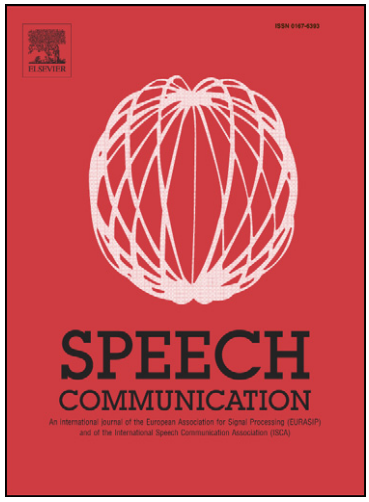
Joining linguistic and statistical methods for Spanish-to-Basque speech translation

Alicia Pérez, M. Inés Torres, Francisco Casacuberta



Joining linguistic and statistical methods for Spanish-to-Basque speech translation

Alicia Pérez, M. Inés Torres, Francisco Casacuberta

PII: S0167-6393(08)00090-3
DOI: 10.1016/j.specom.2008.05.016
Reference: SPECOM 1732

To appear in: *Speech Communication*

Received Date: 14 June 2007
Revised Date: 12 May 2008
Accepted Date: 29 May 2008

Please cite this article as: Pérez, A., Torres, M.I., Casacuberta, F., Joining linguistic and statistical methods for Spanish-to-Basque speech translation, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.05.016

Joining linguistic and statistical methods for Spanish-to-Basque speech translation

Alicia Pérez[(1)] * M. Inés Torres[(1)] Francisco Casacuberta[(2)]

[(1)] *Department of Electricity and Electronics*
Faculty of Science and Technology; University of the Basque Country
48940 Leioa; Spain
e-mail: manes@we.lc.ehu.es

[(2)] *Departament of Information Systems and Computation*
Faculty of Computer Science; Technical University of Valencia
Camí de Vera, s/n
46071 Valencia; Spain
e-mail: fcn@dsic.upv.es

Abstract

The goal of this work is to develop a text and speech translation system from Spanish to Basque. This pair of languages shows quite odd characteristics as they differ extraordinarily in both morphology and syntax, thus, attractive challenges in machine translation are involved. Nevertheless, since both languages share official status in the Basque Country, the underlying motivation is not only academic but also practical.

Finite-state transducers were adopted as basic translation models. The main contribution of this work involves the study of several techniques to improve probabilistic finite-state transducers by means of additional linguistic knowledge. Two methods to cope with both linguistics and statistics were proposed. The first one performed a morphological analysis in an attempt to benefit from atomic meaningful units when it comes to rendering the meaning from one language to the other. The second approach aimed at clustering words according to their syntactic role and used such phrases as translation unit. From the latter approach phrase-based finite-state transducers arose as a natural extension of classical ones.

The models were assessed under a restricted domain task, very repetitive and with a small vocabulary. Experimental results shown that both morphological and syntactical approaches outperformed the baseline under different test sets and architectures for speech translation.

Key words: spoken language translation, stochastic transducer, phrase-based translation model, morpho-syntactic knowledge modeling, bilingual resources

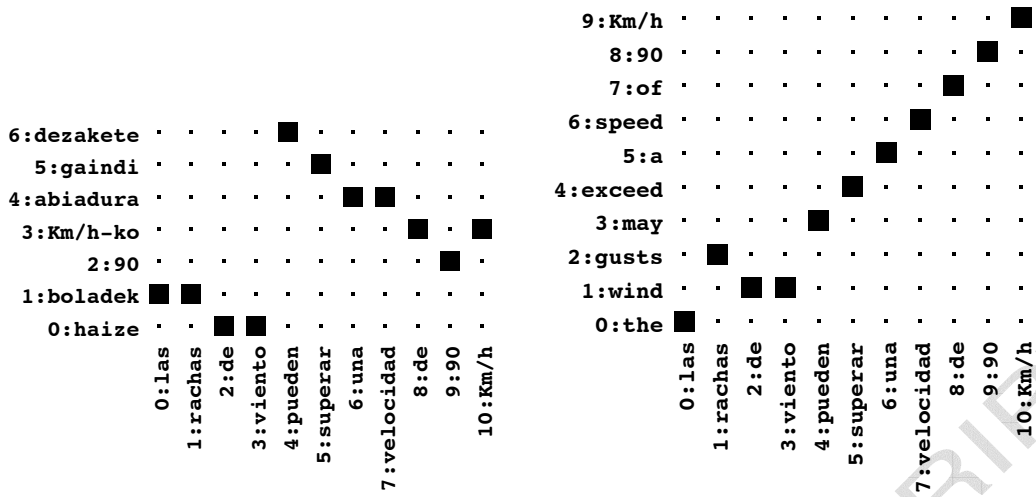
1 Introduction

Speech translation represents nowadays a challenge in natural language processing due to the difficulties of combining speech and translation technologies. The so called statistical framework is, without doubt, a very promising approach for speech and translation modeling. Nevertheless, this approach requires large training material. Furthermore, the scarcity of available linguistic resources associated with minority languages as Basque, Catalan or Galician, has to be faced in advance. The work presented in this paper focuses on Spanish to Basque speech and text translation. Thus, the first stage of this work consisted of the generation of linguistic resources (corpus and tools) for Basque (Pérez et al., 2006b).

Spanish and Basque languages differ significantly in both morphology and syntax (for further details see appendix A). Hence, specific problems have to be faced: on the one hand, Basque is a highly inflected language with 17 cases (for a matter of comparison, in German there are 4 grammatical cases, 7 in Czech and 15 in Finnish); on the other hand, the typical syntactic construction leads to long distance relationships between Spanish and Basque. These characteristics are, somehow, depicted in Fig. 1. The mentioned differences do not occur, to this extent, between Spanish and other Iberian languages, such as Catalan, Galician or Portuguese. Therefore, translation from Spanish into Basque not only does it exhibit academic interest, but also represents a real necessity since both languages are co-official for the 2.5 million inhabitants of the Basque Country. Other tools that aimed at translating text from Spanish into Basque have been previously implemented in the literature such as *Matxin* (Alegria et al., 2007) within the project *Opentrad* (Corbí-Bellot et al., 2005). It makes use of a transfer approach to cope with text translation. Alternatively, this work deals with statistical speech and text translation and, to the authors' knowledge, it is the first approach in the literature related to speech translation between this pair of languages.

Within the statistical framework, we deal with finite-state models, which have been extensively applied to many fields of natural language processing such as language or phonology modeling. They have also been successfully introduced for speech translation within restricted domains (Vidal, 1997; Bangalore and Riccardi, 2002). There are different approaches that cope with machine

* Alicia Pérez
Dep. Electricity and Electronics;
Fac. Science and Technology
University of the Basque Country
48940 Leioa; Spain
e-mail: alicia.perez@ehu.es
Tel: +34 946015364



(a) Spanish-Basque (b) Spanish-English

Fig. 1. Manual alignments of a sentence show the different word order between Basque and Spanish; for a matter of clarity, Spanish-English alignment is also given. That is, both Basque and English sentences (in the abscissa) are the translation of the same Spanish sentence (in the ordinate).

translation using finite state models. On the one hand, (Bangalore and Riccardi, 2001) proposed a two step translation process where first, a mapping from source to target language allowed the lexical selection, and then, a lexical reordering based on a tree-structure arranged the output so that it was grammatically correct. Other works (Knight and Al-Onaizan, 1998; Kumar et al., 2005) formulated a generative source-channel model influenced by the IBM-Models (Brown et al., 1993) but implemented with weighted finite-state transducers. Our work, instead, focuses on a unique model that copes, at the same time, with both transference and, to some extent reordering. It defines a joint probability distribution over all possible translations. This approach leads to the integration of acoustic and translation models in a unique model that can be inferred from samples allowing a single and fast decoding. This integration presents certain similarity with the composition of several finite-state models proposed for speech recognition purposes (Caseiro and Trancoso, 2001).

Taking the aforementioned finite-state models as a baseline, we aimed at improving their performance. Current trends in machine translation suggest that statistical methods could be benefited from linguistic knowledge (Och et al., 2003). Even though there have already been some attempts at exploiting linguistic resources within statistical methods (Nießen and Ney, 2001; Collins et al., 2005; de Gispert et al., 2006), this issue is still an open problem. In this work, morphologic and syntactic knowledge sources have been tightly introduced within statistical translation models. In this framework, two different approaches were proposed and assessed with a series of experimental results in a limited domain but highly practical application task. On the one hand,

a morphological approach attempted at implementing the meaning-transfer models using a categorized target language and then completing the target lexical choice. On the other hand, a syntactical approach tried to build the transducer taking syntactic phrases as translation unit instead of running words. In fact, the state of the art in machine translation suggests the use of phrases as translation unit instead of words (Koehn et al., 2006). As it will be shown, the proposed phrase-based SFST approach is related with a monotonic approach of the commonly used phrase-based models.

All in all, the aim of this paper is to make progress within the field of speech technologies for under-resourced languages as it is the case of Basque. The adopted strategy consists of reinforcing statistical methods by including linguistic knowledge within the finite-state framework. Since the pair of languages under study differ extraordinarily in what word ordering and agglutination concerns (as mentioned in appendix A), new strategies were to be explored. As a by-product, a new phrase-based approach has been formulated and implemented.

The organization of this paper is as follows: the statistical framework is faced in section 2, where speech translation problem is tackled with two different architectures. Stochastic finite-state transducers are developed in section 3, where decoding and learning problems are addressed, in addition, as an extension of those models a phrase-based approach for finite-state transducers is presented. Then, we propose to enrich the mentioned statistical models making use of linguistic features. In particular two approaches are explored in section 4. These approaches have been evaluated within the task described in section 5. Experimental results are given in section 6 and finally, some conclusions and guidelines for future work. For further details with regard to Basque and Spanish languages turn to appendix A.

2 Speech translation

The goal of the statistical speech translation is to search for the likeliest target language string $\widehat{\mathbf{t}}$, given the acoustic representation $\mathbf{x}$ of some source language string.

$$\widehat{\mathbf{t}} = \arg\max_{\mathbf{t}} P(\mathbf{t}|\mathbf{x}) \tag{1}$$

The transcription of the speech into text, is an unknown string $\mathbf{s}$ in the source language that might be considered a hidden variable.

$$\widehat{\mathbf{t}} = \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t}, \mathbf{s}|\mathbf{x}) \tag{2}$$

Applying the Bayes' decision rule:

$$\widehat{\mathbf{t}} = \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t},\mathbf{s}) P(\mathbf{x}|\mathbf{t},\mathbf{s}) \tag{3}$$

Let us assume that the probability of the acoustic signal related to the utterance in the source language has no dependency on the target string once the source string is known i.e. $P(\mathbf{x}|\mathbf{t},\mathbf{s})$ is independent of $\mathbf{t}$. Hence, eq. (3) can be rewritten as:

$$\widehat{\mathbf{t}} = \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t},\mathbf{s}) P(\mathbf{x}|\mathbf{s}) \tag{4}$$

In practice, the sum over all possible source strings in eq. (4) can be approximated by the maximum term involved.

$$\widehat{\mathbf{t}} \approx \arg\max_{\mathbf{t}} \max_{\mathbf{s}} P(\mathbf{t},\mathbf{s}) P(\mathbf{x}|\mathbf{s}) \tag{5}$$

Typically, eq. (5) is implemented in a sub-optimal way, by means of a speech recognizer and a text-to-text translation system in a **decoupled architecture**. Taking into account that $P(\mathbf{t},\mathbf{s}) = P(\mathbf{t}|\mathbf{s})P(\mathbf{s})$, this approach offers the translation of the speech transcription as follows:

(1) Given the acoustic representation $\mathbf{x}$, find its expected text transcription:

$$\widehat{\mathbf{s}} = \arg\max_{\mathbf{s}} P(\mathbf{s}) P(\mathbf{x}|\mathbf{s}) \tag{6}$$

where $P(\mathbf{s})$ is the probability of the string $\mathbf{s}$ according to a language model of the source language.

(2) Translation of $\widehat{\mathbf{s}}$ (the expected transcription of $\mathbf{x}$):

$$\widehat{\mathbf{t}} \approx \arg\max_{\mathbf{t}} P(\mathbf{t}|\widehat{\mathbf{s}}) \tag{7}$$

The serial architecture is the most widely used approach due to the fact that it is independent of the sort of translation paradigm used as both the speech recognition and the translation system are decoupled. Unfortunately, translation models are quite sensitive to input-errors (Sarikaya et al., 2007), thus the translation of two slightly distinct source strings may differ significantly. In short, the approach of eq. (5) by eq. (7) recurs to a strong assumption.

In order to achieve a tighter integration between speech recognition and translation stages, some effort have been made in the literature by using n-best lists (Quan et al., 2005), word-lattices (Saleem et al., 2004) or confusion networks (Bertoldi et al., 2007). Furthermore, joint probability in eq. (5) can be naturally implemented with finite-state transducers as it is well known. In addition, acoustic and translation finite-state models can be efficiently composed in an **integrated architecture** (Casacuberta et al., 2004). In practice (as will be shown in section 6.1), integrated architecture works as a speech

recognition system that makes use of a translation model instead of the usual language model. The same acoustic models, typically *hidden Markov models*, can be used for either speech recognition or speech translation. Our attention is thus focussed on language modeling. In fact, the translation model (under finite-state transducer methodology) involves two language models: the source language model, is the input projection of the transducer, and the target language model the output projection. On the whole, the composition seems to be a robust technique to hierarchically integrate knowledge-sources of different complexity or depth level in either speech recognition (Pereira and Riley, 1997; Caseiro and Trancoso, 2001) or speech translation (Casacuberta et al., 2004). In addition, there are efficient algorithms to carry out on the fly integration of these sort of models at decoding time (Caseiro and Trancoso, 2006).

3 Stochastic Finite-State Transducers

Finite-state transducers are versatile models that count on thoroughly studied efficient implementations for training (Casacuberta and Vidal, 2007) and decoding (Mehryar Mohri and Riley, 2003). Definition and layout for probabilistic finite-state machines (automata and transducers) were comprehensively described in (Vidal et al., 2005a,b), and so we are going to follow that formalism. Though the formal definition is reported next, as an introductory notion a *stochastic finite-state transducer* (SFST) might be roughly described as a finite-state machine where each transition, labeled with an input/output pair, has associated a probability to occur. That is, a Mealy machine with a set of probability distributions involved.

3.1 Definition

An SFST is tuple $\mathcal{T} = \langle \Sigma, \Delta, Q, q_0, R, F, P \rangle$, where:

Σ is a finite set of input symbols (source vocabulary);
Δ is a finite set of output symbols (target vocabulary);
Q is a finite set of states;
$q_0 \in Q$ is the initial state;
$R \subseteq Q \times \Sigma \times \Delta^* \times Q$ is a set of transitions such as $(q, s, \tilde{t}, q')$, which is a transition from the state q to the state q', with the source symbol s and producing the substring $\tilde{t}$;
$P : R \to [0, 1]$ is a transition probability distribution;
$F : Q \to [0, 1]$ is a final state probability distribution;

The probability distributions satisfy the stochastic constraint:

$$\forall q \in Q \quad F(q) + \sum_{s,\tilde{t},q'} P(q, s, \tilde{t}, q') = 1 \tag{8}$$

3.2 Decoding

The goal of statistical machine translation is to find the target language string $\mathbf{t}$ that better matches the source string $\mathbf{s}$. Within SFSTs, the analysis of a source string is carried out by exploring all the possible translation forms. A *translation form*, $d(\mathbf{s}, \mathbf{t})$, is a sequence of transitions in an SFST, $\mathcal{T}$. That is, a path compatible with both the source and the target strings:

$$d(\mathbf{s}, \mathbf{t}) : (q_0, s_1, \tilde{t}_1, q_1)(q_1, s_2, \tilde{t}_2, q_2) \cdots (q_{J-1}, s_J, \tilde{t}_J, q_J)$$

where $\mathbf{s} = s_1 s_2 \ldots s_J$ is a sequence of source symbols and $\mathbf{t} = t_1 t_2 \ldots t_I = \tilde{t}_1 \tilde{t}_2 \ldots \tilde{t}_J$ is a sequence of target substrings ($|\mathbf{s}| = J, |\mathbf{t}| = \sum_{j=1}^{J} |\tilde{t}_j| = I$). According to the SFST model, $\mathcal{T}$, the probability associated with a translation form is the following one:

$$P_{\mathcal{T}}(d(\mathbf{s}, \mathbf{t})) = F(q_J) \prod_{j=1}^{J} P(q_{j-1}, s_j, \tilde{t}_j, q_j) \tag{9}$$

Therefore, the probability of $(\mathbf{s}, \mathbf{t})$ to be a translation pair, is the sum of the probability of all the possible translation forms compatible with that pair.

$$P_{\mathcal{T}}(\mathbf{s}, \mathbf{t}) = \sum_{d(\mathbf{s},\mathbf{t})} P_{\mathcal{T}}(d(\mathbf{s}, \mathbf{t})) \tag{10}$$

As a result, $P(\mathbf{s}, \mathbf{t})$, the distribution involved in eq. (5) can be estimated with an SFST model:

$$P(\mathbf{s}, \mathbf{t}) \approx P_{\mathcal{T}}(\mathbf{s}, \mathbf{t}) = \sum_{d(\mathbf{s},\mathbf{t})} P_{\mathcal{T}}(d(\mathbf{s}, \mathbf{t})) \tag{11}$$

The resolution of the eq. (11) has proved to be a hard computational problem (Casacuberta and de la Higuera, 1999), but it can be efficiently computed by the *maximum approximation*, which replaces the sum by the maximum.

$$P_{\mathcal{T}}(\mathbf{s}, \mathbf{t}) \approx \max_{d(\mathbf{s},\mathbf{t})} P_{\mathcal{T}}(d(\mathbf{s}, \mathbf{t})) \tag{12}$$

Under this maximum approach *Viterbi algorithm* can be used to find the best sequence of states through the SFST given the input string. The translation is obtained by concatenating the output substrings through the optimal path.

3.3 Learning

Stochastic finite state transducers can be automatically learnt from bilingual corpora by efficient algorithms (Casacuberta and Vidal, 2007). The goal is to determine the topological parameters and probabilistic distributions defining a specific translation model for the task under consideration. Most common inference approaches (Casacuberta, 2000; Bangalore and Riccardi, 2002; Matusov et al., 2005) lead up to an intermediate bilingual representation of each pair of sentences in the training data. This representation is obtained by segmenting, under different constraints, the pair of sentences into a string of bilingual phrases: $(\tilde{s}_1, \tilde{t}_1)(\tilde{s}_2, \tilde{t}_2) \cdots (\tilde{s}_m, \tilde{t}_m)$.

In this work GIATI (Grammar Inference and Alignments for Transducers Inference) methodology is used to learn the translation models. This approach restricts the length of the source substrings to one and allows the length of the target substrings to be zero. GIATI was exhaustively described by (Casacuberta and Vidal, 2004) and it can be summarized as follows:

(1) For each bilingual pair $(\mathbf{s}, \mathbf{t}) = (s_1^J, t_1^I)$ from the training corpus, find a monotonic segmentation $(s_1^J, \tilde{t}_1^J)$. Thereby, assign an output sequence of zero or more words to each input word, leading to the so called *extended corpus* (the intermediate bilingual representation of each training pair in terms of a single sequence of bilingual tokens, as previously mentioned). To do so, direct and inverse statistical alignments extracted with GIZA++ free toolkit (Och and Ney, 2003) were considered in this work. Each extended string, $\mathbf{z} \in (\Sigma \times \Delta^*)^*$, satisfies:

$$\mathbf{z} = (s_1, \tilde{t}_1)(s_2, \tilde{t}_2) \ldots (s_J, \tilde{t}_J) \quad where \quad \sum_{j=1}^{J} |\tilde{t}_j| = I$$

(2) Then, infer a stochastic regular grammar. We promote the use of *k-testable in the strict sense* (k-TSS) grammars, which are a subset of regular grammars. Hence, the corresponding k-TSS stochastic finite-state automaton (SFSA) can be automatically learnt from positive samples making use of efficient algorithms based on a maximum likelihood approach (García and Vidal, 1990). k-TSS language models are considered to be the syntactic approach of the well known n-gram models. However, the syntactic approach allows to integrate K k-TSS models (ranging K from 1 to k) along with smoothing in a unique SFSA under a hierarchical back-off structure (Torres and Varona, 2001). Smoothing is of interest in

order to prevent the model from assigning null probabilities due to data sparseness at training stage (Jelinek, 1997).

(3) Finally, split the output sequence from the input word on each edge of the automaton, getting, in this way, the finite state transducer. That is, if a transition in the inferred automaton is like $(q_i, (s, \tilde{t}), q_j)$, then, there is an equivalent edge in the transducer with the input symbol s and the output substring $\tilde{t}$. The probability distributions and the topology learnt in the 2nd step remain unchanged.

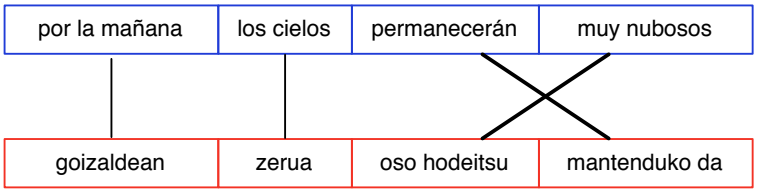
3.4 A novel implementation of monotonic phrase-based models

As shown in this section, one of the contributions of this article is the alternative formulation of a monotonic phrase-based (PB) approach turning to the finite-state (FS) framework. This approach puts together the improvements in translation quality related to PB approach and the flexibility and speed associated with FS models (González and Casacuberta, 2007; Pérez et al., 2007). Furthermore, this approach makes it possible a tight integration of both translation and acoustic models following the methodology proposed for automatic speech recognition in (Caseiro and Trancoso, 2006). In addition, this formulation allows to include morphologic information within the translation model as it will be shown latter.

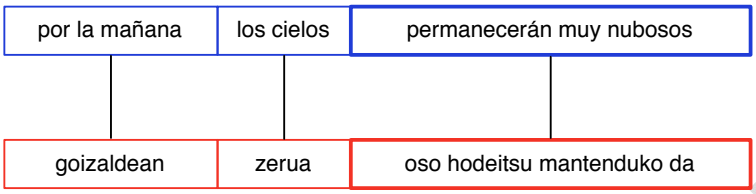
It is usually difficult to capture relationships between languages when long-distance alignments are involved. These distances decrease taking as alignment unit the phrase instead of the word. Furthermore, segmentations might become monotonic, and these are, in fact, the sort of relationships that stochastic finite-state transducers are good at. The mentioned issues are depicted in the Fig. 2 through a pair of sentences in Spanish and Basque whose translation into English is "during the morning skies will remain mostly cloudy".

A monotonic approach usually involves a faster decoding than a non-monotonic one, and indeed time efficiency is an essential issue in speech translation. In what performance concerns, word-based monotonic approach has shown to be a good approach to deal with similar languages. Furthermore, phrase-based monotonic approach shows a wider application scope, thus, it seems to be useful even between languages with long-distance alignments as is the case of Spanish and Basque (at least for tasks such as the one considered in this work). Therefore, phrase-based SFSTs could improve the performance of previous SFSTs when long-distance reorderings are involved.

The phrase-based approach for finite-state transducer was recently introduced in the literature. Let us note that it was the work (Casacuberta and Vidal, 2004) which set a precedent in this investigation line. Alternatively, in (Ku-



(a) Phrase-to-phrase alignment.



(b) Monotonic segmentation at phrase level.

Fig. 2. Phrase-to-phrase alignment and the subsequent monotonic segmentation between Basque (on the bottom), Spanish (on the top).

mar et al., 2005) *alignment template* translation model was implemented making use of weighted finite-state transducers (WFSTs). In (Zhou et al., 2005) the translation process was decomposed in terms of several probability distributions (specifically: output language model, permutation model, fertility model, *NULL insertion* model) and all of them were implemented as WFSTs. Both approaches were carried out by a composition of constituent transducers. Some assumptions were taken in order to infer several components from parallel training data. For instance, given a source sentence, uniform probability distribution over all the compatible segmentations was assumed, as well as for the number of segments to be generated. Alternatively, in this work, for a given training pair only one segmentation is taken into account, and thereby a single SFST is inferred. There is no need of other intermediate models since this model is self contained and copes with both the meaning transference and the arrangement of the output string at the same time. An intrinsic restriction on our approach, however, is its monotonic nature (as earlier mentioned).

Here a natural extension of the transducers described in previous section is proposed so that they cope with phrases instead of running words. To do so, the set of transitions (see section 3.1) is redefined as: $R \subseteq Q \times \Sigma^+ \times \Delta^* \times Q$, where such a transition, $(q, \tilde{s}, \tilde{t}, q')$, links the state q to the q', with the source sub-string $\tilde{s}$ and producing the target substring $\tilde{t}$. In what the inference of such a model regards, first of all, let us extract the set of all the possible segmentations $\mathcal{S}$ for a given pair $(\mathbf{s}, \mathbf{t})$. The segmentation may be related to either statistically or linguistically motivated phrases without any restriction. A specific segmentation for a given pair, denoted as $\sigma \in \mathcal{S}(\mathbf{s}, \mathbf{t})$, can be defined by the number of segments to be made along with the limits of the segments in the source and the target sentences respectively. In this context, segmentations

can be introduced as hidden variables within the joint probability model.

$$P(\mathbf{s},\mathbf{t}) = \sum_{\sigma} P(\mathbf{s},\mathbf{t},\sigma) = \sum_{\sigma} P(\sigma)\cdot P(\mathbf{s},\mathbf{t}|\sigma) \tag{13}$$

Two probability distributions have to be modeled. On the one hand, let us assume to be uniform the segmentation distribution. On the one hand, let us assume a uniform distribution over all possible segmentations. In this context a given segmentation $\sigma \in \mathcal{S}(\mathbf{s},\mathbf{t})$, splits the source and target strings into a number of substrings (m_σ) and indeed it states the limits of each substring: $\sigma = \{\tilde{s}_1\tilde{s}_2\cdots\tilde{s}_{m_\sigma};\tilde{t}_1\tilde{t}_2\cdots\tilde{t}_{m_\sigma}\}$ where $|\tilde{s}_i| > 0$ and $|\tilde{t}_i| \geq 0$. Thus, the probability of a given bilingual pair along with a specific segmentation could be expressed under the common approach for n-gram models (Jelinek, 1997) taking bilingual phrases as language-unit (instead of the typical monolingual running words):

$$P(\mathbf{s},\mathbf{t}|\sigma) \simeq \prod_{i=1}^{m_\sigma} P((\tilde{s}_i,\tilde{t}_i)|(\tilde{s}_{i-n+1},\tilde{t}_{i-n+1}),\ldots,(\tilde{s}_{i-1},\tilde{t}_{i-1})) \tag{14}$$

Once the decoding technique under phrase-based SFST approach has been described, we are going to proceed to describe the learning procedure. This methodology requires to have the training corpus segmented in advance (the segmentation technique used in this work will be presented in section 4.2). Then an SFST is learnt taking the segments as vocabulary units instead of running words. This transducer would be an acceptor of segmented strings, that is, strings built by means of sequences of those segments. Nevertheless, at decoding time the source sentence is not built in terms of segments but in terms of running words. Therefore, the previous model has to be generalized so that the analysis is performed in terms of words. As a segment can be unambiguously converted into words, it is possible to define an equivalence relationship between each segment and a left-to-right finite-state model at word level. Thus, we just proceeded to integrate those word-by-word models within each edge of the phrase based transducer. In brief, the phrase-based transducers we put forward in (González and Casacuberta, 2007; Pérez et al., 2007), kept up with a monotonic approach of the widely used statistical approaches (Koehn et al., 2003).

The final transducer is more restrictive than the one built taking words as units. Bear in mind that with a given history and any input symbol there always exists at least one path in the smoothed transducer with non-zero probability compatible with that symbol. In a smoothed SFST based on words, given a history any word is likely to appear, and therefore, any sequence of words will have a non-zero probability. In the same way, in a smoothed phrase-based SFST any phrase amongst those in the vocabulary is likely to appear,

but not any string of words, only the strings compatible with the phrase-based model will have a non-zero probability.

4 On the use of linguistics within statistical models

Translation between Spanish and Basque presents specific problems, mainly due to the high differences in both morphology and syntax (as mentioned in appendix A). In this work two methods based on linguistic knowledge integration within finite-state transducers were explored.

4.1 Morphological approach

This method attempts at analyzing the morphology of the morphologically richer language (Basque, in this case) and split the words into basic lexically meaningful units (lexemes and morphemes). On account of this, each unit could have a closer counterpart in the other language (Spanish, in this case), and then, a more accurate translations could be obtained. In (Hirsimäki et al., 2006) an unsupervised algorithm for discovering word fragments was presented and evaluated in speech recognition for Finnish. The language model based on those units offered remarkable improvements over the traditional word-based language model. In (Goldwater and McClosky, 2005) an attempt was made at homogenizing Czech and English on the basis of morphological characteristics in order to improve statistical translation models. Bearing in mind that Finish and Czech are highly inflected languages, it may be of interest to explore similar approaches for speech translation into Basque. In fact, previous works (Agirre et al., 2006; Labaka et al., 2007) have studied the benefits of including morphological information in order to improve statistical alignments between Spanish and Basque.

Since the morphemes are the atomic meaningful units no further divisible into smaller units, the statistical translation model could be learnt in terms of such simple units instead of running words. In fact, the number of different morphemes in the studied Basque corpus is around 40% smaller than the number of running words. Then it might be expected that more significant statistics could be collected amongst morphemes rather than amongst words.

To be more specific, the approach was conceived to work in two sequential steps, being each one implemented with an SFST. The first one would render the meaning from the source to the target without taking the grammatical cases into account. That is, this first step would provide a rough representation of the target string. The second step would select the appropriate word-forms

in such a way that the previous string made sense in the target grammar. The aim of this approach is to decompose the translation process into two problems simpler than the original one. In this case, morphology comes to the aid of meaning rendering.

For the *first step*, the SFST was built taking source (Spanish) words as input vocabulary and restricting the target (Basque) vocabulary to basic morphological units with lexical meaning. That is, first of all we aimed at transferring the meaning by assembling a translation model from Spanish into lemmatized Basque. In short, all inflected variants of a word-form share the same lemma. Since lemmatization plays the same role as categorization, other kind categories might also be useful in what comes to rendering the meaning. This process is expressed through eq. (15), where $\mathbf{c}$ stands for the involved sequence of lemmas (or categories, in general). In this work the joint probability involved is modelled with an SFST $\mathcal{T}_1$, that is, $P(\mathbf{c}, \mathbf{s}) \approx P_{\mathcal{T}_1}(\mathbf{c}, \mathbf{s})$.

$$\widehat{\mathbf{c}} = \arg\max_{\mathbf{c}} P(\mathbf{c}|\mathbf{s}) = \arg\max_{\mathbf{c}} P(\mathbf{c}, \mathbf{s}) \tag{15}$$

The first SFST was experimentally proved to be an accurate statistically motivated transfer model, but it still needs for another stage to seek the proper declension cases of the given lemmas.

For the *second step*, a lexical choice has to be made on the basis of both the source string and the categorized target, as expressed through eq. (16).

$$\widehat{\mathbf{t}} = \arg\max_{\mathbf{t}} P(\mathbf{t}|\mathbf{s}, \widehat{\mathbf{c}}) \tag{16}$$

Given the lemmatized representation obtained in the fist step, we did not make further use of the source string. Thus, eq. (16) was approximated by the following one. Once again, the joint probability was modeled with an SFST $\mathcal{T}_2$, that is, $P(\mathbf{t}, \widehat{\mathbf{c}}) \approx P_{\mathcal{T}_2}(\mathbf{t}, \widehat{\mathbf{c}})$

$$\widehat{\mathbf{t}} \approx \arg\max_{\mathbf{t}} P(\mathbf{t}|\widehat{\mathbf{c}}) = \arg\max_{\mathbf{t}} P(\mathbf{t}, \widehat{\mathbf{c}}) \tag{17}$$

It has to be noted that the underlying morphologic analysis was performed by Ametzagaiña group [1] a non-profit organization working on I+D. For further details on this approach and exhaustive experimental results turn to (Pérez et al., 2006a). On the following, this morphology-based approach will be referred to as MB.

Example: the steps involved in the MB approach are here summarized with a sentence extracted from the corpus.

[1] `http://ametza.com`

- Given the source sentence (in Spanish):
  "*por la mañana los cielos permanecerán muy nubosos*"
  whose counterpart in English is
  "*during the morning skies will remain mostly cloudy*"
- The first transducer, $\mathcal{T}_1$, gives as a result the translation of the source sentence into lemmatised Basque:
  "*goizalde zeru oso hodei mantendu izan*"
  word by word, an equivalent in English would be
  "*morning sky remain most cloud* "
- The second stage has to find the proper word-forms for the given lemmas. For the task under consideration there are 4 words in the training corpus that share the lemma *"goizalde"* (the first token within the lemmatised Basque string), to be precise: *goizalde*, *goizaldean*, *goizaldera* and *goizaldez*. That is, there are 4 words compatible with that lemma. The selection of the most likely word-form is carried out taking into account both the whole lemmatised sentence and an output language model (modelled by $\mathcal{T}_2$).
  "*goizaldean zerua oso hodeitsu mantenduko da*"

4.2 Approach based on syntactic phrases

The general framework over phrase-based transducers (described in section 3.4) entails a segmentation technique. In our case the segmentation was syntactically motivated and accomplished by Ametzagaiña group. In this section the phrase extraction methodology is briefly described. The procedure involves a parser for each language under study. That is, the phrases are selected according to monolingual criteria. Even though for Spanish there are open-source analysers (Carreras et al., 2004), for Basque, however, there is none and it had to be developed.

The phrase identification was linguistically motivated and automatically accomplished following the steps listed below. Let us note that this is a domain independent procedure:

- First, a morphologic parsing allows to assign either one or more tags to each word within the corpus. These tags include information about linguistic categories such as declension case, number, definiteness, tense, etc. Besides, the stem and the morphemes are identified. In both Spanish and Basque they were defined around 45 classes with several sub-classes each. Both languages share the majority of them. Nevertheless, there are some tags that are just related to one of the two languages, *ergative* [2] case for instance,

[2] *Ergative:* a case of nouns in a few languages (e.g., Basque and Eskimo) that identifies the subject of a transitive verb.

which does not exist in Spanish. Apart from this, let us emphasize that in this step ambiguity is not removed. As a result, more than one tag-set can be assigned to each word.

- Next, ambiguities have to be removed. For each word, the most likely category-set is chosen so that the words of the sentence shared syntactically compatible categories according to predefined rules for each language. Around 15 rules were defined for each language over their corresponding category set so as to get a high coverage of word-forms. The order in which these items might appear within a specific syntactic function is described as well.
- Finally, linguistic phrases can be identified under an elementary criteria: recursively group all the words that share the same syntactic function whenever the frequency of that group (phrase) in the corpus exceeds a threshold. On the first iteration of this algorithm just noun and verb phrases are distinguished, that is the parsing is quite shallow. Then, regular expressions (such as dates) are also taken into account so as to help the selection of an error-free category. As the iterations go ahead, more and more precise tokens are identified, such as composed stems, periphrastic verbs, etc. along with their syntactic role within the sentence they belong to.

Example: Throughout this example the aforementioned steps are going to be developed taking the following Spanish sentence as input: *"las rachas de viento pueden superar una velocidad de 90 Km/h"* (note that this is the example shown in Fig. 1). In the first step the words are analyzed separately without taking either left or right context into account. This analysis entails a source of ambiguities. For instance, the word *"viento"* in the sentence was given two tag-sets as summarized in Fig. 3. According to these results the word *"viento"* has two different lemmas, *"viento"* and *"ventar"*. The grammar category associated to the former is noun, gender male and number singular. While the latter, *"ventar"*, is a verb in indicative mode and present tense associated with the first person.

```
T1:Lemma=viento[Gram.:noun,Gender:m,Number:singular]

T2:Lemma=ventar[Gram.:verb,Mode:indicative,Tense:present,Person:1]
```

Fig. 3. The first step for the syntactically motivated segmentation resorts to a morphological parsing of individual words in order to categorize them. This step over the word *"viento"* gave as a result two different tag-sets.

The second step takes as input all the category-sets related to all the words within a sentence and by a set of rules defined over the tags, a segmentation is obtained. The output of this procedure over the sentences shown in Fig. 1 leads to the segmentation of Fig. 4. The phrases in one language usually have their counterpart in the other language as a phrase as well, not in a monotonic fashion, however.

	Phrase	Clasification
Spanish	las rachas de viento	Noun Phrase
	pueden superar	Verb Phrase
	una velocidad	Noun Phrase
	de 90 Km/h	Prep. Segment: de
Basque	haize boladek	Noun Phrase: ergative case
	90 Km/h-ko abiadura	Noun Phrase: absolutive case
	gaindi dezakete	Verb Phrase: 0

Fig. 4. Example of segmentation and related information. The classification of phrases provides a natural segmentation. In addition, it might help to the alignments, since usually, the phrases playing the same role represent related entities in both languages. In this example, the verb phrase as a whole, even being in different places for each language, are tightly connected.

A manual inspection shown that more realistic correspondences could be described between these linguistically motivated phrases rather than between running words. For instance, the phrases "las rachas de viento" and "haize boladek", in Spanish and Basque respectively, are both noun phrases and mean *the strong gusts*. The same happens with the verb phrase, both the Spanish and the Basque phrases are translation pairs (meaning *may exceed*). While in the last case, a couple of phrases in Spanish refers to a single one in Basque, being "una velocidad de 90 Km/h" the counterpart of "90 Km/h-ko abiadura" (meaning *a speed of 90 Km/h*).

As a consequence, the alignment unit was not word-form any longer, but linguistic token. Thus, Fig. 5 shows the statistical alignments at phrase level over the example of the Fig. 1. On the following, this phrase-based approach will be referred to as PB.

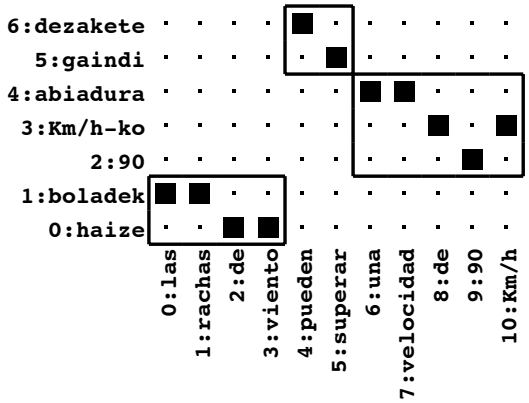


Fig. 5. Alignments at phrase level between Basque and Spanish.

5 Task and corpus

METEUS is a weather forecast corpus (Pérez et al., 2006b) composed by 28 months of daily weather forecast reports in Spanish and Basque. These reports were picked from those published in Internet by the Basque Institute of Meteorology [3]. Initially, the bilingual corpus was aligned at paragraph level and it consisted of 3 865 paragraphs of 54 words on average. RECALIGN, a non-supervised statistical technique (Nevado and Casacuberta, 2004) made the alignments at sentence level possible. A hundred paragraphs, randomly chosen, were evaluated by experts and agreed on the correctness of all of them. From then onwards we worked with the corpus aligned at sentence level.

After the segmentation process, the corpus was randomly divided into two disjoint sets, referred to as *Training* and *Test-1* in Table 1. The former consisted of 14 615 sentences (summing up 191 156 running words) and the latter of 1 500 sentences. In addition, a sub-test of 500 different sentences was extracted from Test-1 under one restriction: it was not allowed to be any coincidence with the training set. The latter test, *Test-2*, was recorded for speech translation purposes by 36 bilingual speakers. Each sentence was uttered by at least 3 speakers (male and female) resulting in 1 800 utterances, or equivalently in terms of time, 3.5 hours of audio signal recorded at 16 KHz.

Due to the nature of the task, the Test-1 set is quite repetitive, and so is the training set. Test-1 is statistically representative of the task whereas Test-2 has shown to be more difficult as can be derived from Table 1. On average, the length of the sentences in both the Training and Test-1 sets is around 13 words per sentence, whereas it is around 17 for the Test-2 set. Perplexity is another indicative feature that shows the bias in Test-2, which is relatively much higher than the one for Test-1. As a result, Test-2 might help to test the robustness of the models under situations biased against the training set.

The most remarkable feature of the corpus may lie in what the size of the vocabulary concerns. Notice that the size of the Basque vocabulary is 38% bigger than the Spanish one due to its inflected nature. Even though this is a medium difficulty task, there is a high data sparseness. As a matter of fact, notice that 26% of the words in the Basque vocabulary have just seen once all over the corpus (see singletons in Table 1). In addition, the 66% of the singletons have to do with different lemmas.

[3] `http://www.euskalmet.net`

4

		Spanish	Basque
Training	Pair of sentences	14 615	
	Different pairs	8 445	
	Running words	191 156	187 195
	Vocabulary	702	1 135
	Singletons	162	302
	Average length	13.1	12.8
Test-1	Pair of sentences	1 500	
	Different pairs	1 173	
	Average length	12.6	12.4
	Perplexity (3-grams)	3.6	4.3
Test-2	Pair of sentences	1 800	
	Different pairs	500	
	Average length	17.4	16.5
	Perplexity (3-grams)	4.8	6.7

Table 1
Main features of METEUS corpus. There is a training set and two test sets. Test-1 was randomly selected from the task, while Test-2 consists of 500 different and training independent sentences with both text and speech representations.

6 Experimental results

6.1 Implementation issues

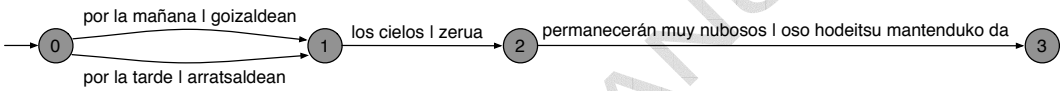
For speech translation purposes, the underlying recognizer used in this work is our own continuous-speech recognition system, which implements stochastic finite-state models at all levels: syntactic, lexical and acoustic-phonetic. These models are integrated on the fly at decoding time within the aforementioned integrated architecture, which is illustrated through Fig. 6. The integration on the fly has shown to be an efficient technique in speech-recognition framework (Caseiro and Trancoso, 2006) and it can be implemented in the same way for speech translation.

Instead of the usual *language model*, we made use of the SFST itself (Fig. 6(a)), which had the syntactic structure provided by a k-testable in the strict sense model, with k=3 and back-off smoothing. Let us remark that our aim in this work is to develop different approaches of this model and thus the remaining

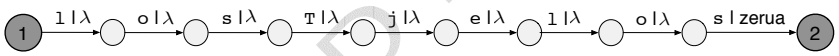
models were not changed.

The *lexical model* consisted of the extended tokens $(\tilde{s}_k, \tilde{t}_k)$ of the SFST instead of the running words involved in a typical language model. The phonetic transcription for each extended token was automatically obtained on the basis of the input projection of each unit $(\tilde{s}_k)$, that is, the Spanish vocabulary (or sub-strings) in this case. The model itself consists of the concatenation of the phonemes involved in a left-to-right finite-state model, as shown in Fig.6(b).

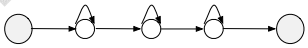
Each phone-like unit was modeled by a typical left to right non-skipping self-loop three-state continuous *hidden Markov model* (referred to as HMM in Fig. 6(c)), with 32 Gaussians per state and acoustic representation. The speech signal database was parametrized into 12 Mel-frequency cepstral coefficients (MFCCs) with delta (ΔMFCC) and acceleration (Δ^2MFCC) coefficients, energy and delta-energy (E, ΔE), so four acoustic representations were defined. A phonetically-balanced Spanish database, called Albayzin (Moreno et al., 1993), was used to train these models.



(a) Finite-state transducer.



(b) On the fly integration of the lexical model.



(c) HMM.

Fig. 6. Speech translation with integrated architecture involves on the fly Integration of several knowledge sources within a single finite-state network. (a) Finite-state transducer built on the basis of the segmentation of Fig. 2. (b) The lexical model consists on the phonetic transcription of the input substring by means of a left-to-right topology (SAMPA is used here). (c) Phone-like units are modeled by typical three-state continuous hidden Markov models (HMMs).

6.2 Evaluation

Three models were trained from the training data: the classical word-based (WB) SFST (introduced in section 3), the morphology-based (MB) approach (see section 4.1) and the phrase-based (PB) approach (see sections 3.4 and 4.2).

All the models were built under the same k=3 k-TSS topology, where the maximum length of the history is 2, as in a 3-gram language model (Jelinek, 1997). No *pre-* or *post-* edit processing was carried out (neither on idioms or numbers or names etc.). Our purpose was to compare the three models under the same circumstances and study their performance for both text and speech translation. The systems were assessed under the commonly used automatic evaluation measures: *bilingual evaluation understudy* (BLEU) (Papineni et al., 2002), NIST (Doddington, 2002), *word error rate* (WER) and *position-independent error rate* (PER).

Text translation results associated with the two test-sets described in section 5 are shown in Table 2. As it was expected, the training independent test set (Test-2) reports worse results than Test-1. All in all, phrase-based (PB) approach outperformed both morpho-syntactic (MB) and word-based (WB) approach for both test sets. An additional experiment was carried out with Moses free toolkit (Koehn et al., 2006), the state-of-the art in phrase-based translation models, taking at random 14 000 pairs from the training set for training purposes and the remaining 615 for tuning. The system was tested with Test-1 set. The obtained results made no statistical difference with respect to those obtained with the phrase-based SFST approach proposed here (specifically a BLEU score of 55.9 was obtained).

		WB	**MB**	**PB**
Text Translation Test-1	BLEU	57.9	60.3	66.1
	NIST	7.6	7.7	8.2
	WER	32.8	31.4	27.6
	PER	27.7	26.6	22.3
Text Translation Test-2	BLEU	41.2	41.6	44.6
	NIST	6.1	6.0	6.4
	WER	47.5	48.0	46.9
	PER	39.4	40.4	38.1

Table 2
Text translation results with the three models described, namely, word-to-word model (WB), morphology-based approach (MB) and syntactically motivated phrase-based approach (PB).

Speech translation results, shown in Table 3, were carried out with both decoupled and integrated architectures using the three models under study. For speech translation, as well as for text translation, syntactically motivated phrase-based SFST approach (PB) turned out to report the best performance. Morphologic knowledge in MB SFST approach, however, just provided slightly improvements over the classical approach (WB).

Comparing the two architectures studied for speech translation, in these experiments the integrated architecture has proved to offer slightly better performance than the decoupled under all the approaches explored (WB, MB and PB). In addition, the integrated architecture is even faster than the decoupled one, since the former occurs in a single decoding stage. These results obtained for a real task and natural corpus agreed with the conclusions extracted in (Casacuberta and Vidal, 2007).

Even though our interest focuses on comparing the performance of different translation models, speech recognition rate is also reported since it plays an important role. Notice that with the decoupled architecture the output of the speech recognition system was the input for the text-translation systems. The speech recognition system was in the same circumstances (3-TSS language model instead of the translation model). Under the integrated architecture, however, speech transcription and translation is simultaneously produced, for this reason each SFST approach offered different recognition results. Slightly better recognition results are obtained with the MB-SFST than with the typical speech recognition system. Another odd fact extracted from the Table 3 is that the PB model offers the best translation results but the worst recognition rate.

		Integrated Arch.			Decoupled Arch.		
		WB	**MB**	**PB**	**WB**	**MB**	**PB**
Speech Recognition	WER	8.3	7.3	9.6	7.9		
Speech Translation Test-2	BLEU	38.5	38.9	40.9	37.4	37.8	40.8
	NIST	5.7	5.8	6.0	5.6	5.6	6.0
	WER	51.3	50.5	49.6	51.2	51.2	50.3
	PER	42.5	41.8	40.4	42.2	42.6	40.3

Table 3
Speech translation results of Test-2 with different approaches, namely, word-to-word model (WB), morphological approach (MB) and syntactically motivated phrase-based approach (PB). Integrated architecture give as a result both the transcription and the translation of speech in a unique decoding step. Decoupled architecture entails two independent steps: first speech recognition and then text-to-text translation of recognized utterances.

Taking text and speech translation results into account (compare Test-2 of Table 2 with Table 3), it is remarkable the small differences in performance. It might be concluded that the errors reflected in speech recognition were not directly propagated into translation errors. Bear in mind that the translation model is by itself an important source of noise. Its error rate is one magnitude order higher than in the case of the ASR. Thus, the SFST might be insensitive to small deviations in the input. In addition, the model is smoothed,

which allows either correct or incorrect input strings while trying to give as a translation only well formed outputs according to the training data.

On the whole, both morphologic and syntactic knowledge introduced within statistical transducers improves the performance for both text and speech translation. Both MB and PB approaches outperform the classical one (WB). The major improvement is associated with the PB approach, were syntactic phrases were taken as translation unit.

7 Concluding remarks and future work

This work deals with Spanish into Basque speech translation, a task that, to the authors' knowledge, had never been previously faced in other works. To do so, stochastic finite-state transducers are considered. The goal of this work is to improve their performance by including linguistic knowledge. Two new approaches were proposed taking into account the specific features of the languages involved, namely high inflection of Basque and strong differences in syntax. On the one hand morphology was exploited and syntax on the other. The approaches focused on modeling rather than on decoding and represent a general methodology that may be applied to any pair of languages.

Morphologically rich languages present data sparsity on the statistics collected over running word n-grams. A lemmatized representation of the language, however, helped to render essential information about the meaning from one language to the other in the MB framework. This rough approach allowed to deal with rather reliable statistics. In the PB approach, linguistically motivated phrases were proposed as translation unit. Thus, the underlying formulation for phrase-based SFST is defined in this work as an extension of previous SFSTs. The phrases considered in the experimental results consisted of a sequence of words that played the same syntactic role within a given sentence. The long-distance alignments between Basque and Spanish were cut down with the phrase-based approach. Therefore, the SFST was able to capture those sort of relationships with more accuracy than using running-words.

Experiments were carried out under a narrow domain framework with small and repetitive vocabulary. Results over different architectures for speech translation show that linguistic knowledge helps to improve the performance of statistical translation models. The improvement of the morphological approach was not as successful as the syntactically motivated phrase-based model, where the relative improvement on BLEU score was over the 6%. Let us mention the development of linguistic resources, tools and corpus, for Basque language as another contribution of this work.

For future work, we consider that rather accurate MB approach could be reached since, in practice, the second stage involved a strong assumption as it discarded the source string. The second stage could be exploited along with a statistical dictionary from source words into lemmatized targets. In addition, other kind of categories, instead of lemmas, could be explored as well. With regard to the phrase based SFST approach, we intend to explore not only syntactically motivated phrases and categories but also statistical ones. Finally, it could be of interest to explore the combination of both morphological and syntactic approaches

A Spanish and Basque languages

Machine translation between Spanish and Basque presents multiple challenges as they differ extraordinarily. They do not have the same origin, in fact, Basque is a pre-Indo-European language of still unknown origin, whereas Spanish is amongst the Romance group of the Indo-European family of languages. Nowadays, Spanish and Basque share official status for the 2.5 million inhabitants of the autonomous community of the Basque Country (Spain). Basque is spoken beyond that region, in some areas of Navarre (Spain), Atlantic Pyrenees (France), Reno (United States of America) etc. It has to bear in mind, however, that Basque is a minority language, while Spanish holds the 4th place in the world in what the number of native speakers concerns (after Mandarin Chinese, Hindi and English). Needless to say, there are significant differences with regard to the amount of available linguistic resources, but in these last years several groups from the university and industry are playing a valuable role in developing linguistic supports for Basque language.

Basque, in contrast to Spanish, is a highly inflected language (both in nouns and verbs). For noun phrases there are 17 cases which are at the same time modified by determiners related to number alternation, and indeed, they are susceptible to throwing together in several recurrence levels. For a matter of comparison, in German there are 4 grammatical cases, 7 in Czech, and 15 in Finnish. With regard to the verbs, apart from the tense, both subject (person and number), direct object and indirect object marks (if any) are part and parcel of verbs. This issue has a significant effect on the size of the vocabulary and the repetition ratio. Typically, given a text in Basque and Spanish (or English), the first one has usually fewer running words on the whole than the latter one but the number of different words (that is, the size of the vocabulary) is normally much higher, having direct repercussions on the statistics collected over word utterances.

As depicted in Fig. 1 long distance alignments are likely to occur between Spanish and Basque. This is due to the fact that typical syntactic construction

in Basque is Subject-Objects-Verb (like Japanese) unlike Spanish (or English) where Subject-Verb-Objects construction is more common. The order of the phrases within a sentence can be changed with thematic purposes. As a matter of fact, the order of the phrases in Basque is *topic-focus*, meaning that the topic is stated first and then the focus immediately before the verb phrase. This characteristics have incidence on the statistical alignment modeling.

Acknowledgments

We would like to thank anonymous reviewers for their criticisms and suggestions.

We would also like to thank *Ametzagaiña* group (`http://www.ametza.com`), and Josu Landa, in particular, for providing us with the morpho-syntactic parse which made possible this work.

This work has been partially supported by the University of the Basque Country under grant 9/UPV 00224.310-15900/2004, by the Spanish CICYT under grant TIN2005-08660-C04-03 and by the research programme Consolider Ingenio-2010 MIPRCV (CSD2007-00018).

References

Agirre, E., de Ilarraza, A. D., Labaka, G., Sarasola, K., September 13-14 2006. Uso de información morfológica en el alineamiento español-euskara. In: Besga, C. I., nano, I. I. A. (Eds.), Actas del XXII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN). ISSN: 1135-5948. Zaragoza (España), pp. 257–264.

Alegria, I. n., de Illarraza, A. D., Lersundi, M., Mayor, A., Sarasola, K., 2007. Transfer-based mt from spanish into basque: Reusability, standardization and open source. In: Lecture Notes in Computer Science. Vol. 4394. Springer, pp. 374–384.

Bangalore, S., Riccardi, G., 2001. A finite-state approach to machine translation. In: Proceedings of the 2nd meeting of the North American Chapter of the Association for Computational Linguistics (NAACL). Sapporo, Japan, pp. 40–47.

Bangalore, S., Riccardi, G., 2002. Stochastic finite-state models for spoken language machine translation. Machine Translation 17, 165–184(20).

Bertoldi, N., Zens, R., Federico, M., 2007. Speech translation by confusion network decoding. In: Proceedings of ICASSP. Honolulu, HA.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Mercer, R. L., 1993.

The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics 19 (2), 263–311.

Carreras, X., Chao, I., Padró, L., Padró, M., 2004. Freeling: An open-source suite of language analyzers. In: In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal, `http://garraf.epsevg.upc.es/freeling`.

Casacuberta, F., 2000. Inference of finite-state transducers by using regular grammars and morphisms. In: Oliveira, A. (Ed.), Grammatical Inference: Algorithms and Applications. Vol. 1891 of Lecture Notes in Computer Science. Springer-Verlag, pp. 1–14, 5th International Colloquium Grammatical Inference -ICGI2000-. Lisboa. Portugal. Septiembre.

Casacuberta, F., de la Higuera, C., 1999. Linguistic decoding is a difficult computational problem. Pattern Recognition Letters 20, 813–821.

Casacuberta, F., Ney, H., Och, F. J., Vidal, E., Vilar, J. M., Barrachina, S., García-Varea, I., Llorens, D., Martínez, C., Molau, S., Nevado, F., Pastor, M., Picó, D., Sanchis, A., Tillmann, C., Jan. 2004. Some approaches to statistical and finite-state speech-to-speech translation. Computer Speech and Language 18, 25–47.

Casacuberta, F., Vidal, E., 2004. Machine translation with inferred stochastic finite-state transducers. Computational Linguistics 30 (2), 205–225.

Casacuberta, F., Vidal, E., 2007. Learning finite-state models for machine translation. Machine Learning 66 (1), 69–91.

Caseiro, D., Trancoso, I., December 2001. Transducer composition for on-the-fly lexicon and language model integration. In: Proceedings ASRU'2001 - IEEE Automatic Speech Recognition and Understanding Workshop. Madonna di Campiglio, Italy, pp. 393–396.

Caseiro, D., Trancoso, I., 2006. A specialized on-the-fly algorithm for lexicon and language model composition. IEEE Transactions on Audio, Speech & Language Processing 14 (4), 1281–1291.

Collins, M., Koehn, P., Kucerova, I., June 2005. Clause restructuring for statistical machine translation. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05). Association for Computational Linguistics, Ann Arbor, Michigan, pp. 531–540.

Corbí-Bellot, A. M., Forcada, M. L., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Alegria, I., Mayor, A., Sarasola, K., May 2005. An open-source shallow-transfer machine translation engine for the romance languages of spain. In: Proceedings of the Tenth Conference of the European Associtation for Machine Translation. Budapest, Hungary, pp. 79–86.

de Gispert, A., Mariño, J. B., Crego, J. M., March 2006. Linguistic knowledge in statistical phrase-based word alignment. Natural Language Engineering 12 (01), 91–108.

Doddington, G., 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. Proceedings of the second international conference on Human Language Technology Research, 138–145.

García, P., Vidal, E., 1990. Inference of k-testable languages in the strict sense and application to syntactic pattern recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (9), 920–925.

Goldwater, S., McClosky, D., 2005. Improving statistical mt through morphological analysis. In: HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Morristown, NJ, USA, pp. 676–683.

González, J., Casacuberta, F., September 14-16 2007. Phrase-based finite state models. In: Proceedings of the 6th International Workshop on Finite State Methods and Natural Language Processing (FSMNLP). Potsdam (Germany).

Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., Pylkkönen, J., 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. Computer Speech and Language 20, 515–541.

Jelinek, F., 1997. Statistical Methods for Speech Recognition. MIT Press, Cambridge, MA.

Knight, K., Al-Onaizan, Y., 1998. Translation with finite-state devices. Vol. 1529 of Lecture Notes in Computer Science. Springer-Verlag, pp. 421–437.

Koehn, P., Federico, M., Shen, W., Bertoldi, N., Bojar, O., Callison-Burch, C., Cowan, B., Dyer, C., Hoang, H., Zens, R., Constantin, A., Moran, C., Herbst, E., 2006. Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding. Tech. rep., Johns Hopkins University, Center for Speech and Language Processing.

Koehn, P., Och, F. J., Marcu, D., May 2003. Statistical phrase-based translation. In: Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL). Edmonton, Canada, pp. 48–54.

Kumar, S., Deng, Y., Byrne, W., December 2005. A weighted finite state transducer translation template model for statistical machine translation. Natural Language Engineering 12, 35–75.

Labaka, G., Stroppa, N., Way, A., Sarasola, K., 2007. Comparing rule-based and data-driven approaches to spanish-to-basque machine translation. Copenhagen.

Matusov, E., Kanthak, S., Ney, H., May 2005. Efficient statistical machine translation with constrained reordering. In: Proceedings of EAMT 2005 (10th Annual Conference of the European Association for Machine Translation). Budapest, Hungary.

Mehryar Mohri, F. C. N. P., Riley, M. D., 2003. AT&T FSM LibraryTM Finite-State Machine Library. `http://www.research.att.com/sw/tools/fsm`.

Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mario, J. B., Nadeu, C., 1993. Albayzin speech database: Design of the phonetic corpus. In: Proc. of the European Conference on Speech Communications and Technology (EUROSPEECH). Berlín, Germany.

Nevado, F., Casacuberta, F., November 2004. Bilingual corpora segmentation using bilingual recursive alignments. In: Actas de las III Jornadas en Tecnologías del Habla, 3JTH. Rthabla, Valencia.

Nießen, S., Ney, H., Sep. 2001. Morpho-syntactic analysis for reordering in statistical machine translation. In: Proceedings of the Machine Translation Summit VIII. Santiago de Compostela, Spain, pp. 247–252.

Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., Radev, D., Jul. 2003. Syntax for statistical machine translation, final report, JHU 2003 summer workshop. `http://www.clsp.jhu.edu/ws2003/groups/translate`.

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

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

Pereira, F. C., Riley, M. D., 1997. Speech Recognition by Composition of Weighted Finite Automata. In: Roche, E., Schabes, Y. (Eds.), Finite-State Language Processing. Language, Speech and Communication series. The MIT Press, Cambridge, Massachusetts, pp. 431–453.

Pérez, A., Torres, M. I., Casacuberta, F., 23-25 August 2006a. Towards the improvement of statistical translation models using linguistic features. In: Proceedings of the FinTAL - 5th International Conference on Natural Language Processing. Lecture Notes in Computer Science 4139. Turku, Finland, pp. 716–725.

Pérez, A., Torres, M. I., Casacuberta, F., April 15-20 2007. Speech translation with phrase based stochastic finite-state transducers. In: Proceedings of the IEEE 32nd International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007). Vol. IV. IEEE, Honolulu, Hawaii USA, pp. 113–116.

Pérez, A., Torres, M. I., Casacuberta, F., Guijarrubia, V., 2006b. A Spanish-Basque weather forecast corpus for probabilistic speech translation. In: Proceedings of the 5th Workshop on Speech and Language Technology for Minority Languages (SALTMIL). Genoa, Italy.

Quan, V., Federico, M., Cettolo, M., 2005. Integrated n-best re-ranking for spoken language translation. Proceedings of Interspeech 2005, 3181–3184.

Saleem, S., Jou, S., Vogel, S., Schultz, T., 2004. Using Word Lattice Information for a Tighter Coupling in Speech Translation Systems. Proc. Int. Conf. on Spoken Language Processing, 41–44.

Sarikaya, R., Zhou, B., Povey, D., Afify, M., Gao, Y., April 15-20 2007. The impact of ASR on speech-to-speech translation performance. In: Proceedings of the 32nd International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007). IEEE, Honolulu, Hawaii USA.

Torres, M. I., Varona, A., 2001. k-TSS language models in speech recognition systems. Computer Speech and Language 15 (2), 127–149.

Vidal, E., Apr. 1997. Finite-state speech-to-speech translation. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 1. Munich, Germany, pp. 111–114.

Vidal, E., Thollard, F., C. de la Higuera, F. C., Carrasco, R., 2005a. Probabilistic finite-state machines - part I. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 27 (7), 1013–1025.

Vidal, E., Thollard, F., C. de la Higuera, F. C., Carrasco, R., 2005b. Probabilistic finite-state machines - part II. IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI) 27 (7), 1025–1039.

Zhou, B., Chen, S., Gao, Y., 2005. Constrained Phrase-based Translation Using Weighted Finite State Transducer. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 1. pp. 1017–1020.