# MONOTONE STATISTICAL TRANSLATION USING WORD GROUPS

**Jesús Tomás[1], Francisco Casacuberta[2]**

[1] Escuela Politécnica Superior de Gandia
Universidad Politécnica de Valencia.
46730 Gandia, SPAIN
jtomas@upv.es

[2] Instituto Tecnològic d'Informàtica
Universidad Politécnica de Valencia.
46071 Valencia, SPAIN
fcn@iti.upv.es

**Abstract**

A new system for statistical natural language translation for languages with similar grammar is introduced. Specifically, it can be used with Romanic Languages, such as French, Spanish or Catalan. The statistical translation uses two sources of information: a language model and a translation model. The language model used is a standard trigram model. A new approach is defined in the translation model. The two main properties of the translation model are: the translation probabilities are computed between groups of words and the alignment between those groups is monotone. That is, the order between the word groups in the source sentence is conserved in the target sentence. Once, the translation model has been defined, we present an algorithm to infer its parameters from training samples. The translation process is carried out with an efficient algorithm based on stack-decoding. Finally, we present some translation results from Catalan to Spanish and compare our model with other conventional models.

## Keywords

Statistical machine translation, words segments, search.

## Introduction

Statistical methods have proven to be valuable in tasks such as automatic speech recognition and natural language processing (Bahl, 1983), and they present a new opportunity for automatic translation. However, current results in automatic translation are far from satisfactory (Onaizan et al., 1999; Tomás & Casacuberta 1999).

We present an approach that attempts to apply Statistical Machine Translation two similar languages such as Spanish and Catalan. We think that approach is also applicable to other pairs of Romanic languages (French, Italian, Portuguese, etc.). Romanic languages allow for great flexibility in the word ordering a sentence. This feature is especially found in Spanish where almost any order of the sentence components can be correct. For example, there are dozens of ways in which the fist sentence can be expressed:

Se requerirá una acción de la Comunidad para la puesta en práctica
Una acción se requerirá de la Comunidad para la puesta en práctica
Se requerirá de la Comunidad una acción para la puesta en práctica
De la Comunidad se requerirá una acción para la puesta en práctica
Para la puesta en práctica se requerirá una acción de la Comunidad
...

It is true that the order in which a Spanish sentence is written can have slight differences in meaning, mainly by emphasizing certain elements in the sentence. Nevertheless, these nuances are not significant to the task we are attempting to carry out.

The translation model we present stars from the following assumption "It is possible to translate a sentence written in one Romanic language to another Romanic language by translating word groups in a monotone way.

To determine whether this statement was meaningful, we performed the following test: A set of 400 sentences was selected at random from different sources in three different Romanic languages- Spanish, Catalan and French. A human translator attempted to do monotone translations groups of five, or less than five, words in order to determine whether it was possible to do the translation in a monotone way.

The results obtained indicate that monotone translations for Catalan to Spanish, Spanish to Catalan, and French to Spanish are always correct. However, monotone translations from Spanish to French are not always possible. The results showed that 10 sentences out of 400 sentences could not be translated with complete accuracy (containing small grammatical errors). Conclusion: Despite the small number of not completely correct translated sentence in French, translating the languages tested using monotone translation is feasible.

We are interested in taking advantage of this property in our translation model. The two main contributions are: the translation probabilities are calculated between groups of words and the alignment is monotone constrained. Other works use similar approaches. In (Vogel et al., 1996) the concept of monotone alignment is used to improve the search. Alignment models based on word groups are introduced in (Epstein et al., 1996; Och et al., 1999; Wang & Waibel, 1997). But, unlike our approach, the lexical model is based on word-to-word correspondence.

The organization of the paper is as follows. First we review the statistical approach to machine translation. Second, we introduce our new translation model. Then, we show the training procedure and propose a search strategy based on stack decoding. Finally, we report some experimental results and compare our models with other conventional models. The system was tested translating a newspaper from to Catalan into Spanish.

## Stochastic Translation.

The goal of statistical translation is to translate a given source language sentence $\mathbf{f} = \mathbf{f}_1 \ldots \mathbf{f}_{|\mathbf{f}|}$, to a target sentence $\mathbf{e} = \mathbf{e}_1 \ldots \mathbf{e}_{|\mathbf{e}|}$. The methodology used (Brown et al, 1993), is based on the definition of a function $\Pr(\mathbf{e}|\mathbf{f})$ that returns the probability of translating the input sentence **f** into the output sentence **e**. Once this function is estimated, the problem can be formulated to compute a sentence **e** that maximizes the probability $\Pr(\mathbf{e}|\mathbf{f})$ for a given **f**. Using Bayes' theorem, we can write:

$$\Pr(\mathbf{e}\,|\,\mathbf{f}) = \frac{\Pr(\mathbf{e})\Pr(\mathbf{f}\,|\,\mathbf{e})}{\Pr(\mathbf{f})} \quad (1)$$

And therefore, statistical translation can be presented as:

$$\mathbf{e}' = \arg\max_{\mathbf{e}} \Pr(\mathbf{e})\Pr(\mathbf{f}\,|\,\mathbf{e}) \quad (2)$$

Equation (2) summarizes the three following matters to be solved:

- An output language model is needed to distinguish valid sentences from invalid sentences in the target language, Pr(**e**).
- A translation model, Pr(**e**|**f**) must be defined.
- An algorithm must be designed to search for the sentence **e** that maximizes this product. The search must be fast and efficient, even at the risk of a suboptimal result.

This approach is very similar to the one used in speech recognition, so we will use many of the techniques which have been developed in this area to solve text translation problems (Jelinek, 1976).

| | | | | | |
|---|---|---|---|---|---|
| Se requerirá | una acción | de la | Comunidad | para la | puesta en práctica |
| É necessária | uma acção | por parte da | Comunidade | para pôr | plenamente em prática |
| Sarà necessaria | un'azione | della | Comunità | per dare | piena attuazione |
| Une action | est nécessaire | au niveau | communautaire | afin de | mettre pleinementen œuvre |
| Action | is required | by the | Community | in order to | implement fully |

Figure 1: Sentence written in five different languages

## Translation Model

We now propose our translation model. The principal innovation of the model is that we try to calculate the translation probabilities of word groups rather than of only single words. Figure 1 shows the same sentence written in five different languages.

As can be seen from this example, we join groups of words that are translated together in a natural way. The other property of our translation model is that the alignment between the word groups is monotone constrained. In the example, we can observe how the first three sentences are monotone translated.

To reduce the model's parameters, we do not consider all possible groups of adjacent words as a word group. Previously, we use an algorithm to identify word groups in a parallel corpus (see below section). Let $\hat{E}$ be the union of the set of word groups obtained as the output of this algorithm, the individual words and the empty word ($e_0$) all of which are from the target language. Each element of $\hat{E}$, $\hat{e}$, can be expressed as a sequence of words $\hat{e}_1,\ldots,\hat{e}_{|\hat{e}|}$. The sentence **e** can be expressed as a concatenation of elements of $\hat{E}$. We denote $\hat{\mathbf{e}} = \hat{\mathbf{e}}_1\,\hat{\mathbf{e}}_2 \ldots \hat{\mathbf{e}}_{|\hat{\mathbf{e}}|}$ ($\hat{e}_i \in \hat{E}$), as a possible breakdown of **e** using word groups of $\hat{E}$.

$$\hat{\mathbf{e}} = \hat{\mathbf{e}}_1\,\hat{\mathbf{e}}_2 \ldots \hat{\mathbf{e}}_{|\hat{\mathbf{e}}|} = \hat{\mathbf{e}}_{1_1}\ldots\hat{\mathbf{e}}_{1_{|\hat{\mathbf{e}}_1|}}\hat{\mathbf{e}}_{2_1}\ldots\hat{\mathbf{e}}_{2_{|\hat{\mathbf{e}}_2|}} \ldots \hat{\mathbf{e}}_{|\hat{\mathbf{e}}|_1}\ldots\hat{\mathbf{e}}_{|\hat{\mathbf{e}}|_{|\hat{\mathbf{e}}_{|\hat{\mathbf{e}}|}|}} = \mathbf{e}_1\,\mathbf{e}_2 \ldots \mathbf{e}_{|\mathbf{e}|} = \mathbf{e} \quad (3)$$

Similar definitions can be made with $\hat{F}$ and $\hat{\mathbf{f}}$ in the source language.

Now we can estimate the translation probability as the addition of all possible alignments between **e** and **f**. We define an alignment, **a**, between **e** and **f**, as the tuple {$\hat{\mathbf{e}}$, $\hat{\mathbf{f}}$ } with $\hat{\mathbf{e}}$ being a possible breakdown of **e** in word groups of $\hat{E}$; and with $\hat{\mathbf{f}}$ being a possible breakdown of **f** in word groups of $\hat{F}$; with the restriction that the numbers of word groups in $\hat{\mathbf{e}}$ should be identical to $\hat{\mathbf{f}}$. We assume the alignment is monotone, thus, the word group $\hat{\mathbf{e}}_i$ is aligned with the word group $\hat{\mathbf{f}}_i$, with i =1…$|\hat{\mathbf{e}}|$.

$$\Pr(\mathbf{f}\,|\,\mathbf{e}) = \sum_{\mathbf{a}} \Pr(\mathbf{f},\mathbf{a}\,|\,\mathbf{e}) = \sum_{\hat{\mathbf{e}}:\hat{\mathbf{e}}=\mathbf{e}} \sum_{\substack{\hat{\mathbf{f}}:\hat{\mathbf{f}}=\mathbf{f};\\ |\hat{\mathbf{e}}|=|\hat{\mathbf{f}}|}} \Pr(\hat{\mathbf{f}}\,|\,\hat{\mathbf{e}}) \quad (4)$$

To calculate the probability of each alignment we use the following expression:

$$\Pr(\hat{\mathbf{f}}\,|\,\hat{\mathbf{e}}) = \prod_{i=1}^{|\hat{\mathbf{e}}|} t(\hat{\mathbf{f}}_i\,|\,\hat{\mathbf{e}}_i) \quad (5)$$

where the parameter $t(\hat{f}\,|\,\hat{e})$ estimates the probability that the word group, $\hat{f}$, be translated to the word group $\hat{e}$. There are the only parameters of our model.

## Training

We obtain the parameters of our translation model by using a training set of parallel sentences (Brown et al, 1993). To simplify the notation, we consider only one parallel sentence. In the training procedure, we need to maximize:

$$\Pr(\mathbf{f}\,|\,\mathbf{e}) = \sum_{\hat{\mathbf{e}}:\hat{\mathbf{e}}=\mathbf{e}} \sum_{\substack{\hat{\mathbf{f}}:\hat{\mathbf{f}}=\mathbf{f};\\ |\hat{\mathbf{e}}|=|\hat{\mathbf{f}}|}} \prod_{i=1}^{|\hat{\mathbf{e}}|} t(\hat{\mathbf{f}}_i\,|\,\hat{\mathbf{e}}_i) \quad (6)$$

subject to the constraints that hold for each $\hat{e}$:

$$\sum_{\hat{f}} t(\hat{f}\,|\,\hat{e}) = 1 \quad (7)$$

To maximize this function, Lagrange multipliers $\lambda_{\hat{e}}$ are introduced in the auxiliary function *h*:

$$h(t,\lambda)=\sum_{\hat{\mathbf{e}}:\hat{\mathbf{e}}=\mathbf{e}}\ \sum_{\substack{\hat{\mathbf{f}}:\hat{\mathbf{f}}=\mathbf{f};\\ |\hat{\mathbf{e}}|=|\hat{\mathbf{f}}|}}\ \prod_{i=1}^{|\hat{\mathbf{e}}|} t(\hat{\mathbf{f}}_i \mid \hat{\mathbf{e}}_i)-\sum_{\hat{e}}\lambda_{\hat{e}}\Big(\sum_{\hat{f}} t(\hat{f}\mid\hat{e})-1\Big) \quad (8)$$

We now calculate the partial derivative of $h$ with respect to $t(\hat{f}\mid\hat{e})$:

$$\frac{\partial h}{\partial t(\hat{f}\mid\hat{e})}=\sum_{\hat{\mathbf{e}}:\hat{\mathbf{e}}=\mathbf{e}}\ \sum_{\substack{\mathbf{f}:\mathbf{f}=\mathbf{f};\\ |\hat{\mathbf{e}}|=|\hat{\mathbf{f}}|}}\ \sum_{i=1}^{|\hat{\mathbf{e}}|}\delta(\hat{f},\hat{\mathbf{f}}_i)\delta(\hat{e},\hat{\mathbf{e}}_i)\,t(\hat{f}\mid\hat{e})\text{-}1)$$

$$\prod_{k=1}^{|\hat{\mathbf{e}}|} t(\hat{\mathbf{f}}_k \mid \hat{\mathbf{e}}_k)\ -\ \lambda_{\hat{e}} \quad (9)$$

where $\delta$ is the Kronecker delta function, which is equal to one when both of its arguments are the same and which is equal to zero otherwise. If we equate this partial derivative to zero the following equation is obtained:

$$t(\hat{f}\mid\hat{e})=\lambda_{\hat{e}}^{-1}\sum_{\hat{\mathbf{e}}:\hat{\mathbf{e}}=\mathbf{e}}\ \sum_{\substack{\hat{\mathbf{f}}:\hat{\mathbf{f}}=\mathbf{f};\\ |\hat{\mathbf{e}}|=|\hat{\mathbf{f}}|}}\ \prod_{i=1}^{|\hat{\mathbf{e}}|} t(\hat{\mathbf{f}}_i \mid \hat{\mathbf{e}}_i)\sum_{i=1}^{|\hat{\mathbf{e}}|}\delta(\hat{f},\hat{\mathbf{f}}_i)\delta(\hat{e},\hat{\mathbf{e}}_i) \quad (10)$$

The parameters that we are interested in estimating $t(\hat{f}\mid\hat{e})$) appear on both sides of equation 10. Thus, we need to use the EM algorithm in an iterative procedure (Brown et al, 1993). We chose as initials values for $t(\hat{f}\mid\hat{e})$, $1/|\hat{F}|$. We can calculate equation 10, using an efficient algorithm based on dynamic programming.

### Identifying Word Groups.

In recent years, some works have been presented to identify word groups in bilingual corpora (Maynard & Ananiadou, 1999; Ahrenberg et al., 1997; Och & Weber, 1998). Our approach to this problem is simple but efficient. We have taken advantage of the fact that almost all of sentence pairs of our corpus had been sequentially translated.

We take a sentence pair from the training set, and we attempt to find the best sequential alignment (**e** and **f**), that minimizes equation 4. At this point, we do not have the parameters of our model. Thus, we need to redefine equation 5. Model 1 presented in (Brown et al, 1993) is used for this purpose:

$$\Pr(\hat{\mathbf{f}}\mid\hat{\mathbf{e}}) = \prod_{i=1}^{|\hat{\mathbf{e}}|} \Pr_{\text{IBM-1}}(\hat{\mathbf{f}}_i\mid\hat{\mathbf{e}}_i) \quad (12)$$

$$\Pr_{\text{IBM-1}}(\mathbf{f}\mid\mathbf{e}) = \prod_{j=1}^{|\mathbf{f}|}\ \sum_{i=1}^{|\mathbf{e}|} t(\mathbf{f}_j\mid\mathbf{e}_i) \quad (13)$$

Where, $t(\mathbf{f}_j|\mathbf{e}_i)$ are the conventional lexical parameters of model 1. Once, we have the best monotone alignment of the sentence pair, we add all the groups of more than one word in $\hat{\mathbf{e}}$ to $\hat{E}$. The groups of more than one word in $\hat{\mathbf{f}}$ are added to $\hat{F}$. If a word group is detected less than ten times, it is erased.

## Search

The aim of search is to find an approximation to sentence **e** that maximizes the product Pr(**e**)Pr(**f**|**e**). The search algorithm is a crucial part in statistical machine translation. Its performance directly affects the quality and efficiency of translation (Wang & Waibel, 1998). In this section, we present a search algorithm based on stack-decoding. This algorithm obtains the translation of maximum probability in a few seconds.

The basic stack-decoding algorithm (Wang & Waibel, 1997) consists of an iterative process: We have a set of partial translation hypotheses comprised of a source sentence prefix sentence. We associate a score for each hypothesis according to the language model and the translation model. In each iteration, we select the hypothesis which has the highest score for extension. If the score of is lower than a threshold, we extend the hypothesis by adding a new word group to the right. The process continues until there are not more sentences to extend. Then the complete hypothesis with the highest score is selected as output.

In our implementation, a partial hypothesis was defined as the triple (*mk*, $\mathbf{e}_1..\mathbf{e}_{lk}$, *g*). *mk* was the number of words in the source sentence that was being considered. $\mathbf{e}_1..\mathbf{e}_{lk}$ was the translation prefix, and *g* was the score of the hypothesis ($g=\Pr(\mathbf{e}_1..\mathbf{e}_{lk})\Pr(\mathbf{f}_1..\mathbf{f}_{mk}|\mathbf{e}_1..\mathbf{e}_{lk})$). For a better performance, a separate stack was used for each hypothesis source sentence length. We stored a hypothesis in a different stack according to the value of *mk*. Thus, we needed the stacks numbered from 0 to |**f**|. Following, we show the search algorithm.

```
Initialize the stack 0 with a null hypothesis
Repeat as long as there are hypotheses to expand
    For each stack from 0 to |f|-1 do
        Pop the hypothesis with the highest score
        If score > threshold of the stack
            Extend the hypothesis
    If a new complete hypothesis has been created,
        Recalculate the stack thresholds
The highest score hypothesis of stack |f| is the output
```

Figure 2: Stack-decoding search algorithm

### Threshold computing.

For each stack, we have a threshold that has been utilised as a pruning criterium. A hypothesis which has a score, which is lower than the threshold of its stack, is eliminated. At the beginning, all thresholds are set to infinite. When a new complete hypothesis has been generated, if its score is greater than the best one so far, then, the thresholds are recalculated. The new threshold of a stack *i* is obtained dividing the score of the new best translation by the value $S_{|\mathbf{f}|-i}$. The value $S_j$ estimates the maximum probability contribution of a suffix of *j* words in any source sentence. These parameters can be pre-calculated with a parallel training set.

The sequential nature of a translation model make the use of a dynamic programming search algorithm (Tillmann et al, 1997; Garcia-Varea et al, 1998). We are interested in exploring this possibility in future work.

## Evaluation

In order to evaluate our model, we carried out some experiments. We used the corpus "El Periódico" obtained from the electronic edition of a general newspaper published daily in Catalan and Spanish.

The training corpus was made up of 10 months of the newspaper. We detected some kinds of words with special properties. If we considered a word was a number, an abbreviation, an acronym or a proper name, we substituted this word with a corresponding label. If a word appeared less than 30 times, it was replaced by the *$unknown* label. Figure 3 presents some statistical information about the corpus after the pre-processing phase.

| | **Spanish** | **Catalan** |
|---|---|---|
| Number of sentences | 643,961 | |
| Number of running words | 7,180,186 | 7,435,016 |
| Vocabulary size | 44,006 | 38,105 |
| Number of *$unknown* | 0.097% | 0.088% |

Table 1: Statistical information of the selected sentences from the "El periodico" corpus.

To learn the language model, we obtained a set of 850,521 Spanish sentences. We selected the trigram model for the system.

Table 2 shows the translation probabilities obtained for the Spanish word 'del' in our model (MonWG) and in the second translation model (IBM-2) present at (Brown et al, 1993).

| MonWG | IBM-2 |
|---|---|
| $t$( del \| del ) = 0.79 | $t$( del \| del ) = 0.70 |
| $t$( de l' \| del ) = 0.18 | $t$( de \| del ) = 0.11 |
| | $t$( l' \| del ) = 0.12 |

Table 2**:** Translation probabilities for "del" word.

To evaluate our translation system, we obtained 221 random sentences in Spanish, with a mean sentence length of 14 words. These sentences were extracted from the same corpus, but they were not sentences that we trained on. A total of 177 correct translations were obtained. Table 3 shows more details about the results and compares our system with the model IBM-2 and a rule-based commercial system (SALT-2).

For the evaluation of the translation quality we used the automatically computable Word Error Rate (WER) and the manually computable Subjective Word Error Rate (SWER). The WER corresponds to the edit distance between the produced translation and a predefined reference translation (Och et al., 1999). The SWER corresponds to the minimum edit distance between the produced translation and any correct translation. The concept of correct translation is subjective, therefore a person has to calculate this measure.

In some cases, the WER measure does not reflect properly the quality of the translations results. Table 3 show no so good WER results for SALT-2. A closer look to SALT-2 translated sentences will show that most of the detected error words will come from translated sentences different from the reference translation but with correct grammatically and meaning. In order to overcome the limitations of the WER measure, we introduce the SWER measure.

| **System** | **WER** | **SWER** | **correct sentence translation** | **incorrect sentence translation** | **translation speed (words/min.)** |
|---|---|---|---|---|---|
| MonWG | 12.4% | 1.6% | 80.1% | 19.4% | 56 |
| IBM-2 | 22.3% | 3.0% | 72.5% | 27.5% | 0.8 |
| SALT-2 | 20.0% | 1.5% | 81.4% | 18.6% | 290 |

Table 3: Sentence Translation Results. MonWG: our system. IBM-2: the second translation model present at (Brown et al, 1993) (one-stack stack-decoding algorithm is used in search (Wang & Waibel, 1997)). SALT-2: rule-based commercial system.

```
        e =  d' altra banda   ,   es va quedar   dels     primers
                   |          |       |           |         |
        f =  por otro lado    ,   se quedó       de los   primeros

 Pr(f|e,a)=      0.85   ·   0.97  ·  0.88    ·    0.97  ·   0.93   = 0.66
```

Table 4: Example of computing probability alignment in MonWG.

## Conclusions

A system for machine translation between the Spanish and Catalan languages has been presented. All components were inferred automatically from training pairs. For the language model, we used a conventional trigram model. For the translation, we presented a new model based on the sequential translation of word groups. A maximum likelihood estimation criterium was used for training the models. A Stack-Decoding algorithm was used for searching. These techniques were tested on the "El Periodico" corpus. Finally, we presented the performance results of the system and compared with others translations models.

In the future, we are interested in the exploration of new approaches that lead to more correct translations. This future works should include more complex alignment translation models and others search algorithms. Furthermore, we are interested in testing our translation model with others Romanic languages.


### Acknowledgements

This work was partially funded by the Spanish CICYT under grants TIC-1FD1997-1433 and TIC2000-1500-C02.


### Bibliographical References


Ahrenberg, L., Andersson, M. & Merkel, M. (1997) A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts. Proc. of the 35th Annual Meetingof the Association for Computational Linguistics, Madrid, Spain.

Bahl, L.R. (1983). A Maximum Likelihood Approach to Continuous Speech Recognition. IEEE Trans. Pattern Analysis, Machine Intellig. VOL. PAM1-5, NO.2, (pp.179—190).

Brown, P.F., Della Pietra, S., Della Pietra, V. & Mercer, R.L. (1993) The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, vol. 19 no. 2, (pp. 263--311)

Della Pietra, M., Epstein, M., Roukos, S. & Ward, T. (1997) Fertility Models for Statistical Natural Language Understanding. Proc. of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain.

Epstein, M., Papineni, K., Roukos, S., Ward, T., & Della Pietra, S. (1996). *Statistical natural language understanding using hidden clumpings*. In Proceedings of the 1996 International Conference on Acoustics, Speech and Signal Processing, Atlanta, Georgia, U.S.A., pp. 176-179.

Garcia-Varea, I., Casacuberta, F., & Ney, H. (1998) *An Iterative, DP-based Search Algorithm for Statistical Machine Translation*. In Proc. Int. Conf. Spoken Language Processing, Sidney, Australia.

Maynard, D. & Ananiadou, S. (1999) Identifying Contextual Information for Multi-Word Term Extraction.

Al-Onaizan, Y. et al. (1999). Statistical Machine Translation. Final report JHU workshop.

Tomás, J. & Casacuberta, F. (1999) A Statistical Spanish-Catalan Translator: A Preliminary Version. Proc. of the VIII Symposium Nacional de Reconocimiento de Formas y Análisis de Imágenes, vol. 1, (pp.103--110), Bilbao, Spain.

Vogel, S., Ney, H. & Tillmann, C. (1996) HMM-Based Word Alignment in Statistical Translation. International conference on Computational Linguistics, (pp. 836--841), Copenhagen, Denmark.

Och, F.J., Tillmann, C. & Ney, H. (1999) Improve Alignment Models for Statistical Machine Translation. In Proc. Of the Joint SIGDAT Conf. On Empirical Methods in Natural Language Processing and Very Large Corpora (pp. 20--28), Univ. of Maryland, Colege Park, MD, USA.

Och, F.J. & Weber H. (1998) Improve Statistical Natural Language Translation with Categories and Rules. in Proc. 35th Ann. Conf. Assoc. Computational Linguistics (pp. 985--989) Montreal, Canada.

Tillermann, C., Vogel, S., Ney, H. & Zubiaga, A. (1997) A DP based Search Using Monotone Alignments in Statistical Translation. Proc. of the 35th Annual Conf. Assoc. Computational Linguistics, Madrid, Spain, (pp. 289--296).

Wang, Y.Y., Waibel, A. (1997) Decoding Algorithm in Statistical Machine Translation. Proc. of the 35th Annual Meetingof the Association for Computational Linguistics, Madrid, Spain.

Wang, Y.Y., Waibel, A. (1998) Fast Decoding for Statistical Machine Translation. In Proc. Int. Conf. Spoken Language Processing.

Wang, Y.Y., Waibel, A. (1998) Modeling with Structures in Statistical Machine Translation, *Proceedings of the COLING / ACL98*