of the full sentence.

Figure 2 shows an example with an enriched tree for the source sentence in Spanish *"Una enorme mancha de hidrocarburo aparece en una playa de Vila-Seca"* ("A huge oil slick appears on a beach in Vila-Seca"). The main verb, *"aparece"* ("appears"), is the root node of the dependency tree. The chunk *"aparece"* is translated alone as *"azaldu"* ($local_{SMT}$) and as *"agertu da"* ($local_{SMT-C}$) when extracted from the whole sentence translation (*"orban handi bat, hidrocarburo hondartza batean* **agertu da** *vila-seca"*). Focusing on the node *"en una playa"* ("on a beach"), we see that there is only one SMT translation, since $local_{SMT}$ and $local_{SMT-C}$ coincide. The $full_{SMT}$ translation in that node produces a translation for the complete text span corresponding to the subtree dominated by the node (*"en una playa de Vila-Seca"*).

The number of actual SMT translations incorporated at every node can be variable. One can use one or more SMT systems and incorporate n-best translations for any of them. In our experiments, we used the two individual SMT systems from Section 3.1, SMT_b and SMT_g. As for the n-best translations, we restricted ourselves to $n = 1$, as longer n-best lists did not produce significantly better translations.

3.2.2 Monotone decoding

The modules applied after tree enrichment are transfer and generation (see Figure 1). Only minor modifications were required, basically to keep the SMT translations introduced by the tree enrichment module. At the end of the generation process, we have a tree structure

defining the order of the chunks to construct the translation. But for each chunk there are available both the RBMT translation and a set of SMT translations. At this point, one needs to apply a final process to decide which is the most adequate translation for each chunk to construct the final translation of the complete sentence. This is similar to the search process conducted by an SMT decoder but simplified, since reordering is not allowed. One can, therefore, use a standard statistical decoder (Moses in our case) for the monotone decoding. This kind of decoder naturally deals with differences in scope of the candidate translation ($local_{SMT}$ vs. $full_{SMT}$). The set of features can also be simplified. All the available candidates have already been chosen by any of the systems as the preferred translation for the corresponding chunks. Features such as the language model, the individual system that produced a candidate chunk translation, or the number of individual systems that propose the same chunk translation should be more important for the final decoding.

Our basic set of features is made up of seven features[3]. We can divide them into those common to standard SMT decoding and those depending on the individual systems producing candidate translations:

SMT standard features

1. Language Model (LM): the same n-gram target language model used in the SMT systems
2. Word Penalty (WP): count of words used in the translation
3. Phrase Penalty (PP): count of phrases used in the translation

Source/Consensus features

1. Counter: number of systems that generated an identical candidate chunk translation
2. SMT: indicates whether the candidate is generated by an SMT system or not
3. RBMT: indicates whether the candidate is generated by the RBMT system or not
4. BOTH: When both individual systems (SMT and RBMT) generates an identical translation candidate, count of translated source words, otherwise zero. Using the count of source words instead of the count of phrases allows the decoder to treat phrases differently according their length

Note that our approach for the final decoding is substantially different from that of Federmann et al. [20]. We do not try to identify chunks that the RBMT system translates incorrectly to substitute them with their SMT counterparts. But let the decoder select among all the available options. The more informed the decoder is the better the selection will be. Besides, we do not face all the problems associated with the alignments between systems because, as previously said, the tree is enriched with candidates that are obtained by running SMT systems for each of the segments (or subtrees) given by the RBMT system. We also use local translations extracted from broader context translations, but always within the same system.

4 Experimental Work

In this section we start by describing the corpora used to train the statistical and hybrid systems as well as some relevant system development details. Then, we devote a complete

[3] Note that due to the log-linear approach use in Moses, the features should be stored in the phrase-table as exponential of e. The features are presented in the way the decoder will see them.

	Sentences	Tokens Spanish	Tokens Basque
EHUBooks	39,583	1,036,605	794,284
Consumer	61,104	1,347,831	1,060,695
ElhuyarTM	186,003	3,160,494	2,291,388
EuskaltelTB	222,070	3,078,079	2,405,287
Total	491,853	7,966,419	6,062,911

Table 1 Statistics on the bilingual collection of parallel corpora.

subsection to present each of the experiments carried out and discuss the corresponding results.

Corpora. We used a heterogeneous bilingual corpus including four Basque–Spanish parallel corpora: (*i*) six reference books translated manually by the translation service of the University of the Basque Country (EHUBooks); (*ii*) a collection of 1,036 articles of the Consumer Eroski magazine[4] written in Spanish along with their Basque translation (Consumer); (*iii*) translation memories mostly using administrative language developed by Elhuyar[5] (ElhuyarTM); and (*iv*) a translation memory including short descriptions of TV programmes (EuskaltelTB). The entire dataset makes a total of 491,853 sentences with 7,966,419 tokens in Spanish and 6,062,911 tokens in Basque. Table 1 shows some statistics on the corpora, specifying the number of sentences and tokens per collection. It is worth noting that the bilingual collection is rather small compared to the sizes of the parallel corpora available for the language pairs with more resources (e.g, the seventh version of Europarl Corpus[6] contains almost 50 million words per language pair, six times larger than what we have for Spanish-Basque), and, therefore, we might expect some sparseness in pure statistical approaches. Note as well that the number of tokens on the Basque side is much lower compared to Spanish. This is due to the rich morphology and the agglutinative nature of the language.

The training corpus is composed by the above described bilingual collection, which heavily relies on administrative documents and descriptions of TV programmes. For development and testing we separated a subset of the administrative corpus for the *in-domain* evaluation and selected a fresh collection of news documents for the *out-of-domain* study, totalling three sets: (*i*) ADMIN devel and (*ii*) ADMIN test, extracted from the administrative documents and containing 1,500 segments each with a single reference; and (*iii*) NEWS test, containing 1,000 sentences collected from Spanish newspapers with two human references. Additionally, we collected a 21-million-word monolingual corpus, which together with the Basque side of the parallel bilingual collection, built up a 28-million-word corpus to train the Basque language model. This monolingual corpus is also heterogeneous and includes text from two new sources: a Basque corpus of Science and Technology (*ZT* corpus) and articles published by the Berria newspaper (*Berria* corpus).

System development. The statistical systems were developed as explained is Section 3.1. The quality scores presented in the tables of results are all obtained after series of 11 parameter tuning executions with MERT. Due to the randomness of the starting point in the

[4] http://revista.consumer.es

[5] http://www.elhuyar.org/hizkuntza-zerbitzuak/EN/Services

[6] http://statmt.org/europarl/

optimisation process, results vary slightly from one execution to another. In order to increase the robustness of the results we ran 11 of them and selected the weights from the run that got the median of BLEU scores. BLEU is chosen because it is the metric used by MERT to optimise parameters and it is computed on regular text without any segmentation at morpheme level. This choice has been done in order to avoid the propagation of the segmentation errors into the evaluation. The same process is repeated in the hybrid system. The MERT optimisation for the different sets of features that are used in the monotone decoding is also done 11 times. We also allow this final decoder to use phrases of any length. In this way, both small chunks and the complete translation of a sentence by an individual system can be chosen for the final translation.

Finally, we optimised our systems using MERT and MIRA [10] to study the dependence on the optimisation method. Although the concrete scores obtained with the two algorithms slightly differ, they do not represent any change in the conclusions one can draw. Therefore, in order to ease the reading of the paper and especially the tables of results, we only present the evaluation of the systems developed using MERT.

4.1 Results of the hybrid system

Table 2 shows the comparative results of the three individual systems (Matxin, SMT_b, SMT_g) and the hybrid architecture described in Section 3 (SMatxinT). Results are presented on the two test sets, namely, the in-domain corpus ADMIN and the out-of-domain corpus NEWS. Several automatic evaluation measures are provided, which will be used throughout all the experimental automatic evaluation: WER [36], PER [49], TER [46], BLEU [42], NIST [12], GTM [35] (concretely GTM-2, with the parameter associated to long matches $e = 2$), Meteor [5] (concretely MTR-st, i.e. using a combination of exact matching and stem matching), Rouge [32] (concretely RG-S*, i.e. a variant with skip bigrams without max-gap-length) and ULC, which is a normalised linear combination of all the previous measures [24]. All measures have been calculated with the ASIYA toolkit[7] for MT evaluation [25].

SMatxinT is evaluated with a monotone decoding (*m*), as presented in Section 3, and also allowing reordering (*r*) using MOSES distortion. The purpose of allowing this reordering is to check the assumption that the order given by Matxin is adequate for the translation. For the sake of comparison, we also include the results of Google Translate[8]. Finally, to have an idea of the quality upper bound of the hybrid architecture resulting from the three individual systems, we calculated an *oracle* system by selecting the best achievable translations with the SMatxinT translation models. To do so, we ran SMatxinT and calculated *n*-best translation lists ($n = 10,000$) for every segment[9], from which the best BLEU-performing translations were selected as the output for the Oracle. As with hybrid models, oracles were calculated with and without reordering (*m* and *r*, respectively).

Several conclusions can be drawn from Table 2. The most relevant aspect is that SMatxinT outperforms all individual MT systems. The improvement is not large, but is consistent along all evaluation metrics considered. Per test corpora, we observe that SMatxinT quality improvement with respect to the individual systems is larger in the NEWS corpus, which corresponds to the out-of-domain test set. In the ADMIN in-domain corpus we observe a

[7] http://nlp.lsi.upc.edu/asiya/

[8] http://translate.google.com/, translations obtained on the 29th of April, 2013.

[9] Larger *n*-best lists did not improve significantly the BLEU score.

	WER	PER	TER	BLEU	NIST	GTM-2	MTR-st	RG-S*	ULC
					ADMIN corpus				
Matxin	84.66	63.01	83.56	7.47	3.81	18.45	14.52	10.76	28.12
SMT_b	75.97	49.80	70.48	16.62	5.63	25.31	21.20	22.93	51.78
SMT_g	77.68	50.22	71.73	15.23	5.49	24.62	20.98	23.31	50.10
SMatxinT(*m*)	75.07	48.53	69.44	17.32	5.72	25.90	21.83	24.68	53.99
SMatxinT(*r*)	74.54	48.66	68.99	17.47	5.77	26.05	21.82	24.88	54.38
Google	81.77	59.54	78.37	8.84	4.18	19.80	15.63	12.94	33.15
Oracle(*m*)	66.40	40.64	58.93	23.91	7.07	31.08	26.75	33.48	71.46
Oracle(*r*)	65.44	41.03	58.21	25.81	7.15	32.22	26.94	34.25	73.49
					NEWS corpus				
Matxin	76.04	53.18	73.57	14.29	6.05	22.62	20.27	15.90	41.02
SMT_b	77.52	51.26	68.49	15.93	6.45	23.64	21.59	16.39	44.56
SMT_g	78.71	52.86	68.93	15.21	6.43	23.44	21.84	17.66	44.20
SMatxinT(*m*)	76.09	50.29	66.70	17.14	6.72	24.58	22.52	18.66	48.00
SMatxinT(*r*)	76.03	50.12	66.63	17.18	6.73	24.59	22.51	18.52	48.03
Google	78.29	56.47	70.66	13.15	5.69	21.62	18.51	13.32	37.11
Oracle(*m*)	65.50	40.85	53.61	26.52	8.34	30.47	28.29	28.59	69.61
Oracle(*r*)	64.76	41.57	53.42	29.16	8.40	31.86	28.33	28.71	71.44

Table 2 Automatic evaluation of the three individual systems (Matxin, SMT_b, SMT_g) and the hybrid architecture (SMatxinT) for both test corpora (in-domain ADMIN and out-of-domain NEWS). Several automatic evaluation measures are provided. SMatxinT is evaluated with a monotone decoding (*m*) and also allowing reordering (*r*). For comparison, we include the results of Google Translate and the Oracle system, resulting from selecting the best achievable translations by SMatxinT.

large difference between the performance of Matxin and the SMT individual systems, in favour of the latter. This is a well-known phenomenon of automatic lexical-matching evaluation metrics overestimating the quality of statistical systems in in-domain test sets [23]. Despite the huge differences (e.g., BLEU score varies from 7.47 (Matxin) to 16.62 (SMT_b)), the hybrid system, working on the basis of Matxin analysis and word order, is able to take advantage of the combination and consistently improve results over the best individual SMT system.

Note that apparently the SMT systems do not experience a drop in translation quality when going from the in-domain corpus to the out-of-domain one, but this is only an effect of having two references in the NEWS corpus, compared to one in ADMIN. In fact, a significant quality drop exists both in the SMT and SMatxinT systems. Matxin, is generally keeping the quality of the translation in the out-of-domain dataset (in practice doubling the absolute BLEU score due to the larger number of references).

Compared to the Oracle scores, we see that there is still a large room for improvement. For instance, according to the BLEU score on the NEWS corpus, SMatxinT is only recovering 1.21 BLEU points from the 10.59 gap between the best individual system and the Oracle. Regarding the monotone versus reordering-based decoding of SMatxinT, the differences in performance are very small. This point backs our selection of the RBMT system as the basis for fixing word order in the translation, allowing a simpler and faster decoding for the hybrid model. The study on the oracles is further supporting this selection, since the quality upper bounds achieved by Oracle (*r*) are not dramatically better than the ones with fixed order (*m*).

Focusing on the NEWS corpus, in which all individual systems are compared under fairer conditions by the automatic evaluation metrics, we see that all individual MT systems

System	Phrases			Tokens/Phrase		
	SMT	Matxin	Both	SMT	Matxin	Both
			ADMIN corpus			
SMatxinT(*m*)	2,587 (60.2%)	177 (4.1%)	1,538 (35.8%)	7.7	2.9	1.4
Oracle(*m*)	4,290 (58.4%)	679 (9.2%)	2,374 (32.3%)	3.4	2.6	1.2
			NEWS corpus			
SMatxinT(*m*)	2,498 (53.6%)	324 (7.0%)	1,838 (39.4%)	5.9	3.2	1.5
Oracle(*m*)	3,796 (52.0%)	1,074 (14.7%)	2,417 (33.2%)	3.1	2.7	1.3

Table 3 Number of phrase pairs and average number of tokens per phrase coming from the two individual systems ('SMT' and 'Matxin'). Chunks appearing in both systems are counted under the 'Both' column. The name of the systems corresponds to those in Table 2.

(Matxin, SMT_b, SMT_g) overcome Google Translate according to all evaluation metrics. This fact indicates that we depart from strong individual baseline systems in this study. Obviously, results of SMatxinT system significantly improve the results of Google Translate.

4.2 Further analysis of the hybrid translation output

In this section we further analyse the output of the hybrid system in two directions. First, we observe the origin of the phrase pairs used to construct the output translations. Second, we analyse the performance of the hybrid system on subsets of the test corpus according to the correctness of the syntactic analysis.

Table 3 presents the absolute counts and percentages of the origin of the phrases used to construct the output translations. It is worth noting that the meaning of "phrase" here is more general than in the individual SMT system, since they can correspond to any fragment of the input that is present in the enriched tree of the hybrid system, and their translations by either SMT or Matxin individual systems. The SMT and Matxin translations of the fragments can be identical in some cases. When these fragments are used in the translation we count them under the 'Both' column. The table also contains information on the average length per phrase. Results in both test corpora are presented for SMatxinT with monotone decoding and its corresponding Oracle.

Several conclusions can be drawn from Table 3. First, it seems clear that SMatxinT strongly favours the usage of SMT-source phrase pairs over the pairs coming from Matxin alone (both in number and percentage). This may be caused by the fact that the individual SMT translator performs better than Matxin in our test sets (especially on the in-domain ADMIN corpus), but also because the decoder for SMatxinT is the same in nature as that of the individual SMT system, with only a few extra features to account for the origin of phrase pairs and their consensus. Second, it is worth mentioning that SMatxinT tends to translate by using very long phrase pairs from SMT compared to the fragments coming from RBMT (e.g., 7.7 vs. 2.9 tokens per phrase in ADMIN). In the out-of-domain NEWS corpus both previous effects are slightly diminished. Third, the Oracle shows that better translations are possible by using more phrase pairs from Matxin (e.g., in NEWS, the absolute number of fragments is multiplied by 3.8 and the percentage is more than doubled). These solutions imply the use of more, and consequently shorter, phrase pairs to construct translations. In particular, those coming from SMT are shortened from 7.7 to 3.4 tokens per phrase on average. Again, this effect is slightly diminished in the NEWS corpus. Finally, we want to

#UTrees	Δ(-WER)	Δ(-PER)	Δ(-TER)	Δ(BLEU)	Δ(NIST)	Δ(GTM-2)	Δ(MTR-st)	Δ(RG-S*)	Δ(ULC)
Any	0.90	1.27	1.04	0.70	0.09	0.59	0.63	1.63	2.79
0	2.64	1.44	2.66	2.00	0.19	1.77	1.21	3.19	5.39
1,2	1.07	1.39	1.39	0.75	0.11	0.79	0.73	1.28	2.84
>2	0.29	0.68	0.38	0.35	0.04	0.14	0.42	0.56	1.29

Table 4 Variations in the quality of the translations defined as the difference between the score of the hybrid system (without reordering) and the best individual system. These Δ differences are calculated for every evaluation metric used in previous tables. Each row restricts results to a subset of sentences with a certain number of unrooted subtrees ('#Utrees'). '0' indicates a successfully analysed sentence, '*n*' indicates an incomplete parsing with output a forest with *n* unrooted subtrees. 'Any' corresponds to the results on the complete test set already provided in Table 2.

note that the same conclusions can be drawn from the SMatxinT versions with reordering (SMatxinT(*r*) and Oracle(*r*)). Numbers are not included for brevity.

Our hybrid system is tightly tied to the syntactic analysis of Matxin when deciding which fragments can play a role in the translation and which is their linear order. Departing from erroneous parse trees should negatively affect SMatxinT performance substantially. In the rest of this subsection we will break down the SMatxinT results into separate subsets of the test set according to the quality of the input parsing. Performing a manual evaluation of the syntactic quality of all trees in the tests sets would have been too labour intensive. As a rough estimation, we took a simple heuristic rule to classify parse trees into quality classes. When the FreeLing syntactic analyser is unable to provide a unique complete tree it outputs a forest with several unrooted subtrees. The more of these unmatched subtrees in the analysis the worse we can assume the quality is.

Table 4 shows the above described analysis of results for the ADMIN corpus[10]. Three quality groups are defined depending on the number of unmatched syntactic subtrees: '0' (i.e., successfully analysed sentences), '1 or 2' and 'more than 2'. These categories result in example subsets of similar size. For comparison, we include the row 'Any', corresponding to the results on the complete test set. The numbers presented in the table are calculated, for every evaluation measure, as the difference in score between SMatxinT and the best of the individual translators (either SMT or Matxin). As expected, we clearly observe that most of the improvement resides in the sentences where the parser successfully produces a single parse tree, and that this gain decreases as the number of unmatched subtrees increases. As a consequence, improving parsing quality is very important to improve SMatxinT performance. In the next section we explore the addition of an alternative parse tree to increase robustness of the hybrid system.

4.3 Multiple syntactic trees

In Section 4.2, we have seen that the higher the quality of the syntactic parsing is, the higher the quality of the hybrid translation. As described so far, SMatxinT makes use of a single parse tree to produce the final grammatical structure of the sentence. Therefore, one of the weaknesses of the system is that in case of a parsing failure the hybrid translation might be strongly limited and, most probably, the chosen translation will be that of the statistical system (recall that the full Matxin, SMT_b and SMT_g translations are always available for the final monotone decoding). In this section we introduce the structure and translation options

[10] Results on the out-of-domain NEWS corpus are similar and not included for brevity.

given by a second alternative parse tree, as a way to test the importance of having parsing diversity to increase the robustness of the hybrid system.

The parser used by SMatxinT is FreeLing, a grammar-based dependency parser. FreeLing cannot provide the *n*-best parse trees for a given input sentence, so it cannot be used to generate diversity. The second parser we introduced is MaltParser [37]. This is also a dependency parser but, contrary to FreeLing, it is machine-learning-based and allows for confidence scores to rank the parses. Unfortunately, it cannot provide *n*-best lists either, so its contribution was also reduced to a single extra parse tree. The first step for the integration is to obtain the mixed dependency/constituency parses Matxin needs. To do that, the parses provided by MaltParser are augmented with constituency information derived from the dependencies. A small number of rules has been defined to determinate if a node and its dependant are part of the same constituent or not. In order to avoid further modifications we retrained the Malt models based on the same morphosyntactic tagset. Even so, each parser follows different guidelines to parse some complex syntactic structures, such as coordination and subordinate clauses. Therefore, some transfer rules needed to be modified to deal with the most important differences introduced by Malt. Due to the high cost of this adaptation (in human labour), we followed a minimalistic approach by investing the minimum amount of hours to produce a compatible transfer module compatible with Malt. Given that, Malt-based Matxin is a worse system than the original Matxin, but good enough to serve the purpose of this *proof of concept* experiment.

Once Malt parser is integrated into Matxin, the source can be translated with the two variants independently. A simple modification to Moses' MERT script allows to optimise the weights of the log-linear model with respect to the two translations simultaneously[11]. In this way, the final translations with Moses' monotone decoding are comparable for the two systems and the new hybrid system simply chooses the best one according to its score. This is the simplest approach, in which the two variants of Matxin (FreeLing- and Malt-based) are used separately to produce two SMatxinT translations, which are then ranked to select the most probable one.

Table 5 shows the results of this experiment. The first two rows in each corpus section contain the results of Matxin using either FreeLing ($\text{Matxin}_{(F)}$) or Malt ($\text{Matxin}_{(M)}$) parsers. As expected, $\text{Matxin}_{(F)}$ systematically outperforms $\text{Matxin}_{(M)}$ according to all evaluation metrics in both corpora.[12] More interestingly, using the combination of both parsers allows $\text{SMatxinT}_{(F+M)}$ to produce systematically better results than the best single-parser counterpart $\text{SMatxinT}_{(F)}$, especially in the NEWS corpus. In this case, the quality raises by 0.51 BLEU points and 1.17 points of the ULC averaged metric. This improvement is statistically significant according to paired bootstrap resampling [29] and the p-value of the two translations coming from the system is only of 0.008. The differences are not large but they are consistent across metrics and corpora. This is remarkable in our opinion given the shallow integration of Malt into Matxin, which produced a fairly weak translation system $\text{Matxin}_{(M)}$. Compared to the best rule-based system, the improvement is of course larger. $\text{SMatxinT}_{(F+M)}$ outperforms $\text{Matxin}_{(F)}$ by 10.09 and 3.36 BLEU points in the in-domain and out-of-domain corpora, respectively.

[11] In every run of MERT the development set is translated by a system and this generates an n-best list of translations. In our case we have two systems that generate two n-best lists. These two lists are joined and sorted at every run so that the minimisation process proceeds as usual but with the translations of both systems.

[12] The same happens with the hybrids: $\text{SMatxinT}_{(F)}$ is consistently better than the hybrid version constructed with $\text{Matxin}_{(M)}$ (results not included in the table for brevity and clarity reasons).

	WER	PER	TER	BLEU	NIST	GTM-2	MTR-st	RG-S*	ULC
					ADMIN corpus				
Matxin$_{(F)}$	84.66	63.01	83.56	7.47	3.81	18.45	14.52	10.76	28.47
Matxin$_{(M)}$	85.08	63.62	84.21	6.86	3.66	17.96	13.88	10.18	26.93
SMatxinT$_{(F)}$	75.07	48.53	69.44	17.32	5.72	25.90	21.83	24.68	54.39
SMatxinT$_{(F+M)}$	74.94	48.37	69.25	17.56	5.76	26.00	21.93	24.77	54.77
Oracle$_{(F)}$	66.40	40.64	58.93	23.91	7.07	31.08	26.75	33.48	71.89
Oracle$_{(F+M)}$	64.95	39.96	57.58	25.01	7.18	31.75	27.25	34.55	74.06
					NEWS corpus				
Matxin$_{(F)}$	76.04	53.18	73.57	14.29	6.05	22.62	20.27	15.90	39.97
Matxin$_{(M)}$	76.16	54.41	74.07	13.80	5.88	22.24	19.66	14.98	38.32
SMatxinT$_{(F)}$	76.09	50.29	66.70	17.14	6.72	24.58	22.52	18.66	46.92
SMatxinT$_{(F+M)}$	75.10	49.65	65.80	17.65	6.78	24.80	22.71	19.16	48.09
Oracle$_{(F)}$	65.73	40.91	53.83	26.48	8.32	30.42	28.29	28.62	68.41
Oracle$_{(F+M)}$	64.01	40.22	52.11	28.55	8.52	31.49	28.91	29.66	71.46

Table 5 Comparative automatic evaluation of: (*i*) Matxin using FreeLing and Malt parsers (Matxin$_{(F)}$ and Matxin$_{(M)}$, respectively), (*ii*) the hybrid architecture using either FreeLing (SMatxinT$_{(F)}$) or the combination of both parsers (SMatxinT$_{(F+M)}$). The last two rows show the *Oracle* systems for the last two cases of the hybrid system.

Finally, Table 5 also shows the evaluation of the Oracle system over the hybrid translators. In a coherent way with respect to the previous findings, we observe a larger room for improvement with the system that makes use of both parsers (Oracle$_{(F+M)}$). In summary, in this experiment we have seen that, even though the translation models based on Malt are clearly weaker than the ones using FreeLing, introducing parsing diversity through the Malt models produces more translation alternatives leading to higher translation quality by the hybrid systems. In an ideal case, one would like to incorporate an arbitrarily large number of parse trees in the hybrid system. This could be done by using more parsers or, even easier, by producing *n*-best lists of parse trees, by using a statistical parser with this capability.

4.4 Additional features for the hybrid decoder

As a final experiment, we made an attempt to include new features in the statistical decoder other than the basic ones considered in the previous sections. The motivation is to try to overcome the excessive SMT-prone bias introduced by the statistical decoder of SMatxinT, by using some linguistically motivated features. This behaviour was observed and discussed in Subsection 4.2 (Table 3). The new features considered are divided in two categories: those related to lexical probabilities and those related to the syntactic properties of the phrases.

Lexical probability features. Two feature types are defined in this category.

1. *Corpus Lexical Probabilities* (both directions): This feature is based on the lexical probability commonly used in SMT (IBM-1 model). But, since morpheme-based SMT (SMT$_g$) and Matxin systems are able to generate alignments not seen in the training corpus, a modification was needed to treat unknown alignments. Concretely, those alignments that were not present in the corpus are just ignored. Those words for which all their alignments are ignored receive the probability corresponding to the NULL alignment.

	WER	PER	TER	BLEU	NIST	GTM-2	MTR-st	RG-S*	ULC
					ADMIN corpus				
SMatxinT	74.94	48.37	69.25	17.56	5.76	26.00	21.93	24.77	62.36
SMatxinT++	73.81	48.62	68.25	17.54	5.80	26.16	21.90	24.86	62.84
					NEWS corpus				
SMatxinT	75.10	49.65	65.80	17.65	6.78	24.80	22.71	19.16	62.43
SMatxinT++	74.73	49.30	65.40	17.53	6.81	24.80	22.70	19.20	62.64

Table 6 Comparison of SMatxinT to the corresponding version enriched with lexical and syntactic features (SMatxinT++) for the two test sets under study.

Unknown words that would not be present in the IBM-1 probability table use a default NULL alignment probability (10^{-10}).

2. *Dictionary Lexical Probabilities* (both directions): These are also translation probabilities at word level, but instead of estimating them on the training corpus, they are extracted from the Matxin bilingual dictionary. This is not a probabilistic dictionary, so to estimate actual probabilities we used heuristic rules depending on the number of different word senses and their order in the dictionary entry. The same mechanism used for the unknown alignments in Corpus Lexical Probabilities is used here.

Syntactic features. Three feature types are defined in this category. The source's syntactic information is based on Matxin's tree structure, while the target is obtained using the in-house developed shallow parser [2].

1. *Syntactic Correctness*: Binary feature that indicates whether the candidate translation forms a correct sequence of linguistic phrases or not, according to the target shallow parser. Given that the candidate translations correspond to syntactic constituents in the source, it is expected that the translations form a syntactically correct chunk as well.
2. *Source-Target Chunk Proportion*: Number of chunks identified in the source segment divided by the number of chunks in the target translation candidate. This feature aims to capture the syntactic differences between source and candidate translations. Although a one-to-one correspondence cannot be expected, it is also true that a large difference in the number of syntactic chunks identified in both segments would probably indicate problems in the translation.
3. *Phrase Type & Source*: According to our previous experience, each individual system translates better different types of phrases. For example, Matxin usually gets better verb-chain translations, while SMT better translates the noun-phrases due to its better lexical selection. In order to allow the hybrid decoder to distinguish between them, one feature for each phrase type (noun-phrase, verb-chain, etc.) is added that gets a positive value (+1, which is converted into e^{+1}) if the translation is generated by the SMT system or negative (-1, converted into e^{-1}) if it is generated by Matxin.

Table 6 shows the results obtained by the hybrid system enriched with lexical and syntactic features (SMatxinT++) compared to its basic version (SMatxinT). The monotone decoding for SMatxinT++ incorporates a total of 22 features, compared to the 7 features of the basic version. It can be observed that the inclusion of these features do not significantly vary the performance of the system. SMatxinT++ obtains an increment of 0.48 ULC points in the ADMIN corpus and 0.21 in the NEWS corpus. However, this improvement mainly

comes from WER. Metrics such as BLEU or METEOR do not really discriminate between the two translation systems.

Although we are using a set of linguistically motivated features, the behaviour of the hybrid system does not vary much with respect to the basic version. We also observed that the number of chunks coming from Matxin in the final translation does not increase significantly, and keeps being much smaller compared to the proportion of chunks coming from SMT alone (see Table 3). It remains to be studied whether the problem resides in the feature design, in their implementation or instead, is more structurally tied to the decoding approach we used. We plan to investigate all these aspects in the future as well as considering the possibility of introducing different decoding strategies (e.g., in the line of using confusion networks or Minimum Bayes-Risk decoding) closer to the system output combination approaches.

5 Manual Evaluation

In order to contrast the results obtained with the automatic evaluation exercise we conducted two human evaluations. First, we present a subjective judgment of translation quality by means of pairwise comparison of the two individual systems (SMT_b and Matxin) and the hybrid one (SMatxinT).[13] Second, we discuss a task-oriented evaluation by means of Human-targeted Translation Error Rate (HTER) [46], where each automatic translation is compared against a reference created by post-editing the given sentence.

The same test-set has been used in both evaluations, we selected fifty sentences of each test corpora, for a total of one hundred samples. These sentences were selected randomly but from a pre-selected subset of sentences satisfying the two following conditions: (*i*) sentence length is between 6 and 30 words and (*ii*) at least one of the individual systems achieves a segment-level BLEU score above the median (for that system and evaluation corpus). With the length constraint we wanted to discard too easy and too difficult sentences. The requirement on minimum translation quality was set to avoid considering cases in which the two individual systems produced very bad translations, and thus, to concentrate on examples in which the hybrid system has real potential for combining the output of both individual systems. This is especially critical, given the generally low absolute BLEU scores obtained by the individual systems when translating into Basque. Manual evaluation is costly and we ultimately tried not to waste effort in assessing useless examples.

5.1 Pairwise comparison

Six Basque native speakers were asked to evaluate 100 pairs of alternative translations each (50 from the in-domain test set and 50 from the out-domain). Since there are 3 systems to compare over 100 examples this makes a total of 300 pairwise comparisons. Our evaluators decided on 600 comparisons. Therefore, each pair was evaluated by two different evaluators, allowing for the calculation of agreement rates between them. The distribution of the examples among evaluators was done randomly at the level of pairwise comparisons.

For each pairwise comparison, the evaluator was presented with the source sentence and the automatic translations of two of the systems. The goal of the evaluator was to assess which one of the two alternative translations is better, allowing to decide on a tie when

[13] For the hybrid system we used the $\text{SMatxinT}_{(F+M)}$ variant presented in Table 5.

	agreement	weak disagreement	disagreement
# cases	215	70	15
percentage	71.7%	23.3%	5.0%

Table 7 Agreement between evaluators in the manual evaluation. Weak disagreement is considered when one of the evaluators preferred one system while the other considered both of the same quality.

	SMT_b vs. Matxin	SMT_b vs. SMatxinT	Matxin vs. SMatxinT
Best system		*ADMIN* corpus	
System1	35	16	41
Same quality	16	55	12
System2	49	29	47
Best system		*NEWS* corpus	
System1	10	14	53
Same quality	11	42	22
System2	79	44	25

Table 8 Pairwise manual evaluation of the individual and hybrid systems (SMT_b, Matxin and SMatxinT) for both test corpora (in-domain ADMIN and out-of-domain NEWS). All human assessments are considered to be independent.

both translations are of the same quality. In the cases in which the evaluator expressed a preference for one of the systems, he was asked to explain why it is better by selecting one or more of the following quality aspects: *lexical selection*, translation *adequacy*, *syntactic agreement* (e.g., subject-verb agreement), *morphology* (e.g., missing or incorrect suffixes), *word order*, or *verb formation* (accounting for any error that may happen in the verb phrase). If non of them was applicable, the evaluator was allowed to chose a generic 'other' category and explain the situation in a open text box.

Table 7 shows the agreement rates obtained in this human evaluation exercise (in absolute number of cases and also percentages). In total, there are 300 assessments. In 71.7% of them, the two evaluators agreed on the assessment (they showed preference for the same system or considered that the two translations were indistinguishable in quality). On the other side, in 5% of the assessments they disagreed, that is, each evaluator preferred a different system. In the rest of cases, 23.3%, one evaluator expressed a preference for one system while the other considered that both translations were comparable. We refer to this situation as *weak disagreement*, since they are not reflecting a truly contradictory decision, and they should by consider differently. Assessing the quality of automatic translations is a difficult task even for humans. Its definition is difficult and the subjectivity of evaluators plays an important role. Overall, the agreement rates obtained cannot be considered bad, since only 5% of the cases correspond to real disagreements between evaluators.

Tables 8 and 9 show the results of the manual evaluation for each system pair and corpus, essentially counting the number of times each system is preferred over the other and the times the quality of both is indistinguishable. Table 8 presents results considering all quality assessments, i.e., for every translation pair we do not aggregate the quality assessments by the two evaluators, but we consider them as independent counts, even if they disagree. In Table 9 the same results are presented but aggregating the two assessments for each translation pair, i.e., each example is counted only once. We distinguished 6 cases. When the two evaluators agree, the outcome can be a win for system 1, system 2, or tie. When the two evaluators weakly disagree, we still consider that the outcome is favourable to either

	SMT_b vs. Matxin	SMT_b vs. SMatxinT	Matxin vs. SMatxinT
Best system		*ADMIN* corpus	
System1 (agreement)	12	5	15
System1 (weak)	8	6	3
Same quality	1	21	2
System2 (agreement)	20	11	17
System2 (weak)	6	7	5
Disagreement	3	0	8
Best system		*NEWS* corpus	
System1 (agreement)	4	5	20
System1 (weak)	0	4	11
Same quality	0	16	4
System2 (agreement)	33	19	10
System2 (weak)	11	6	3
Disagreement	2	0	2

Table 9 Pairwise manual evaluation of the individual and hybrid systems (SMT_b, Matxin and SMatxinT) for both test corpora (in-domain ADMIN and out-of-domain NEWS). Multiple human assessments are aggregated at the level of translation pair.

one or the other system (noted *'weak'* in the table). Finally, there is the situation in which the two evaluators disagree, which we consider independently of all others.

As one can see from the tables, the manual evaluation partly contradicts the previous evaluation performed with automatic metrics (Section 4). Unlike in the automatic evaluation setting, Matxin is considered to be a much better system than SMT_b both in the in-domain and out-of-domain test sets. Also SMatxinT is able to beat SMT_b in both scenarios, with a especially large difference in the out-of-domain set. When comparing Matxin with SMatxinT (third column), we observe that in the in-domain test set the differences are not large (with a slight advantage for SMatxinT), but for the out-of-domain corpus the clear winner is Matxin, thus contradicting the automatic evaluation, which situated the hybrid system on top in every scenario.

By comparing the human and the automatic evaluations, it is clear that: (*i*) Matxin's quality was underestimated by all the automatic measures, and (*ii*) the severe quality drop of SMT_b on the out-of-domain test was not properly captured by the automatic measures. The use of these wrongly biased automatic measures at development and optimisation stages made our hybrid system to prefer the partial translations from SMT_b over the translation choices offered by Matxin. As a result, the performance of the hybrid system was clearly improved according to the automatic metrics, but the actual performance was hurt according to the human assessments. One conclusion from this study is that having automatic metrics that correlate well with the human perception of translation quality is paramount at development stages to obtain reliable hybrid systems.

Table 10 provides a summary of the features which led the human evaluators to prefer one translation over the other at every pairwise comparison. This is important qualitative information to identify the strong and weak points of every system. The columns in the table contain the number of times human evaluators selected each of the quality features for a particular system in winning situations compared to an alternative system.

In the in-domain corpus (ADMIN), Matxin achieved better results than SMT_b in all features except *lexical selection*, which, taking into account the differences in evaluation, can be considered the biggest strength of SMT_b. Most of the differences in favour of Matxin are

	SMT_b	Matxin	SMT_b	SMatxinT	Matxin	SMatxinT
ADMIN corpus						
lexical selection	18	16	1	10	12	18
adequacy	4	7	4	9	10	16
agreement	1	8	3	3	5	3
morphology	12	16	3	10	14	17
order	22	32	5	13	12	9
verb	6	7	0	1	9	3
other	7	11	4	8	7	7
NEWS corpus						
lexical selection	4	24	1	15	9	19
adequacy	4	38	7	18	4	1
agreement	1	29	0	0	10	1
morphology	4	37	1	9	29	11
order	4	20	6	17	21	6
verb	0	15	3	12	7	3
other	0	0	1	1	8	1

Table 10 Aspects of translation quality which made evaluators to prefer one system over the other.

not large, the biggest ones occurring on *agreement* and *order* aspects of quality. Similarly, SMatxinT improves over SMT_b in all features except *syntactic agreement* and *verb formation*, where both systems achieve very similar results. Remarkably, SMatxinT is clearly better than SMT_b in *lexical selection*. Finally, comparing Matxin and SMatxinT we can observe that they have some complementary strong points. Matxin performs better on *verb formation*, while SMatxinT strengths lie on *lexical selection* and *translation adequacy*.

In the out-domain corpus (NEWS) both Matxin and SMatxinT are largely better than SMT_b in all quality aspects. Regarding the comparison Matxin vs. SMatxinT we observe that Matxin is generally much better (especially on *syntactic agreement* and *word order*), except for *lexical selection*, where SMatxinT is clearly preferred. Again, this fact points to lexical selection as the most important strength of SMatxinT.

Finally, in order to have a better understanding of the divergences between manual and automatic evaluation, we inspected some of the manually evaluated sentences. On the one hand, the example presented in Figure 3(*a*) shows the expected behaviour, where SMatxinT manages to properly outperform both individual systems. On the other hand, Figure 3(*b*) shows an example where Matxin's translation is preferred by humans over the other two systems, even when it achieves a worse segment-level BLEU score. Some differences in word formation —hyphen separation between acronyms and their suffixes (*eebk* vs. *ebb-k*), and the use of a periphrastic verb instead of its synthetic form (*esaten du* vs. *dio*)— hurt the automatic evaluation of Matxin output. Compared to Matxin, SMatxinT contains more severe errors from the human perspective, including uninflected or badly inflected words. Nonetheless, the automatic evaluation metrics were unable to capture this.

5.2 HTER evaluation

In addition to the subjective pairwise comparison, we also conducted a task-oriented evaluation based on Human-targeted Translation Error Rate (HTER [46]). This metric is a semi-automatic measure in which humans do not score translations directly, but rather generate a new reference translation by post-editing the MT output. The post-edited translation might

Example (*a*)

Source:	una enorme mancha de hidrocarburo aparece en una playa de vila-seca
Ref. 1:	hidrokarburo orban erraldoia agertu da vila-secako hondartza batean
Ref. 2:	hidrokarburo-orban handia agertu da vila-secako hondartza batean
SMT_b:	orban handi bat , hidrocarburo hondartza batean agertu da vila-seca
Matxin:	hidrokarburozko orban oso handi bat vila-secaren hondartza batean agertzen da
SMatxinT:	hidrokarburo orban handi bat agertu da vila-secako hondartza batean

Example (*b*)

Source:	gonzalez dice que el ebb decidió que el pnv de gipuzkoa enmendase el impuesto de sociedades
Ref. 1:	gonzalezek dio ebbk erabaki zuela gipuzkoako eajk sozietateen gaineko zerga ordaintzea
Ref. 2:	gonzalezek dio ebbren erabakia izan zela gipuzkoako eajk sozietateen gaineko zerga aldatzea
SMT_b:	gonzalez esan ebb erabaki zuen gipuzkoako eajk duen enmendase sozietateen gaineko zerga
Matxin:	gonzalezek esaten du ebb-k erabaki zuela gipuzkoaren eaj-k sozietateen gaineko zerga zuzen zezala
SMatxinT:	gonzalezek esan ebb erabaki zuen gipuzkoako eajko sozietateen gaineko zerga zuzen zezala

Fig. 3 Examples extracted from the NEWS corpus with the source sentence, two human references and the three automatic translations output by SMT_b, Matxin and SMatxinT.

	ADMIN	NEWS
Matxin	47.17	42.69
SMT_b	37.32	51.52
SMatxinT	36.86	44.56

Table 11 HTER scores of the individual and hybrid systems (SMT_b, Matxin, and SMatxinT) for both test corpora (in-domain ADMIN and out-of-domain NEWS). The lower the scores the better the quality.

be closer to the MT output, but should pair the fluency and meaning of the original reference. This new targeted reference is then used as the reference translation when scoring the MT output using Translation Edit Rate (TER). This metric is inversely correlated with quality. The lower the score the shorter the distance to the reference, which indicates higher quality.

The corpora and the systems involved in this evaluation were the same set of 100 sentences used in the previous pairwise evaluation, 50 from each corpora (ADMIN and NEWS) translated with the three systems (Matxin, SMT_b, and SMatxinT). These 300 translations were post-edited by three different professional translators. To avoid post-editor bias the sentences were uniformly divided among post-editors. Each post-editor corrected one third of the translations of every system. None of them corrected the same source sentence twice.

Table 11 shows the HTER results obtained for each system, in the two different corpora. The out-of-domain scores align well with the results obtained in the pairwise comparison, that is, Matxin is the preferred system, followed by SMatxinT, and SMT_b, which clearly obtains the worst scores. Interestingly enough, the results obtained in the in-domain corpus show a slightly different pattern. The scores by Matxin are worse than those obtained in the pairwise comparison, and the system obtains the worst scores among the three evaluated systems, by a large margin. This differs from the pairwise comparison results, but matches the results from the automatic evaluation. Consistenly with all previous evaluation, the hybrid system, SMatxinT, is the one that obtains the best in-domain scores.

We further analysed the source of the discrepancy between the two manual evaluations (pairwise and HTER–based) with Matxin in the in-domain corpus by manually inspecting several cases. We first saw that for the sentences where the evaluators preferred the Matxin translation over the SMT translation, the difference in HTER score between the two sys-

tems is small and can express a preference for either. In contrast, when the SMT translation is preferred, the HTER score also shows a clear preference for this system. In our understanding, the discrepancies in the evaluation occur due to two main reasons: (*i*) there is not a direct relation between the importance of the errors, as perceived by human evaluators, and the number of edits needed to correct them. In general Matxin fixes some errors by the SMT that are judged very important by humans and which may override, in terms of overall quality, the effect of other minor mistakes committed by Matxin on the same sentences (and which SMT might not commit). However, these cases lead to a very similar number of edits in TER. (*ii*) since HTER is based on words, it is unable to detect some improvements that are identified by humans in the pairwise comparison (e.g., correctly selected but badly inflected lemmas).

6 Conclusions

In this article we described SMatxinT, a hybrid machine translation architecture which combines rule-based machine translation and phrase-based statistical machine translation individual systems. Our approach builds on two main assumptions: (*i*) the RBMT is generally able to produce grammatically better translations, so we used its analysis and transfer modules to produce the backbone of the translation; (*ii*) SMT-based local translation alternatives and statistical decoding should improve lexical selection and fluency of the final translation. Additionally, the SMT component works as a back off for the cases in which the RBMT fails at producing good translations due to parsing errors. For that, longer SMT translations even corresponding to the full source sentence are made available to the hybrid decoder.

We evaluated our system on two different corpora for a pair of distant languages, Spanish and Basque, being the latter an agglutinative language with a very rich morphology. The hybrid system outperformed the individual translation systems on both benchmark corpora and across a variety of automatic evaluation measures for assessing translation quality. Results also confirmed that working with the structure proposed by the RBMT system was a good choice. Including reordering in the hybrid decoder provided only minuscule improvements over the monotone version. Even the oracle decoding was not much better when reordering was considered. We also verified that, as expected, the improvement of the hybrid system mainly takes place on syntactically well parsed sentences. As a result of this output analysis, we explored two additional modifications over the basic architecture. First, we worked on providing more robustness against parsing errors by incorporating another statistical parser and performing the optimisation of the hybrid system jointly on the output of both parsers. Results confirmed the improvement of the hybrid system when increasing parsing diversity. Second, some linguistically-motivated features for the hybrid decoder were also explored in order to compensate the hybrid decoder for its preference to select SMT-based longer translations. Unfortunately, the results on the usefulness of such features were inconclusive.

Finally, we also carried out two kinds of human evaluation (a subjective pairwise comparison and an HTER–based evaluation) on a subset of the test examples. Their results partly contradicted the automatic evaluation. Although in the in-domain corpus humans also prefer the translations from the hybrid system, in the out-of-domain test corpus, they preferred the translations from the RBMT system over the hybrid and statistical translations (in this order). The main reason is that the automatic evaluation metrics largely overestimate the quality of the SMT system (compared to RBMT) according to the human assessments [7]. Of course, this is also reflected in the behaviour of the hybrid system, showing a strong preference towards SMT translations. On the out-of-domain corpus, the large drop in per-

formance of the SMT individual system exacerbates this problem. Nonetheless, some interesting qualitative conclusions can be extracted from the manual evaluation regarding the strongest and weakest aspects of every system in the comparison. Finally, the comparison between the two manual evaluation schemes lead also to interesting conclusions, pointing out some limitations of the HTER metric.

This work leaves two important open issues, which certainly deserve further research. First, we should explore more thoroughly the usage of additional features for the hybrid decoding. Oracle-based evaluations told us that there is still a large room for improvement in the space of solutions explored by the hybrid system, mainly in the direction of combining smaller and less SMT-centred translation units. The second aspect refers to the evaluation of translation quality. We measured and optimised our hybrid system directly using automatic evaluation metrics, which we managed to improve in both test corpora. Unfortunately, these measures were not well aligned with human assessments. Accurately modelling the human perception of translation quality with automatic measures is still an open problem in the MT community. Having efficient evaluation measures well-correlated with humans is fundamental in order to guide the optimisation of the hybrid architecture and avoid blind system development.

Improving the Spanish–Basque translation system is another line that will receive our attention in the near future. On the one hand, we will use a hierarchical system, such as Moses-chart, instead of the plain phrase-based SMT system to build SMatxinT. On the other hand, we would like to exploit more parse trees in the form of *n*-best lists and explore the usage of stronger components from the *Itzulzailea*[14] translation service. This is a RBMT system for Spanish-Basque working on the grounds of Lucy technology. Finally, it is worth mentioning that we are also interested in applying this hybrid approach to other language pairs and translation scenarios.

Acknowledgements The authors are grateful to the anonymous reviewers of the initial version of this article for their insightful and detailed comments. They contributed significantly to improve this final version.

This work has been partially funded by the Spanish Ministry of Science and Innovation (OpenMT-2 fundamental research project, TIN2009-14675-C03-01) and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 247914 (MOLTO project, FP7-ICT-2009-4-247914).

References

1. Aduriz, I., Aldezabal, I., Alegria, I., Artola, X., Ezeiza, N., Urizar, R.: EUSLEM: a Lemmatiser / Tagger for Basque. In: Proceedings of the 7th Conference of the European Association for Lexicography (EURALEX'96), pp. 17–26. Göteborg, Sweden (1996)
2. Aduriz, I., Aranzabe, M.J., Arriola, J.M., de Ilarraza, A.D., Gojenola, K., Oronoz, M., Uria, L.: A cascaded syntactic analyser for basque. In: Computational Linguistics and Intelligent Text Processing, pp. 124–134. Springer (2004)
3. Alegria, I., Díaz de Ilarraza, A., Labaka, G., Lersundi, M., Mayor, A., Sarasola, K.: An FST Grammar for Verb Chain Transfer in a Spanish-Basque MT System. In: A. Yli-Jyrä, L. Karttunen, J. Karhumäki (eds.) Proceedings of the 5th International Workshop on Finite-State Methods and Natural Language Processing (FSMNLP 2005, Helsinki, Finland), *Lecture Notes in Computer Science*, vol. 4002, pp. 87–98. Springer (2006)
4. Alegria, I., Díaz de Ilarraza, A., Labaka, G., Lersundi, M., Mayor, A., Sarasola, K.: Transfer-Based MT from Spanish into Basque: Reusability, Standardization and Open Source. Lecture Notes in Computer Science **4394**, 374–384 (2007)

[14] http://www.itzultzailea.euskadi.net/traductor/portalExterno/text.do

5. Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan (2005)
6. Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., Specia, L.: Findings of the 2013 Workshop on Statistical Machine Translation. In: Proceedings of the Eighth Workshop on Statistical Machine Translation, pp. 1–44. Association for Computational Linguistics, Sofia, Bulgaria (2013)
7. Callison-Burch, C., Osborne, M., Koehn, P.: Re-evaluating the role of bleu in machine translation research. In: 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 249–256. Association for Computational Linguistics, Trento, Italy (2006). URL http://aclweb.org/anthology-new/E/E06/E06-1032
8. Carreras, X., Chao, I., Padró, L., Padró, M.: Freeling: an Open-Source Suite of Language Analyzers. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), pp. 239–242. Lisbon, Portugal (2004)
9. Chen, Y., Eisele, A.: Hierarchical Hybrid Translation between English and German. In: V. Hansen, F. Yvon (eds.) Proceedings of the 14th Annual Conference of the European Association for Machine Translation (EAMT 2010), pp. 90–97. Saint-Raphaël, France (2010)
10. Cherry, C., Foster, G.: Batch Tuning Strategies for Statistical Machine Translation. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT '12, pp. 427–436. Montréal, Canada (2012)
11. Costa-Jussà, M.R., Farrús, M., Mariño, J.B., Fonollosa, J.A.R.: Study and comparison of rule-based and statistical catalan-spanish machine translation systems. Computing and Informatics **31**(2), 245–270 (2012)
12. Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In: Proceedings of the 2nd Internation Conference on Human Language Technology (HLT), pp. 138–145. San Diego, CA, USA (2002)
13. Dove, C., Loskutova, O., de la Fuente, R.: What's Your Pick: RbMT, SMT or Hybrid? In: Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA 2012). San Diego, CA, USA (2012)
14. Eisele, A., Federmann, C., Saint-Amand, H., Jellinghaus, M., Herrmann, T., Chen, Y.: Using Moses to Integrate Multiple Rule-Based Machine Ttranslation Engines into a Hybrid System. In: Proceedings of the Third Workshop on Statistical Machine Translation, pp. 179–182. Association for Computational Linguistics, Columbus, Ohio (2008)
15. Enache, R., España-Bonet, C., Ranta, A., Màrquez, L.: A Hybrid System for Patent Translation. In: Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT12), pp. 269–276. Trento, Italy (2012)
16. España-Bonet, C., Labaka, G., Díaz de Ilarraza, A., Màrquez, L., Sarasola, K.: Hybrid Machine Translation Guided by a Rule-Based System. In: Proceedings of the 13th Machine Translation Summit (MT-Summit), pp. 554–561. Xiamen, China (2011)
17. Federmann, C.: Results from the ML4HMT Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation. In: Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011) and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT-11), pp. 110–117. Barcelona, Spain (2011)
18. Federmann, C.: Hybrid Machine Translation Using Joint, Binarised Feature Vectors. In: Proceedings of the 20th Conference of the Association for Machine Translation in the Americas (AMTA 2012), pp. 113–118. San Diego, CA, USA (2012)
19. Federmann, C., Chen, Y., Hunsicker, S., Wang, R.: DFKI System Combination Using Syntactic Information at ML4HMT-2011. In: Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011) and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT-11), pp. 104–109. Barcelona, Spain (2011)
20. Federmann, C., Eisele, A., Chen, Y., Hunsicker, S., Xu, J., Uszkoreit, H.: Further Experiments with Shallow Hybrid MT Systems. In: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pp. 77–81. Association for Computational Linguistics, Uppsala, Sweden (2010)
21. Federmann, C., Hunsicker, S.: Stochastic Parse Tree Selection for an Existing RBMT System. In: Proceedings of the Sixth Workshop on Statistical Machine Translation, pp. 351–357. Association for Computational Linguistics, Edinburgh, Scotland (2011)
22. Federmann, C., Melero, M., Pecina, P., van Genabith, J.: Towards Optimal Choice Selection for Improved Hybrid Machine Translation. Prague Bulletin of Mathematical Linguistics **97**, 5–22 (2012)

23. Giménez, J., Màrquez, L.: Linguistic Features for Automatic Evaluation of Heterogenous MT Systems. In: Proceedings of the Second Workshop on Statistical Machine Translation, pp. 256–264. Association for Computational Linguistics, Prague, Czech Republic (2007)
24. Giménez, J., Màrquez, L.: A Smorgasbord of Features for Automatic MT Evaluation. In: Proceedings of the Third Workshop on Statistical Machine Translation, pp. 195–198. The Association for Computational Linguistics, Columbus, Ohio (2008)
25. Giménez, J., Màrquez, L.: Asiya: an Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. The Prague Bulletin of Mathematical Linguistics **94**, 77–86 (2010)
26. Habash, N., Dorr, B., Monz, C.: Symbolic-to-Statistical Hybridization: Extending Generation-Heavy Machine Translation. Machine Translation **23**, 23–63 (2009)
27. Heafield, K., Lavie, A.: Voting on N-Grams for Machine Translation System Combination. In: Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010). Denver, Colorado, USA (2010)
28. Hunsicker, S., Chen, Y., Federmann, C.: Machine Learning for Hybrid Machine Translation. In: Proceedings of the Seventh Workshop on Statistical Machine Translation, pp. 312–316. Association for Computational Linguistics, Montréal, Canada (2012)
29. Koehn, P.: Statistical Significance Tests for Machine Translation Evaluation. In: Proceedings of EMNLP 2004. Barcelona, Spain (2004)
30. 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., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pp. 177–180. Prague, Czech Republic (2007)
31. Labaka, G.: EUSMT: Incorporating Linguistic Information to SMT for a Morphologically Rich Language. Its Use in SMT-RBMT-EBMT Hybridization. Ph.D. thesis, University of the Basque Country (2010)
32. Lin, C.Y., Och, F.J.: Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04), Main Volume, pp. 605–612. Barcelona, Spain (2004)
33. Matusov, E., Ueffing, N., Ney, H.: Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. In: Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), pp. 33–40. Trento, Italy (2006)
34. Mayor, A., Alegria, I., Díaz de Ilarraza, A., Labaka, G., Lersundi, M., Sarasola, K.: *Matxin*, An Open-Source Rule-Based Machine Translation System for Basque. Machine Translation **25**(1), 53–82 (2011)
35. Melamed, I.D., Green, R., Turian, J.P.: Precision and Recall of Machine Translation. In: Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), pp. 61–63. Edmonton, Canada (2003)
36. Nießen, S., Och, F.J., Leusch, G., Ney, H.: An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In: Proceedings of the 2nd International Conference on Language Resources and Evaluation, pp. 39–45. Athens, Greece (2000)
37. Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., Marsi, E.: Maltparser: a Language-Independent System for Data-Driven Dependency Parsing. Natural Language Engineering **13**(2), 95–135 (2007)
38. Och, F.J.: Minimum Error Rate Training in Statistical Machine Translation. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), pp. 160–167. Sapporo, Japan (2003)
39. Och, F.J., Ney, H.: Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL), pp. 295–302. Philadelphia, Pennsylvania, USA (2002)
40. Oflazer, K., El-Kahlout, I.D.: Exploring Different Representation Units in English-to-Turkish Statistical Machine Translation. In: Proceedings of the Second Workshop on Statistical Machine Translation, pp. 25–32. Association for Computational Linguistics, Prague, Czech Republic (2007)
41. Okita, T., Rubino, R., Genabith, J.v.: Sentence-Level Quality Estimation for MT System Combination. In: Proceedings of the Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT, pp. 55–64. COLING'12, Mumbai, India (2012)
42. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. Philadelphia, Pennsylvania, USA (2002)
43. Sánchez-Cartagena, V.M., Sánchez-Martínez, F., Prez-Ortiz, J.A.: Integrating shallow-transfer rules into phrase-based statistical machine translation. In: Proceedings of the XIII Machine Translation Summit, pp. 562–569. Xiamen, China (2011)

44. Sánchez-Martínez, F., Forcada, M.L.: Inferring shallow-transfer machine translation rules from small parallel corpora. Journal of Artificial Intelligence Research **34**, 605–635 (2009). 00000
45. Sánchez-Martínez, F., Forcada, M.L., Way, A.: Hybrid rule-based example-based MT: feeding apertium with sub-sentential translation units. In: M.L. Forcada, A. Way (eds.) Proceedings of the 3rd Workshop on Example-Based Machine Translation, pp. 11–18. Dublin, Ireland (2009)
46. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A Study of Translation Edit Rate with Targeted Human Annotation. In: Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas (AMTA 2006), pp. 223–231. Cambridge, Massachusetts, USA (2006)
47. Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit. In: Proceedings of the Seventh International Conference of Spoken Language Processing (ICSLP2002), pp. 901–904. Denver, Colorado, USA (2002)
48. Thurmair, G.: Comparing Different Architectures of Hybrid Machine Translation Systems. In: Proceedings of the Machine Translation Summit XII, pp. 340–347. Ottawa, Ontario, Canada (2009)
49. Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., Sawaf, H.: Accelerated DP Based Search for Statistical Translation. In: Proceedings of the Fifth European Conference on Speech Communication and Technology, pp. 2667–2670. Rhodes, Greece (1997)
50. Tyers, F.M., Sánchez-Martínez, F., Forcada, M.L.: Flexible finite-state lexical selection for rule-based machine translation. In: Proceedings of the 16th Annual Conference of the European Association for Machine Translation, pp. 213–220. Trento, Italy (2012). 00004
51. Xu, J., Uszkoreit, H., Kennington, C., Vilar, D., Zhang, X.: DFKI Hybrid Machine Translation System for WMT 2011: on the Integration of SMT and RBMT. In: Proceedings of the Sixth Workshop on Statistical Machine Translation, pp. 485–489. Association for Computational Linguistics, Edinburgh, Scotland (2011)