# Use of Rich Linguistic Information to Translate Prepositions and Grammatical Cases to Basque

**Eneko Agirre, Aitziber Atutxa, Gorka Labaka, Mikel Lersundi,**
**Aingeru Mayor, Kepa Sarasola**
IXA Group. University of the Basque Country
aingeru@ehu.es

## Abstract

This paper presents three successful techniques to translate prepositions heading verbal complements by means of rich linguistic information, in the context of a rule-based Machine Translation system for an agglutinative language with scarce resources. This information comes in the form of lexicalized syntactic dependency triples, verb subcategorization and manually coded selection rules based on lexical, syntactic and semantic information. The first two resources have been automatically extracted from monolingual corpora. The results obtained using a new evaluation methodology show that all proposed techniques improve precision over the baselines, including a translation dictionary compiled from an aligned corpus, and a state-of-the-art statistical Machine Translation system. The results also show that linguistic information in all three techniques are complementary, and that a combination of them obtains the best F-score results overall.

## 1 Introduction

Since the first Machine Translation (MT) systems up to today's, performing well the translation of the prepositions is relevant for any MT system; Japkowicz and Wiebe (1991) claimed that doing it correctly is difficult because prepositions cannot be translated in a systematic or coherent way. Koehn (2003) remarked the importance of the correct translation of prepositions and he also reported that the main reason for noun phrase (NP) and prepositional phrase (PP) mistranslations consists of choosing wrong leading preposition.

Translation of prepositions is even more complex when the verb phrase and prepositional phrase structures differ widely in the languages involved in translation (Naskar and Bandyopadhyayn, 2006). This is what happens when translating from Spanish or English into Basque.

This paper explores the problem of translating prepositions heading verbal complements into target language equivalents. Although we focus on Spanish to Basque translation, the evaluation methodology and techniques can be applied to other language pairs. In Basque syntactic functions like subject, object and indirect objects are marked by case-suffixes. In this work postpositions and grammatical cases have been homogeneously treated, therefore it covers not only the translation of Spanish prepositions, but also how to choose the correct grammatical case corresponding to Spanish subjects, objects and indirect objects. Note that in most of the cases Spanish subjects and objects are not marked by any surface word or special case marking. Thus, besides the Spanish prepositions, we also explore the translation of the *zero preposition* corresponding to the grammatical cases of subject and object.

Given an existing open-source rule-based machine translation (RBMT) system called *Matxin* (Mayor, 2007; Alegria et al., 2007), we propose and evaluate three different techniques for translating Spanish prepositions and syntactic functions into Basque. These techniques use rich linguistic information like verb/postposition[1]/head-word dependency triples, verb subcategorization and manually coded selection rules based on lexical, syn-

[1] When we use here the word postposition, we would like to refer to grammatical cases and postpositions

tactic and semantic information. While the latter rules have been coded manually, the first two resources have been automatically extracted from monolingual corpora.

One important contribution of this paper is the evaluation methodology. Previous work (Husain et al., 2007; Gustavii, 2005) on preposition translation measured only accuracy gains with respect to simple baselines, and focused on small sets of frequent prepositions. Our methodology measures both precision and recall over all prepositions occurring in a small corpus of randomly chosen sentences. Once the evaluation corpus has been compiled, the evaluation is fully automatic.

The results of this paper shows that all proposed techniques improve over the baselines, including a translation dictionary compiled from an aligned corpus, and over a full-fledged statistical machine translation (SMT) system. The results also show that the linguistic information in all three techniques is complementary, and a combination of them obtains the best results overall.

In the next section of this paper we describe the RBMT system used, followed by a small review of related work on preposition translation. We then present the linguistic knowledge used. Section 5 presents the different baselines and techniques to translate prepositions. Our evaluation methodology is proposed in Section 6, which is followed by Section 7 with the results. Finally, Section 8 is devoted to conclusions and future work.

## 2 Preposition translation in RBMT

The last decade has seen the raise of SMT techniques, and less research on rule-based techniques. Nevertheless, translation involving a less-resourced language poses serious difficulties for SMT, specially caused by the smaller size of parallel corpora. Morphologically-rich languages have also been proved to be difficult for SMT, as shown in (Koehn and Monz, 2006), where SMT systems lag well behind commercial RBMT systems. At present, domain-specific translation memories for Basque are no bigger than two or three million words, much smaller than corpora used for other languages (the Europarl parallel corpus, for instance, has ca. 30 Mwords). Having limited digital resources, the rule-based approach is suitable for the development of an MT system for Basque, along with a focus on the enhancement of the core RBMT system with statistical and linguistic information.

The freely available open-source *Matxin* system is the first MT system available for Basque. It is a rule-based transfer system based on deep syntactic analysis. which currently translates from Spanish into Basque, and is currently being adapted to the English-Basque pair. The current development status shows that it is useful for content assimilation, for text understanding indeed, but that it is not yet suitable for unrestricted use in text dissemination.

*Matxin* has been evaluated and compared with the state-of-the-art corpus-based *Matrex* MT system (Stroppa et al., 2006; Labaka, 2007) translating from Spanish to Basque. The evaluation was performed using the *edit-distance* metric (Przybocki et al., 2006), based on the *HTER* (human-targeted translation edit rate) presented in (Snover et al., 2006), and the comparative results have shown that *Matxin* performs significantly better: 43.60 vs. 57.97 in the parallel corpus where *Matrex* was trained, and 40.41 vs. 71.87 in an out-of-domain corpus.

The preposition translation module of *Matxin* is located in the structural transfer phase and uses the information carried over from the syntactic analysis and lexical transfer modules. The system currently uses *Freeling* analyzer for Spanish (Atserias et al., 2006). The output of the preposition translation module is later used in subsequent modules in the structural transfer and generation phases. Note that errors from previous modules affect the quality of the preposition translation phase, and this makes the separate evaluation of preposition translation a difficult task. We will get back to this problem in Section 6.

## 3 Related work

Koehn (2003) envisages MT as a divide and conquer task where improving NP/PP translation will carry an improvement of the whole system. That study concluded that the main source of re-ranking errors in NP/PPs translation was the inability to correctly predict the phrase start (preposition or determiner) without context; it can sometimes only be resolved when the English verb is chosen and its subcategorization is known.

There are two main approaches to disambiguate prepositions (Mamidi, 2004; Alam, 2004; Trujillo, 1992): context based (used in transfer systems and more suitable for languages that are structurally different) and concept based (used in interlingua systems and more suitable for languages which are very close). Most of the systems are context based and they use transfer rules given with semantic information for the nouns which are head and complement of the preposition.

(Miller, 2000) argued that statistical models for preposition selection must take into account not only affinities between verbs and prepositions, but affinities between prepositions and nouns functioning as their complement as well.

(Husain et al., 2007) describes an approach to automatically select from two Indian languages the appropriate lexical correspondence of English simple prepositions. They use a set of rules that deal with syntactic and lexical-semantic constraints on the head and complement of the preposition. The results showed relative improvements greater than 20% in precision when compared to the default sense, but the experiments were conducted with just 6 high frequency prepositions. The algorithm was tested on 100 sentences for each preposition The input to the implemented system had been manually checked and corrected to make sure that there were no errors in the PP attachment given by the parser and no mistakes in phrasal verb identification.

(Naskar and Bandyopadhyayn, 2006) describes how the prepositions are handled in an English-Bengali MT system. As in Basque, there is no concept of preposition in Bengali. English prepositions are translated using inflections and/or postpositional words. The choice of the appropriate inflection depends on the spelling of the complement of the preposition and the choice of the postpositional word depends on its semantic information, obtained from the *WordNet*. They don't report any evaluation.

(Gustavii, 2005) corrected the preposition translations using a TBL classifier. She used aligned bilingual corpus data to infer her classifiers. Her evaluation is performed giving translation accuracy for only the six most frequent prepositions in the training corpus. She used a subset of 3 million tokens of the Swedish-English Europarl corpus, 90% for training and 10% for testing. The relative total improvement is of 12,45% (75,5% accuracy for the baseline and 84.9% for her system). However the applicability of the strategy is limited to relatively similar languages, as the ones of that study (Swedish and English). In fact the system avoids inducing rules where a preposition should be changed to some other part-of-speech, or where it should be completely removed. So this approach is not useful to translate from Spanish to Basque.

| Freq. | Transitivity | Postpositions |
|---|---|---|
| 4289.78 | transitive | ABS,ERG |
| 1534.24 | intransitive | ABS |
| 975.31 | transitive | ABS,ERG,INE |
| 476.70 | intransitive | ABS,INE |
| 166.68 | transitive | ABS,ERG,INS |

Table 1: Subcategorization for verb *ikusi* (*to see*).

## 4 Acquisition of rich linguistic information from corpus

Before showing our specific techniques for preposition translation, we briefly present the linguistic resources used, and how they were automatically acquired from Basque monolingual corpora.

### 4.1 Verb subcategorization

One of the information sources used for this experiment was an already existing subcategorization dictionary, initially built with the purpose of making attachment decisions for a shallow parser on its way to full parsing (Atutxa, forthcoming). For each of the 2,571 verbs this dictionary lists information about possible postposition and grammatical case combinations, transitivity, and estimated frequency of each combination. Table 1 shows the most frequent patterns in the dictionary entry for verb *ikusi* (*to see*), including estimated frequency, transitivity and postpositions (including grammatical cases) [2].

This dictionary was automatically built from raw corpora, comprising a compilation of 18 months of news from *Euskaldunon Egunkaria* (a newspaper written in Basque). The size of the corpus is around 780,000 sentences, approximately 10 Mwords. From the 5,572 different verb lemmas in the corpus, the subcategorization dictionary was compiled for the 2,751 verbs occurring at least 10 times.

The corpus was parsed by a chunker (Aduriz et al., 2004) which includes both named-entity and multiword recognition. The chunker uses a small grammar to identify heads, postpositions and verb attachments of NPs and PPs. The grammar was developed based on the fact that Basque is a head

[2] ABS : absolutive case (can be subject or object depending on transitivity). ERG : ergative (subject with transitive verbs). INE : inesive. INS : instrumental. DAT : dative. ALA : allative.

final language and it includes a distance feature as well. Phrases were correctly attached to the verb with a precision of 0,78. Note that the auxiliary verb in Basque allows to unambiguously determine the transitivity of the main verb. Given the fact that Basque is a three-way pro-drop language (subject, object and indirect object can be elided), cases of elided arguments were recovered from the auxiliary verb in most of the cases. The only exception were unergative verbs (e.g. *lo egin* – to sleep), which incorporate the missing argument. Statistical thresholds were used to reduce the errors caused by unergative verbs and wrong verb attachment decisions.

### 4.2 Verb/postposition/head-word dependency triples

Verbal subcategorization can be also modeled using attested (verb, dependency, head word) triples. The postposition can be used as the type of the dependency. In contrast to the subcategorization dictionary, and given that the headword is also kept, these triples are bound to be more sparse. Due to sparseness, the statistical threshold used for subcategorization acquisition proved to be ineffective, and it was devised an alternative acquisition method.

Only dependencies from the preverbal position of each clause were extracted. This position is the focus position of Basque, and the probability that a phrase at this position is attached to the verb just behind is quite high (up to 0.93 precision). Given the fact that Basque is a free word order language, and provided it is used a large enough corpus, it can be expected all arguments of a given verb to appear at the preverbal position in some attested sentence. This way, most of the potential arguments of a verb would be attested in the preverbal position, and therefore be captured as licit arguments of the verb. Table 2 shows the top triples for verb *ikusi* (*to see*). Attested headwords in the example include also elided pronouns and named-entities (of types PERSON, LOCATION, ORGANIZATION).

## 5 Strategies for preposition translation

In this section we present both the dictionary and aligned corpora baselines, alongside our three methods to translate prepositions: a context based approach using manually coded selection rules, and the use of subcategorization information or dependency triples to disambiguate the prepositions heading verbal complements.

| Freq. | Postposition | Head word |
|---|---|---|
| 70 | ERG | PRONOUN |
| 36 | ABS | PRONOUN |
| 30 | ERG | PERSON |
| 16 | INE | LOCATION |
| 13 | ABS | talde (group) |
| 11 | ABS | LOCATION |
| 9 | ABS | ORGANIZATION |
| 9 | ABS | partidu (match) |

Table 2: Dependency triples for verb *ikusi*.

### 5.1 Baselines

The baseline dictionary uses the preposition translations in the *Elhuyar* dictionary (Elhuyar, 2000), the most popular Spanish-Basque dictionary. The first postposition is taken as the preferred translation.

The aligned corpora baseline was constructed applying Giza++ (Koehn et al., 2003) to the *Consumer* magazine parallel corpus (Alcazar, 2006). This corpus contains 60,000 parallel sentences in Spanish (1.3 Mwords) and Basque (1 Mwords). The Basque part of the corpus was morphologically analyzed and segmented, i.e. word forms were split into their lemma and postposition (e.g.: *etxetik* (from the house) → *etxe* (the house) + *tik* (from)). After preprocessing the Basque sentences, we aligned the text automatically and extracted for each Spanish preposition its most frequent corresponding Basque postposition. This alignment technique proved to be superior to word-base alignment (Agirre et al., 2006). For a given Spanish preposition, the most frequent alignment was chosen as its Basque translation.

Note that these techniques do not tackle the translation of subject and object *zero prepositions* into Basque postpositions. In both baselines prepositions are always translated in the same way, irrespective of the context of occurrence of the preposition.

### 5.2 Selection rules

The preposition dictionary used as baseline above contains 351 Spanish prepositions (18 simple and 333 compound) plus what we call *zero preposition* for subject and object, and the possible Basque postpositions (462 in total) into which they can be translated. We have manually coded 89 selection rules to select the appropriate equivalent for the ambiguous prepositions.

| Prep. | Postpos. | Rule |
|---|---|---|
| a | INE | ./[@nounPOS=’Zm’] |
| a | DAT | - |
| a | ABS | - |
| a | ALA | ./[@si=’cc’] |

Table 3: Rule for the Spanish preposition *a*.

The rules contain lexical, syntactic and semantic information about the parent of the PP, and about the words in the PP (mainly the head).

Selection rules select or discard possible postpositions for one preposition, and can thus return, in general, more than one postposition. In the case of multiple suggestions, another method would be used to choose among those returned by the selection rules.

For example, given the sentence *Los venden a tres euros* (They sell them for three euros), the possible translations for the preposition *a* are the cases INE, DAT, ABS and ALA, as we can see in Table 3[3]. The rule that selects INE is applied because the *part-of-speech* of the head of the prepositional phrase is *Zm* and thus the selected translation will be INE: *Hiru eurotan saltzen dituzte*.

### 5.3 Verb subcategorization

Given a source sentence, the system accesses its syntactic analysis (provided by *Freeling* Spanish parser) and retrieves the verbs and a list with their dependent NPs and PPs. We process each verb in turn. For each of the NPs and PPs, the dictionary is used to retrieve all possible translations of the prepositions, building a data structure that contains the main verb and a list of potential translations for each of its dependent NPs and PPs. We also retrieve the translation of the main verb as produced by the lexical selection modules of *Matxin*. The algorithm then examines the subcategorization patterns of the translation of the verb, starting from the most frequent one, until it finds a pattern that matches the aforementioned data-structure.

For instance, given a source sentence like *yo he visto a tu madre* (I have seen your mother), we retrieve the main verb (*visto* - seen) and two dependents: the subject NP (*yo* – I) and the direct object which in Spanish uses the preposition *a* (*a tu madre* – your mother). The possible translations for the zero preposition are ABS, ERG and INE. The possible translations for *a* are ABS, DAT, ALA and INE. Given the translation of the verb, *ikusi* as suggested by *Matxin*, we can now access its subcategorization patterns from the dictionary as described in Section 4.1. The most frequent pattern for *ikusi* is (transitive, ABS,ERG), as shown in Table 1. As this subcategorization frame matches the example (ERG for *yo* and ABS for *a tu madre*) ERG and ABS grammatical cases are selected as translation in Basque. This information would be passed onto the generation module of *Matxin*.

### 5.4 Verb/postposition/head-word dependency triples

The algorithm in this case is very similar to that used in the subcategorization method. For each verb in the source sentence, we generate a data structure with the translation of the verb, and the list of dependent NPs and PPs, with the possible translation postpositions for each. Here we also add the translations into Basque of all heads of NPs and PPs.

Contrary to subcategorization, we treat each dependent NP and PP independently, one at a turn, choosing the most frequent dependency triple which matches the translation of the verb, one of the translations of the postposition and the translation of the noun. In other words, we choose the postposition which occurs first in the triples for this verb and head-word combination.

We will illustrate this example with a different example. Given the source sentence *El se conecta a Internet* (He connects to the Internet), we focus on the translation of the *a* preposition. *Matxin* translates *conecta* as *konektatu* and *Internet* as *Internet*. Given the set of possible translations for *a* (ABS, DAT, ALA and INE), the list of triples containing *konektatu* and *Internet* is checked, and the ALA postposition is chosen as the most frequent one for those.

### 5.5 Combination of techniques

Given a set of single techniques for preposition translation, we can combine them in several ways. Most of the techniques above have partial recall (i.e. they sometimes are not able to choose a single best translation), due mainly to sparse data problems. We therefore decided to combine them in cascade, one after the other, disambiguating in each step the prepositions which had not been translated in the previous one. We tried several combinations, as will be shown in the following

[3] *si*: syntactic information. *cc* circumstancial complement. *Zm*: tag for currency.

| Phrase | Preposition | Postposition |
|---|---|---|
| El mensaje | - | ABS |
| por correo | por | INS |
| a su amiga | a | DAT |

Table 4: An example of the gold standard.

section, but the cascade always orders the techniques according to their precision in the test set.

## 6 Evaluation framework

We ruled out the use of *Bleu* because, as pointed in (Callison-Burch et al., 2006), it cannot be always used to identify the improvements of the aspects of the translation. In our case, it is impossible to establish how much the *Bleu* score should rise or drop to detect significant improvements in the translation of prepositions.

We designed the evaluation framework in order to provide automatically both precision and recall for all prepositions. To create the gold standard, we selected 300 sentences at random from a parallel corpus of newspapers and technical reports. As our evaluation had to isolate the preposition translation task, the output of previous modules in the MT engine for each sentence was examined and if there was any mistake that affected the preposition translation (e.g. in the source text analysis or in the verb transfer), we discarded the sentence. In the remaining 54 sentences there were 80 Spanish prepositions and 81 syntactic functions (subject, direct object and indirect objects) to translate.

Table 4 shows an example of the gold standard. For the sentence *El mensaje ha sido enviado por correo a su amiga* (The message has been sent by mail to her friend) we coded the correct postposition for the prepositions (included the *zero preposition* in subject) of these three phrases: *El mensaje* (The message), *por correo* (by mail), *a su amiga* (to her friend).

## 7 Evaluation results

Table 5 shows, for each strategy, the number of correctly translated postpositions and the total number of postpositons translated (both correctly and incorrectly), alongside the overall number of cases in the test case. Precision, recall and F-score (actually, $F_1$) are also included. Significance ranges for F-score have been computed using bootstrap resampling for 95% confidence. Given the small size of the dataset, the significance ranges are quite large, over 5 percentage points on all cases.

The first set of rows shows the results for the baselines. We can see that the dictionary performs better than the translations coming from the aligned corpus, which was an unexpected finding. Both baselines return a translation in all cases, and have recall identical to the precision.

The second set of rows describes the performance of each of the techniques proposed in this paper. The manually coded selection rules method has the highest precision, but it scores second in recall and F-score. Subcategorization obtains the lowest precision from the three techniques, but the best recall and F-score. The precision of all of our techniques improves over the baselines, but, due to the fact that they don't always provide a translation, recall and the F-score are lower.

Regarding combination, the third set of rows presents several cascades of techniques. Combining single techniques with the first sense baseline basically provides full coverage and improves recall, providing non-significant improvements on F-score for rules and triples, and statistically significant improvement for subcategorization. The pairwise combination of two techniques gets good precision, but not full coverage, and F-score is similar to the 1st sense baseline. On the same set of results the cascade of all three methods is reported to have very high precision and recall.

The last four rows report the results for pairwise and three-wise combinations of the techniques with the 1st sense baseline. The improvement is consistent in all combinations, and the best result is for the combination of all.

Given the small number of examples only a few performance differences are statistically significant. Below we list the pairs of results (among those which have full coverage, i.e. those using 1st sense) that are statistically significant:

1st sense < a+b+c+1st
a+1st < a+b+c+1st
b+1st < a+b+c+1st

Regarding the comparison among techniques, and although the differences are not statistically significant, the combinations that use subcategorization are the ones performing best, and it is always the single technique which improves most in each combination class. This is further enforced by the fact that a+1st and b+1st perform significantly worse than a+b+c+1st, while the difference

| | Correct | Translated | Overall | Precision | Recall | F-score | Signif. |
|---|---|---|---|---|---|---|---|
| Baselines | | | | | | | |
| Dictionary | 109 | 161 | 161 | **67.70%** | **67.70%** | **67.70%** | ±6.26 |
| Alignment Dict. | 101 | 161 | 161 | 62.73% | 62.73% | 62.73% | ±5.98 |
| Techniques | | | | | | | |
| Rules (a) | 73 | 83 | 161 | **87.95%** | 45.34% | 59.84% | ±6.73 |
| Triples (b) | 54 | 62 | 161 | 87.10% | 33.54% | 48.43% | ±7.40 |
| Subcat (c) | 84 | 107 | 161 | 78.50% | **52.17%** | **62.69%** | ±6.78 |
| Combinations | | | | | | | |
| a+1st | 110 | 161 | 161 | 68.32% | 68.32% | 68.32% | ±6.64 |
| b+1st | 111 | 161 | 161 | 68.94% | 68.94% | 68.94% | ±6.30 |
| c+1st | 116 | 161 | 161 | 72.05% | 72.05% | 72.05% | ±5.42 |
| a+b | 87 | 98 | 161 | **88.78%** | 54.04% | 67.18% | ±6.09 |
| b+c | 89 | 112 | 161 | 79.46% | 55.28% | 65.20% | ±6.41 |
| a+c | 99 | 124 | 161 | 79.84% | 61.49% | 69.47% | ±6.11 |
| a+b+c | 103 | 125 | 161 | 82.40% | 63.98% | 72.03% | ±5.48 |
| a+b+1st | 115 | 161 | 161 | 71.43% | 71.43% | 71.43% | ±5.92 |
| b+c+1st | 118 | 161 | 161 | 73.29% | 73.29% | 73.29% | ±5.91 |
| a+c+1st | 117 | 161 | 161 | 72.67% | 72.67% | 72.67% | ±5.68 |
| a+b+c+1st | 121 | 161 | 161 | 75.16% | **75.16%** | **75.16%** | ±5.70 |

Table 5: Overall results of baselines, single techniques and combinations.

| | Correct | Translated | Overall | Precision | Recall | F-score | Signif. |
|---|---|---|---|---|---|---|---|
| $\text{SMT}_{wordforms}$ | 60 | 161 | 161 | 37.27% | 37.27% | 37.27% | ±6.84 |
| $\text{SMT}_{segmented}$ | 82 | 149 | 161 | 55.03% | 50.93% | 52.90% | ±6.35 |

Table 6: Results for SMT systems trained with word forms and segmented words

between c+1st and a+b+c+1st is not significant.

Table 6 shows the results obtained by two state-of-the-art full-fledged SMT systems, one of them was trained using Basque word forms for alignment, and the other using Basque segmented words (see Section 5.1). The whole sentences were translated and then the postpositions related to the translated phrases were compared with the gold standard. Their results are clearly lower than those obtained with each of the three simple strategies or any of their combinations.

## 8 Conclusions and future work

In this work, three techniques that use rich linguistic information to translate grammatical cases and prepositions heading verbal complements have been implemented and successfully evaluated in the context of an RBMT system for an agglutinative language with scarce resources. They are based on verb/postposition/head-word dependency triples, verb subcategorization and manually coded selection rules based on lexical, syntactic and semantic information. The first two resources have been automatically extracted from monolingual corpus, that obviously is easier to collect than parallel corpus. As traslation involving a less resourced language poses serious dificulties for pure SMT, we think these two techniques based on monolingual corpus statistics are opening new ways to integrate rule-based and statistical-based techniques in MT languages with fewer digital resources.

A new methodology of evaluation has been designed. It allows to automatically measure precision and recall against a gold standard. Even if our test corpus is not very large, it is comparable with those used in related work, and the F-scores show that some of the improvements are statistically significant.

The proposed techniques improve precision over the baselines, including a translation dictionary compiled from an aligned corpus, and over a full-fledged SMT system. The results also show that the linguistic information in all three techniques is complementary, and a combination of them obtains the best results overall.

In the near future we plan to collect larger linguistic resources to obtain better information on verb subcategorization and verb/postposition/head-word triples, so we could improve our present results. We also plan to enlarge the gold standard and to evaluate the relevance of our techniques in overall translation quality, using the edit-distance metric (Przybocki et al., 2006). We would also like to use the output of SMT systems in the combined system.

## Acknowledgments

This research was supported in part by the Spanish Ministry of Education and Science (OpenMT: Open Source Machine Translation using hybrid methods, TIN2006-15307-C03-01; RICOTERM-3, HUM2007-65966.CO2-02) and the Regional Branch of the Basque Government (AnHITZ 2006: Language Technologies for Multilingual Interaction in Intelligent Environments, IE06-185). Gorka Labaka is supported by a PhD grant from the Basque Government (grant code, BFI05.326). Consumer corpus has been kindly supplied by Asier Alcázar from the University of Missouri-Columbia and by Eroski Fundazioa.

## References

E. Agirre, A. Díaz de Ilarraza, G. Labaka and K. Sarasola. 2006. *Uso de información morfológica en el alineamiento Español-Euskara*. XXII Congreso de la SEPLN.

Y. S. Alam. 2004. *Decision Trees for Sense Disambiguation of Prepositions: Case of Over*. HLT-NAACL 2004: Workshop on Computational Lexical Semantics . Boston, Massachusetts, USA. ACL.

A. Alcázar. 2006. *Towards linguistically searchable text*. Proceedings of BIDE 2005. Bilbao.

I. Alegria, A. Díaz de Ilarraza, G. Labaka, M. Lersundi, A. Mayor and K. Sarasola. 2007. *Transfer-based MT from Spanish into Basque: reusability, standardization and open source*. LNCS 4394. 374-384. Cicling 2007.

I. Aduriz, M. J. Aranzabe, J. M. Arriola, A. Díaz de Ilarraza, K. Gojenola, M. Oronoz and L. Uría. 2004. *A Cascaded Syntactic Analyser for Basque.* In Gelbukh, A (ed.) Computational Linguistics and Intelligent Text Processing. Springer LNCS 2945.

J. Atserias, B. Casas, E. Comelles, M. González, L. Padró and M. Padró. 2006. *FreeLing 1.3: Syntactic and semantic services in an open-source NLP library*. Proceedings of the 5th LREC (2006). Genova. Italia.

C. Callison-Burch, M. Osborne and P. Koehn. 2006. *Re-evaluating the role of BLEU in Machine Translation Research*. Proceedings of EACL-2006.

Elhuyar 2000. *Elhuyar Hiztegia*. Published by Elhuyar Hizkuntz Zerbitzuak.

E. Gustavii. 2005. *Target language preposition selection - an experiment with transformation based learning and aligned bilingual data*. Proceedings of the 10th EAMT conference. May 2005. Budapest.

S. Husain, D.M. Sharma and M. Reddy. 2007. *Simple preposition correspondence: a problem in English to Indian language machine translation*. Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions. Prague, Czech Republic. 28 June 2007. pp.51-58.

N. Japkowicz and J. Wiebe. 1991. *A System for Translating Locative Prepositions from English into French*. Proceedings of the Meeting of the ACL. 153-160.

P. Koehn. 2003. *Noun Phrase Translation*. PhD. Thesis.University of Southern California.

P. Koehn, F. Och, and D. Marcu (2003). *Statistical phrase based translation*. In Proceedings of HLT-NAACL 2003, pp. 48-54, Edmonton, Canada.

P. Koehn and C. Monz. 2006. *Manual and Automatic Evaluation of Machine Translation between European Languages*. Proceedings of the Workshop on SMT. ACL. June 2006. New York City. pp. 102–121.

G. Labaka, N. Stroppa, A. Way and K. Sarasola. 2007. *Comparing Rule-Based and Data-Driven Approaches to Spanish-to-Basque MT*. Proceedings of the MT-Summit XI. Copenhagen.

R. Mamidi. 2004. *Disambiguating Prepositions for Machine Translation using Lexical Semantic Resources*. Proceedings of the 'National Seminar on Theoretical and Applied Aspects of Lexical Semantics' organized by Centre of Advanced Study in Linguistics. Hyderabad.

A. Mayor. 2007. *Matxin: Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz*. PhD. Thesis. (In Basque). University of the Basque Country.

K. Miller. 2000. *The lexical choice of prepositions in machine translation*. PhD. Thesis. Upenn.

S.K. Naskar and S. Bandyopadhyayn. 2006. *Handling of Prepositions in English to Bengali Machine Translation*. Proceedings of the EACL workshop on Prepositions, Hyderabad.

M. Przybocki, G. Sanders and A. Le. 2006. *Edit distance: a metric for Machine Translation evaluation*. Proceedings of the LREC-2006. Genoa, Italy.

M. Snover, B Dorr, R. Schwartz, L. Micciulla and J. Makhoul. 2006. *A study of translation edit rate with targeted human annotation*. Proceedings of the Association for Machine Translation in the Americas..

N. Stroppa, D. Groves, A. Way and K. Sarasola. 2006. *Example-Based Machine Translation of the Basque Language*. Proceedings of the 7th conference of the AMTA. pp.232–241. Boston.

A. Trujillo. 1992. *Locations in the Machine Translation of Prepositional Phrases*. Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in MT of Natural Languages. Montreal, Canada.