# A Hybrid Approach to Word Order Transfer in the English-to-Vietnamese Machine Translation

Dien Dinh, Nguyen Luu Thuy Ngan, Do Xuan Quang, Van Chi Nam

IT Faculty, Vietnam National University of HCM City

HCM City, Vietnam

ddien@saigonnet.vn

## Abstract

Word Order transfer is a compulsory stage and has a great effect on the translation result of a transfer-based machine translation system. To solve this problem, we can use fixed rules (rule-based) or stochastic methods (corpus-based) which extract word order transfer rules between two languages. However, each approach has its own advantages and disadvantages. In this paper, we present a hybrid approach based on fixed rules and Transformation-Based Learning (or TBL) method. Our purpose is to transfer automatically the English word orders into the Vietnamese ones. The learning process will be trained on the annotated bilingual corpus (named EVC: English-Vietnamese Corpus) that has been automatically word-aligned, phrase-aligned and POS-tagged. This transfer result is being used for the transfer module in the English-Vietnamese transfer-based machine translation system.

# 1 Introduction

In a syntactic-transfer-based machine translation system, the transfer stage plays an important role and strongly affects on the translation result. Structural tree transfer is converting the syntactic trees of the source language (which is English in this experiment) into the ones of the target language (here is Vietnamese). For example, the English sentence "*I bought an old book*", after going through the morphological and syntactic analysis layer, is as follows:

Sentence	Ι	bought	an	old	book
Lemma	Ι	buy	а	old	book
POS	PRO	V	DET	ADJ	N
Syntax	[] <sub>NP</sub>	[	[		] <sub>NP</sub> ] <sub>VP</sub>

Table 1. Information of the English analysis

After collecting this linguistic information, we carry out the English-Vietnamese tree tranfer stage as illustrated in the figure below:

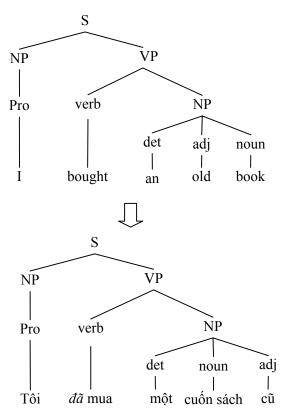


Figure 1. Transfer of English vs. Vietnamese trees

As in the example above, we see: in English noun phrases, the adjective precedes the noun, on which is opposite in almost Vietnamese cases. Due to the characteristics of language typology, in English (an inflectional language), the grammatical meaning is expressed under the inflection grammatical method as in:  $buy \rightarrow bought$ . In Vietnamese (an isolated language), we use function words instead, such as " $d\tilde{a}$ " to show a past event  $mua \rightarrow d\tilde{a} mua$  (Can, 1998).

Word order is the main grammatical method of Vietnamese and it is also the considerable difference between English and Vietnamese. Therefore, in our research, TBL is chosen to extract transfer rules that aim to change the order of grammatical compositions. Our paper is organized into 6 sections. Section 2 summarizes some prior arts. Our approach is presented in section 3. In this section we describe the TBL algorithm, training corpus and its format as well as the typical transfer rules that were extracted from our model. The evaluation and our experiments are given in section 4.

# 2 The Word Order Transfer approaches

In order to solve this problem, many approaches have been applied, for example: rule-based

approach, Statistical machine translation (SMT), corpus-based learning (TBL, MBL,...), and intermediate representations, etc. The following is the summary of those approaches.

# 2.1 Rule-based approach

This is a rather effective approach and was applied in the earliest days (since 1960s-1970s). In this approach, linguistic experts think out the transfer rules for each of the grammar rule used in analysis of the source language. For example: in order to change the orders of adjectives and nouns in English-Vietnamese translation (Dien, 1996), they can use the rule: "(E) NP  $\rightarrow$  Det Adj Noun  $\Rightarrow$  (V) NP  $\rightarrow$  Det Noun Adj". This method provides changes like reordering, insertion and deletion of parts in the right hand side of the same production rule (R1 to R2 in figure 2). As a result, it is impossible to change a node to another level or another governor. As R1 to R3 in figure 2:

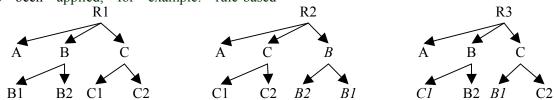


Figure 2: Ability of syntactic transfer by rules  $R1 \rightarrow R2$ : B swap for C and B1 swap for B2 (feasible);  $R1 \rightarrow R3$ : B1 swap for C1 (infeasible)

# 2.2 Statistical machine translation approach

This approach will be trained on bilingual corpora that have been word-aligned and draws transposition probabilities of the source words in comparison with the target words (Brown et al., 1993). According to this approach, translating an English sentence e to a desired Vietnamese sentence v is the same as finding a Vietnamese sentence v so that:

$$v = \arg \max P(v) * P(e | v)$$
, where:

- P(v) is the probability of the language model

-  $P(e \mid v)$  is the probability of the translation model

The reason why they calculate P(v) is to find the sentence, which satisfies Vietnamese,

grammar and, of course, has correct word-order. They use the following formula to compute this probability (due to the complexity of computation, we can reduce this n-gram model to tri-gram or bigram model):

$$P(v) = P(v_1 v_2 \dots v_l) = P(v_1) * P(v_2 | v_1) * P(v_3 | v_1 v_2) * \dots * P(v_l | v_1 v_2 \dots v_{l-1})$$

# 2.3 Corpus-based approach

This method will learn from the bilingual corpora that were word-aligned, phrase-aligned, POStagged and parsed, then produce the transposition rules between the two languages. There are many methods following this approach, e.g. STAG (Synchronous Tree-Adjoining Grammar) (Sheiber, 1990); dominance-preserving algorithm and LCA theory (Least Common Ancestor) (Meyers, 1998); etc.

#### 2.4 Intermediate representation approach

Instead of directly transferring as the abovementioned approaches, they also use other methods, which use a certain intermediate representation, for example, "case-frames" or QLF (Quasi-Logical Form) (Turhan, 1998); etc. This approach is effective in typical sentences (grammatical standard) but not for unrestricted sentences.

# 3 Our approach

While studying and researching the approaches in transfer-based machine translation, we explored different approaches separately. Then we recorded the advantages and disadvantages of each approach. As we discuss briefly in the above section, each pure approach has some gaps that affect strongly to the quality of word order permutation process. That gives us the motivation to establish a new approach, which is more effective and robust in order to break the neck of the disadvantages and take advantages of each approach. We propose a hybrid approach based on the combination of using fixed rules and machine learning method TBL to extract rules from bitexts. TBL is applied on the annotated English-Vietnamese bilingual corpus (named EVC) to extract rules which correct errors of the Rule-based transfer results. With this method, we have to solve such problems as: training corpus, baseline tagging, rule templates, etc. First, we briefly review TBL algorithm.

#### 3.1 Transformation-Based Learning

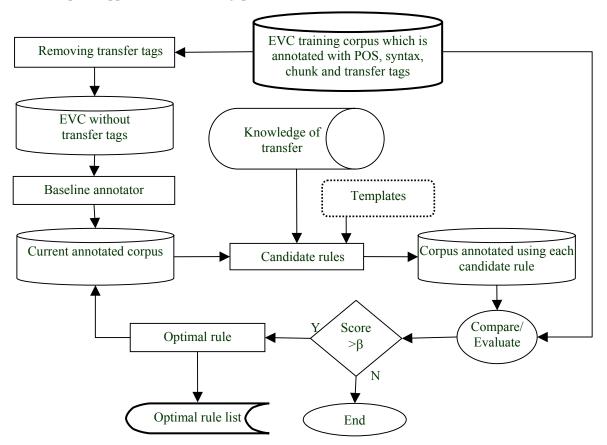


Figure 3: Flowchart of TBL- algorithm in learning transfer rules

The Transformation-Based Learning (or TBL) was proposed by Eric Brill in 1993 in his doctoral thesis (Brill, 1993) on the foundation of structural linguistics of Z.S.Harris. TBL has been successfully applied in various natural languages processing (mainly the annotating ones). A striking particularity of TBL in comparison with other learning methods is perceptive and symbolic: the linguists are able to observe, intervene in all the learning, implementing processes as well as the intermediary and final results. Besides, TBL allows the inheritance of the tagging results of another system (considered as the baseline or initial tagging) with the correction on that result based on the transformation rules learned through the training period.

TBL is active in conformity with the transformation rules in order to change wrong tags into right ones. All these rules obey the templates specified by human. In these templates, we need to regulate the factors affecting the tagging. In order to evaluate the optimal transformation rules, TBL needs the training corpus (the corpus to which the correct tags have been assigned, usually referred to as the golden corpus) to compare the result of current tagging to the correct tag in the training corpus. In the executing period, these optimal rules will be used for the language tagging to the new corpus (in conformity with the sorting order) and this new corpus must also be assigned with the

baseline tags similar to that of the training period. These language tags can be morphological ones (sentence boundary, word boundary), POS tags, syntactical tags (phrase chunker), sense tags, grammatical relation tags, and in this paper, it is transfer tags.

### **3.2** The Training Corpus

This is the 5-million-word English-Vietnamese bilingual corpus which was automatically wordaligned, phrase-aligned (Dien, 2002a) and POStagged (Yarowsky, 2001; Dien, 2003). This corpus was collected from many different bilingual text resources (books, dictionaries, corpora) belonging to science, technology, common conversation, ... For example:

\*D02:01323: Jet *planes* fly about nine miles high. +D02:01323: Các máy bay phản lực bay cao khoảng chín dặm.

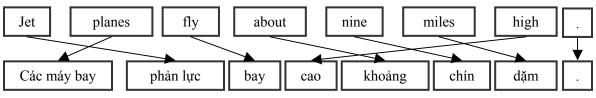


Figure 4. An example of word-alignment in the English-Vietnamese bilingual corpus

English	Jet	planes	fly	about	nine	miles	high
Vietnamese	phån lực	Máy	bay	khoång	chín	dặm	cao
		bay					
POS	NN	NN	V	IN	CD	NN	RB
Chunk	[	]NP	[	[	[	]NP	]PP]VP

Table 2. Part-of-speech, syntax, and chunk result in bilingual corpus

# 3.3 Baseline tagging

We use fixed transfer rules, going with each production rule which was used in parsing stage. Basing on production rules used in English sentences, we look up the rule which can map into production rules for Vietnamese sentence.

Example: The English sentence: "I have already read that interesting book." will be transferred (by using fixed rules in table 3) as follows: [[I/Tôi]NP [have/ε <already/rồi [read/đọc<<that/đó interesting/thú vi>ADJP book/cuốn sách>NP ]VP >VP ]VP .]

We will have the Vietnamese sentence "*Tôi đã* đọc cuốn sách thú vị đó rồi.".

English rules	Vietnamese rules
$S \rightarrow [NP VP]$	$S \rightarrow [NP VP]$
$NP \rightarrow [ADJP NP]$	$NP \rightarrow \langle ADJP NP \rangle$
$NP \rightarrow [NP1 NP2]$	$NP \rightarrow \langle NP1 NP2 \rangle$
$VP \rightarrow [have VP]$	$VP\!\!\rightarrow\![\mathcal{E}VP]$
$VP \rightarrow [already VP]$	VP→ <vp rồi=""></vp>
	$S \rightarrow [NP VP]$ $NP \rightarrow [ADJP NP]$ $NP \rightarrow [NP1 NP2]$ $VP \rightarrow [have VP]$

Table 3: Example for fixed transfer rules

Notes: we made use symbol [a b] to indicate the normal word order (a then b) and symbol  $\langle a b \rangle$  to indicate the inverted word order (b then a). In those fixed rules, we also implemented  $\varepsilon$  (deletion) and particle (insertion).

#### 3.4 Word Order Factors

In transferring word orders from English into Vietnamese, we found following factors which maybe affect to its transposition:

- Typology characteristics: this factor has been used in fixed rules in baseline transfer.
- Length of phrase: count on the number of words in a phrase (node).
- Part-Of-Speech (POS)
- Morphology
- Syntactic role

All these above-mentioned factors (except the first one) will be included in templates of TBL.

### 3.5 Rule templates

Rule templates in this problem will use word order factors from knowledge resources relating to transposition (out of knowledge resources was used in baseline tagging stage above). Generally, we use the following template:

$$\underset{\substack{LEN \ LEN \ LEN \ LEN \ LEN \ LEN \ LEN \ POS \ POS \ POS \ SYN \ Matha{}^{SEM} N_a^{SEM} [N_{a1}^{SEM} - N_{a2}^{SEM} - N_{am}^{SEM}] \Longrightarrow \\ \bigcup_{\substack{i,j,k=0 \ i\neq j\neq k}}^{m} N_0 [N_{ai} - N_{aj} - \chi - N_{ak}]$$

Where:

- *Word* is the variable presenting the morphology (such as: of, account, ...) of a word;
- *LEN* is the variable presenting the length of a node (the number of words in a node);
- POS is the variable presenting the part-ofspeed {N, V, A, ...} of a word or syntactic label of a node {VP, NP, ...};
- *SYN* is the variable presenting the syntactic role {Sub, Obj, Modifier, Goal, ...} of a node;
- *SEM* is the variable presenting semantics label {HUM, ANI, PLA, ...} of a node;
- $\chi$ : is the inserted particle.

The subscript presents the relative position of that part in the parent node, where m is the farthest position searched from the beginning. N<sub>a</sub> is the lowest node that dominates every transferring in rules. We examine the following example (for the convenience, we omit the superscript of the knowledge using in transferring)

$$N_{a}[N_{a1}[N_{a11}-N_{a12}]-N_{a2}[N_{a21}-N_{a22}]] \Longrightarrow$$
$$N_{a}[N_{a22}-N_{a2}[N_{a21}]-N_{a1}[N_{a12}-N_{a11}]]$$

<u>Remark</u>: Not every factor *Word*, *LEN*, *POS*, *SYN*, and *SEM* are necessary simultaneously. Each child node  $N_{ai} \in N_a$  can contain child nodes  $N_{aix}$  with the information that was defined (recursively) similarly to  $N_{ai}$ .

The transfer rule above has the syntactic tree type in figure 5 below:

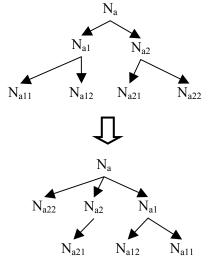


Figure 5. An example of Word Order transfer rule

#### 3.6 Corpus format

Format of the training corpus is built based on the language characteristics of English which strongly affect the word order in Vietnamese as we discuss in the above section. For the convenience in adding language characteristics, an English sentence is presented by two lines. The first line is marked with the asterisk (\*) and the second one is marked with the plus sign (+) as the following:

\* (S1 (S (NP (DT This)[4])[3](VP (AUX is)[6](NP (NP (DT a)[9](JJ good)[10](NN type)[11])[8] (PP (IN of)[13] (NP (NN book)[15])[14])[12])[7])[5](..)[16])[2])[1]

+ (This){Sub}<1> (is){V}<1> (a good type of book){Comp}<5> (.){End}<1>

The pair of brackets [.] presents the identification of word in English sentence. The pair of brackets {.} presents the syntactic role and the pair of brackets <.> presents the length of parts in the sentence. Note that the length characteristic was presented explicitly to simplify the learning process.

There are two kinds of corpus in the learning process, the golden and the training corpora. The golden corpus was built semi-automatically from the English-Vietnamese bilingual corpus, which was word-aligned (Dien,2002a), and then reviewed by linguists, whereas the training one was built based on the baseline tagging with fixed rules.

For the pair of English-Vietnamese sentence:

(E) "This is a good type of book ."

(V) "Đây là một dạng sách tốt ."

The corpus has the following format:

Raw corpus:

\* (S1 (S (NP (DT This)[4])[3](VP (AUX is)[6](NP (NP (DT a)[9](JJ good)[10](NN type)[11])[8] (PP (IN of)[13] (NP (NN book)[15])[14])[12])[7])[5](..)[16])[2])[1]

+ (This){...}<1> (is){...}<1> (a good type of book){...}<5> (.){...}<1>

Training corpus

\* (S1 (S (NP (DT This)[4])[3](VP (AUX is)[6](NP (NP (DT a)[9](NN type)[11] (JJ good)[10])[8] (PP (IN of)[13] (NP (NN book)[15])[14])[12])[7])[5](..)[16])[2])[1]

+ (This){Sub}<1> (is){V}<1> (a good type of book){Comp}<5> (.){End}<1>

Golden corpus

\* (S1 (S (NP (DT This)[4])[3] (VP (AUX is)[6] (NP (NP (DT a)[9] (NN type)[11] (NP (NN book)[15])[14] (JJ good)[10])[8] (PP (IN of)[0])[12])[7])[5] (..)[16])[2])[1] + (This){Sub}<1> (is){V}<1> (a good type of book){Comp}<5> (.){End}<1>

According to the above example, our program will extract the following transfer rule:

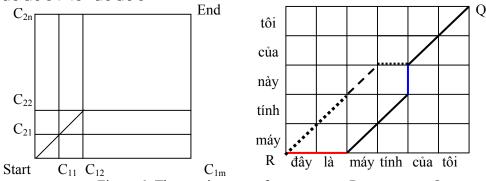
N1(NP){Comp}<5>[N11(NP)[N111(DT)

#### **4** Experiment – Results

We have implemented the model above to transfer English word order into Vietnamese ones in English-Vietnamese machine translation system. The experimental results as the following:

#### 4.1 Evaluation Results

Starting from the idea of Keh-Yih Su *et al.* (1992) about estimation measure of English-Chinese translation system, we change the problem of comparison between Vietnamese sentence transferred by machine with the one translated by human being into the problem of finding the shortest route between two points: from the departure point R and the target point Q. Where  $R = \{c_{11}, c_{12}, ..., c_{1m}\}$  is the Vietnamese sentence translated by machine having *m* words, and  $Q = \{c_{11}, c_{12}, ..., c_{1n}\}$  is the sentence translated by human being having *n* words.





The distance is calculated by

 $D = w_d * n_d + w_i * n_i + w_r * n_r + w_s * n_s$ 

Where:  $n_d$ ,  $n_i$ ,  $n_r$ , and  $n_s$  are the number of deleting (Del), inserting (Ins), replacing (Rep), and swapping (Swap) (NOP: no changing), and  $w_d$ ,  $w_i$ ,  $w_r$  and  $w_s$  are corresponding weights (this weights

are depended on language and experience). In our work,  $w_d=1$ ,  $w_i=5$ ,  $w_r=5$  and  $w_s=6$ .

From  $R \rightarrow Q$ , we have many routes, and different costs, for example:

R	đây	là	máy	tír	ıh	của	tôi	we have: $n_d=1$ , $n_r=3$
Q	máy	tính	này			của	tôi	D=1×1+5×3=16
Op1	REP	REP	REP	DF	EL	NOP	NOP	(the dash line)
R	đây	là	máy	tính		của	tôi	we have: $n_d=2$ , $n_i=1$
Q			máy	tính	này	của	tôi	D=1×2+5×1=7
Op2	DEL	DEL	NOP	NOP	INS	NOP	NOP	(the solid line)

Two routes above (may have many other routes) correspond to two different graphs in the right, where, the horizontal line corresponding to DEL, vertical: INS, diagonal: REP or SWAP. The cost which is used to edit a transferred sentence into the "golden" one is compute based on the following formula:  $k=m/D_{min}$ , where  $D_{min}$  is the shortest route (lowest distance) from R to Q, and m is the length of sentence. The defining of D<sub>min</sub> conducts to the "salesman problem", which is very familiar in dynamic programming technique. In our English-Vietnamese word order transfer problem, k is also considered as the error rate in word order transfer. In the case above, the shortest distance between R and Q is 7 (the solid line), so the transferred sentence "Đây là máy tính của tôi." consumes the editing cost k=2.54 (i.e. we still have to consume 2.54 units/word to transform R into Q).

#### 4.2 Transfer result estimation measurement

In our experiment, we use the CADASA corpus which is built from the English-Vietnamese bilingual books of "Come to the World of Microcomputers" (published by CADASA). This corpus is organized into 24 text files containing 8553 pairs of English-Vietnamese sentences. 8053 pairs were used for the training phase and the rest (500 pairs) were saved for testing. The training corpus needs to be annotated with part-of-speech tags, syntactic tags, etc., which is the important information for the transfer process. The training process took totally 36 hours and 29 minutes and 1260 transfer rules were extracted. In reality, we only use the first 565 rules in the rule list for testing thanks to observing the effects of number of rules used.

Total number of sentence	500
Total number of Vietnamese word	10660
Total number of insertions	891
Total number of deletions	961
Total number of replacements	1320
Total number of swaps	429
Total cost for the corpus	14590

Cost per sentence	29.18			
Cost per Vietnamese word	1.37			
Table 5: Statistic summary of distance measure				

	Baseline using	Transfer using TBL
	fixed rules	C
Total cost	18868	14590
Cost per sentence	37.74	29.18
Cost per	1.77	1.37
Vietnamese word		

Table 6: Performance of our transfer model

We apply the transfer rules extracted from learning process on the rest 500 pairs of CADASA corpus. The cost for changing a transferred sentence into the corresponding human translation will be computed as presented in 4.1. Our experiment result is summarized in Table 6. We found that the rules extracted from TBL training process can improve the quality of transfer process so that the transfer result can be closer to the human translation. If we assume that the word order of English and Vietnamese are similar, we can make a simple baseline tagging by using continuous orders. However, this solution is proved to be not only much more time-consuming but also lower precise when compared with our approach. In this experiment, it took us 47 hours and 48 minutes for learning and 1398 rules were extracted.

Unlike our approach, to our knowledge, all other English-Vietnamese MT systems made use of fixed rules for word-order transferring without learning from corpora (equivalent to our baseline tagging). Marginal phenomenon, which is a difficult problem with fixed-rule approach, is also solved with our approach. For example,

(E) "I have NP(JJ many) (NNS books))"

**(V1)** "Tôi có NP(NNS <u>sách</u>) (JJ <u>nhiều</u>))" (applied fixed transfer-rule NP(JJ NNS)  $\rightarrow$  NP(NNS JJ))

(V2)"Tôi có NP((JJ <u>*nhiều*</u>) (NN <u>sách</u>))" (applied two transfer-rules in the rule list extracted from

bilingual corpus "N1(NP)(N11(JJ) N12(NNS))  $\rightarrow$ N1(NP)(N12(NNS) N11(JJ)) and "N1(NP)(N11(NNS) N12(JJ many))  $\rightarrow$ N1(NP)(N12(JJ) N11(NNS))".

In Vietnamese (V2) has the right structure.

# 5 Conclusion

As above, we have presented the machine learning approach TBL for learning transfer rules from English-Vietnamese bilingual corpus in order to transfer word orders from English into Vietnamese. We have exploited the characteristics of TBL algorithm when putting the stage of transferring by fixed rule into baseline tagging stage. Besides, with the learning algorithm TBL, we have exploited the knowledge in linguistics, the homogeneous and heterogeneous features in syntactic structure, the word order between two linguistic typologies: inflection (for English) and isolation (for Vietnamese) in general, and between two languages in particular. These homogeneous and heterogeneous features were put into out model under two forms:

- Transfer rules in baseline tagging stage: basing on the basically heterogeneous features between two languages, which were discovered from the research results of comparative linguistics. These rules are used to solve core-phenomena.

- Transfer rules drawn from learning from bilingual corpus: in order to discover marginphenomena in transferring syntactic trees between two languages.

In the future, we will improve rule templates in order to draw more and more effective rules.

# **6** References

- Eric Brill. 1993. A Corpus-based approach to Language Learning. PhD-thesis, Pennsylvania Uni., USA.
- Eric Brill.1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. Computational Linguistics, 21(4), pp. 543-565.
- P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. *The mathematics of statistical machine translation: Parameter estimation.* Computational Linguistics, 19(2):263-311.

- Dinh Dien. 1996. *Rule-based English-to-Vietnamese Machine Translation*. Master thesis in Computer Science, University of Natural Sciences – Vietnam National University of HCMC.
- Dinh Dien, Hoang Kiem, Thuy Ngan, Xuan Quang, Quoc Hung, Phu Hoi, Van Toan. 2002a. *Word alignment in English–Vietnamese bilingual corpus*. Proceedings of EALPIIT'02, HaNoi, Vietnam, pp.3-11.
- Dinh Dien, Hoang Kiem. 2002b. Building a training corpus for word sense disambiguation in the English-to-Vietnamese Machine Translation. Proceedings of Workshop on Machine Translation in Asia, COLING-02, Taiwan, 9/2002, pg.26-32.
- Dinh Dien, Hoang Kiem. 2003. *POS-Tagging for English – Vietnamese Bilingual Corpus.* Accepted in Workshop on Parallel Texts at HLT-NAACL-03, Edmonton, Canada.
- Adam Meyers, Roman Yangarber, Ralph Grishman, Catherine Macleod and Antonio MorenoSandoval. 1998. *Deriving Transfer Rules from Dominance-Preserving Alignments*. Proceedings of the 36th ACL, Monreal, Canada, pg. 843-847.
- Geoffrey Sampson. 1995. English for the Computer: The SUSANNE Corpus and Analytic Scheme. Clarendon Press (Oxford University Press).
- Stuart Sheiber and Yves Schabes. 1990. Synchronous Tree Adjoining Grammars. Proceedings of the 13th COLING-90, Helsinki, ACL.
- Keh-Yih Su, Ming-Wen Wu, Jing-Shin Chang. 1992. A New Quantitative Quality Measure for Machine Translation System. Proceedings of COLING-92, Nantes, France, pp. 433-439.
- Cigdem Keyder Turhan. 1998. Structural Transfer in an English-to-Turkish Machine Translation System. PhD-thesis, Department of Computer Engineering, The Middle-East Technical University, Turkey.
- David Yarowsky and Grace Ngai. 2001. Induce, Multilingual POS Tagger and NP bracketer via projection on aligned corpora. Proceedings of NAACL-01.
- Can Nguyen Tai, *Ngữ pháp tiếng Việt*, The National University of Hanoi Publisher, 1998.