Machine Learning of Language Translation Rules

J. Tenni & A. Lehtola &
C. Bounsaythip & K. Jaaranen
VTT Information Technology
P.O.B. 1201, Espoo, FIN-02044 VTT, Finland
Jarno.Tenni@vtt.fi

ABSTRACT

The purpose of this paper is to present learning methods for creating language translation rules from multilingual text samples. The languages concerned are controlled languages, i.e. they are domain specific sublanguages with ambiguities eliminated by restricting the vocabulary and syntax. Learning methods presented here enable a supervised, human-assisted learning of generalised translation rules, thus making it faster and easier to adapt our machine translation system to new languages.

INTRODUCTION

This paper describes the machine learning approach developed at VTT for the discovery of translation rules for translating from a controlled language (CL) to another. Our Webtran machine translation software implements supervised machine learning to help adaptation to new language pairs [12]. The rules we are dealing with are rewrite rules consisting of feature constructs. The translation software is piloted for translating mail-order product descriptions from Swedish to Finnish.

CLs are domain specific human sublanguages that have limited vocabulary and restricted syntax. Control is used to minimise ambiguities in the texts. This enables, e.g., accurate fully automatic MT and text mining, which both are extremely difficult with unrestricted natural languages.

The use of CLs is growing in multilingual information systems. The approach has been successfully used to enhance the quality of translation [10], and also readability, comprehensibility, and maintainability of text originals. Examples include the TITUS system, which was designed for storing and translating abstracts on textiles from French to English, German and Spanish [9]. Also, the restricted vocabulary and telegraphic style syntax used in the weather forecast bulletins explain the success of TAUM-METEO [17]. Simplified English is used by AECMA [6] and others [1], [5], [13]. Currently Scania is implementing ScaniaSwedish for the preparation of truck maintenance manuals in controlled Swedish [2], [15].

Many WWW sites contain language that is almost controlled. This is the case also with product descriptions in several mail-order catalogues. The use of CL permits maintenance of the original catalogue in one language only and accurate translations can be obtained even in real-time. The CL approach requires that the catalogue maintenance process must include language check. Efficiency of the grammar acquisition and maintenance process is crucial for the usefulness of CL approach in practice.

In this paper we first outline our translation rule formalism. After that we present the learning approach we have developed for retrieving translation rules. Then we describe each learning method and present the benchmark results obtained when modelling the language used in a product catalogue.

AUGMENTED LEXICAL ENTRIES AND LEXICON

In Webtran, the controlled language definition is defined using a lexicon and rules, which we call Augmented Lexical Entry-rules (ALE-rules)[11]. The lexicon lists all the allowed words and the ALE-rules define allowed phrase and sentence structures and their translation relations.

In the lexicon words can have three kinds of properties: morphological, semantic and translation information. Morphological properties describe syntactic features (word class, inflection...). A word is also assigned with a semantic class. The classification depends on the use domain of the sublanguage. E.g. for the domain of women's clothes, as analysed from a mail-order catalogue, we are currently using 26 different semantic classes, like "product", "property", "colour" etc. Translation information tells the basic form of the corresponding word in another language. Only needed properties must be entered, so it is not necessary to define fully all the words appearing in texts. E.g., a word that does not appear in any generalised entry does not need semantic classification.

ALE-rules define the allowed sentences or phrases with their translation. Rules can be either surface form entries

(describing phrase that is always repeated in the same manner) or generalised rules. In the latter some or all words are not surface words but rather variables with restricting features. These variables are matched by any word that matches the restrictions. Restrictions can be semantic and/or morphological.

```
[women.cloth.shape.7
        [se ^(B){product} i form av ett (A){shape}]
        [fi (A){gen} muotoinen ^(B){nom}]
]
```

Figure 1 An example ALE-rule

Figure 1 defines a phrase, where the first word in Swedish (the se-part) belongs to "product"-class and is followed by surface part "i form av ett", which is followed by a word that belongs to "shape"-category. The rule is made for sentences like "a skirt with a shape of a bell". The translation for this is shown in the "fi"-part. The up-arrows in the phrases mark the governing words for dependency parsing purposes. The Finnish translation shows also the inflections of the words. The variables (A) and (B) are bound when the rule matches an excerpt of the input source text. The translations of the bound words are retrieved from the lexicon when output is generated in the target language.

MACHINE LEARNING METHODS

The learning methods described in this article are used for two purposes: creating ALE-rules and defining new words for the lexicon. Figure 2 shows the overall architecture of our supervised translation grammar learning system.

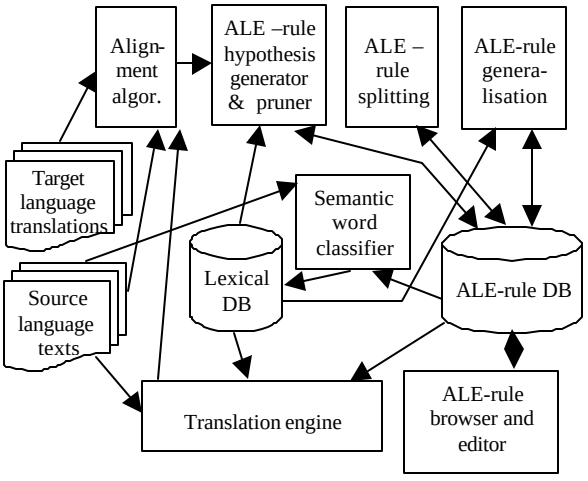


Figure 2 Architecture of the translation grammar learning system

Rules are created using three methods: alignment, split and generalisation. First, the alignment extracts translation equivalents from the example material. The equivalents are then transformed into ALE-rule-formalism using the ALE-rule hypothesis generator. The rules, whose matching score exceeds threshold value, are moved into ALE-rule DB and others are pruned. The split-method cuts the long rules, which are usually a result of alignment-method, into smaller phrases, so that their usage is extended. The generalisation method modifies the rules, so that some words in the rules are replaced by variables with feature restrictions and thus the rules can be used in a wider context. The existing controlled language definition can be applied, e.g., for inferring semantic information for new words found in text samples.

The translation engine can be used with an incomplete rule base to reveal those text excerpts that are not covered. The learning process can then focused on these excerpts.

All four methods are human assisted, i.e., their results are checked and expanded by a human. The user can verify and edit the rules using the ALE-rule browser and editor.

Sentence Alignment

Alignment is used for extracting translation relations from texts provided in two languages.

In ordinary translations, one sentence in the source language does not always get translated into one sentence (called substitution or 1-1 translation). Other types of translations include contraction (2 source sentences translate into one target sentence, 2-1), expansion (1-2), deletion (1-0), insertion (0-1) and merging (2-2). Merging could include more sentences (2-3,3-2,3-3,...), but they are very rare, and such cases do not exist in our material.

Alignment consists of three phases: sentence segmentation, sentence comparison and optimisation.

The sentence segmentation deals with the problem of finding sentence ends, i.e. separating the dots that mark sentence ends from the dots in another functions: abbreviations, in numbers etc. The sentence comparison calculates the matching probability of sentences in the source and target languages. The optimisation finds the optimal sentence correspondences from the matrix of matching probabilities.

Several methods have been applied to the problem, like Gale and Church- method length-based and alignment type probability alignment [7], cognates-based alignment [14] and statistical alignment [4].

For the CL approach, we have developed a new dictionary-based method for sentence alignment. With the dictionary, we onwards mean a bilingual lexicon providing translation equivalencies of words.

Sentence segmentation heuristics of Webtran, is based on dot-neighbourhood as the meaning of dot is concluded from the neighbouring words or characters of the dot. For detailed information, see [16].

The alignment method goes through sentence pairs in source and target languages and calculates the matching probabilities for different sentence combinations. The probabilities are saved in a matrix. As metrics we use character-wise length of the sentence, and matching based on the dictionary and the semantic classes of words.

The semantic class and the dictionary matching scores for whole sentence are obtained through summing individual word comparison values. Length score is directly the relative lengths of the sentences. The overall score is gained by weighting and summing individual comparison values so that the overall score for sentence is between 0 and 1. This weighting of different comparison values enables the method to adapt to the situation where it is being used. For instance, at the beginning phase, where the dictionary is small, the length value can be important, whereas at the later phase, the comparison can be mostly based on dictionary and semantic classes.

As the translation of a sentence typically tends to proportional to the original language, the lengths of the sentences are also compared. Measure is the relative ratio of sentence lengths, that is, the shorter length (in characters) divided by the longer length.

Words are matched on two criteria: semantic and dictionary values.

Semantic matching is based on the definitions of semantic classes for words in the lexicon.. As the semantic classes used in Webtran are not hierarchical, the semantic classes can be compared directly.

Dictionary matching is similar with the semantic matching, with the exception that the testing will be done against (possible) multiple translations. Multiple translations are possible when the controlled language definition is not finished and it is still ambiguous. When doing the dictionary matching the words are first reduced to their basic forms.

Evaluation of the sentence pairs found in alignment is based on the scores given in matching. Alignment results are optimised using the dynamic optimisation-algorithm.

Split operation

The results of the alignment are typically too long and too specific to be used as such for CL definition. For this reason, aligned sentences are split into phrases according to the current ALE-rule base.

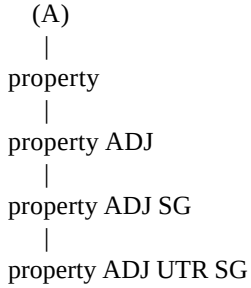
Those "phrase entries" usually present some semantic entity, so they often start with some special items that divide the sentence semantically. These items are retrieved from language specification and typically include comma and prepositions or conjunctions that start phrases in the ordinary natural language. By finding these words in the sentences, it is possible to divide the sentences into phrases that can be used for CL definition. When splitting a multilingual entry same amount of split points must be found in all language versions of the entry, even though it is not necessary that the splitpoints match according to the dictionary. After the splitpoints are found in both languages, each part between those points is moved into a separate ALE-rule. Dictionary based verification for each part could be used, even though the order of parts rarely changes, as can be noticed from our results-section.

Tree generalisation

The phrases created by splitting are usually too specific to enable compact and easily manageable language definition. For instance, each word with class “colour” would require its own entry in the CL definition. With generalisation, the words that share a similar function in the text can be handled with just a few entries. When new words are added to the vocabulary, they are categorised into semantic classes, and then processed without adding any new entries.

When creating a CL definition with sentence alignment, the results in the first phase are specific to the samples. The next phase is the generalisation process. The task of generalising translation entries is to extend the usability of entries so that they match also new cases where the structure of the sentence remains same but words appearing vary in some part(s) of the sentence. Furthermore, generalisations can be used when there is a need to diminish the number of entries of a controlled language, so that the entrybase is easier to manipulate.

We have applied an attribute-oriented phrase generalisation method in this task. Method has been previously used by Han et al. in discovering new and non-trivial information that is implicitly contained in the data of a relational database [8]. They concentrated mostly on finding quantitative rules from relational databases. We have applied and developed the method to suit sentence generalisation.



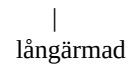


Figure 3 Tree structure generated for a Swedish word

When doing tree generalisation, the surface form entries are thought of as forests consisting of several trees. Each tree corresponds to one surface form word. Tree structures have words at their leaves and at the root they have a variable with no restrictions, and so the root matches every word. In between the leaves and the root there are nodes containing feature sets in the order of increasing generality when navigating towards the root. An example of surface form entry and a tree like structure generated from it is shown in Figure 3.

First, trees whose correspondent can be found on the other language, are chosen to be generalised. These trees are used in the generalisation by shifting from leaves towards the root until a proper level of generality is reached. This is evaluated by tracking the number of surface word forms that match the feature set. This matching word value is calculated both for each word and for whole sentence. Generalisation is continued until threshold values are reached. When generalisation produces duplicates, they are removed and in this way the entrybase gets smaller.

Semantic classification

The adaptive part of the system makes the lexicon fit to new texts and the current language definition. We use a semantic classification of words for this task. Prerequisite for this method is that generalised language definition exists. This method finds out the parts of the new texts that do not comply with the language definition, and tries to find out requirements for words in those parts, so that the parts would match the grammar definition.

The first task is to translate the text used as teaching material in order to find sentences that are not fully translated. Then these found sentences are split using the split operation described earlier. The next step is to translate all these split sentences one at the time and those parts that are not fully translated (some split parts may be translated) are saved for classification. These saved parts are called "non-translated parts".

The next task is to process the language definition in order to find from the language definition generalised rules, which have both variables and surface form words in their source part. The surface form words in the entries try to eliminate ambiguous classification. Then the actual comparison begins. The "non-translated parts" are compared with "general entries" one at the time. Whenever it is possible for a "non-translated part" to match an "general entry", the needed word-property-pair, gained from general entry, is saved as requirement. If all words in the part match the corresponding words in the entry, then all requirements are saved as suggestion. For each word-property-pair the suggestion with the number of times the requirement was set, are saved. When all part-entry- pairs are gone through, begins qualification. Only qualified requirements are shown to the user. In the qualification, for each word-property-pair, a qualification score is calculated by comparing the times that suggestion was made and the number of different semantic class suggestions for the word. If qualification score equals or exceeds user set threshold value then the suggestions are shown to the user.

TESTS AND RESULTS

Sentence Alignment Tests

Our sentence alignment-method was tested with Ellos catalogue. The idea was to test these alignment methods - sentence and split alignment - with the real material used in Ellos Ltd. Testing had three phases: material preparation, automatic alignment and evaluation.

The preparation phase was performed by a human, who checked the texts. The only changes to the texts were removal of some texts that were not part of the product descriptions. No descriptions were left out, that is, the tests included all 109 descriptions we had for one catalogue. The next phase was automatic alignment. The machine aligned texts in non-interactive mode, and the results were saved for evaluation. The final phase was the evaluation of the alignment results. In evaluation, a human checked the rules that had been created from the sample texts by the alignment program.

In tests, the catalogue was split into five parts, based on their usage on manual language modelling work. The first two test sets (test sets 1 and 2), were used throughout in manual language modelling, so their lexicon was very well covered in the dictionary. The next two sets (test sets 3 and 4) were less used in language modelling, but the most of their lexicon was covered in the dictionary. The final set (test set 5) was not used in manual language modelling at all.

In the evaluation, only unique sentence pairs were counted. This decreases the percentage of correct pairs, as for instance, in test case 5, all the duplicates (for both the whole and split sentences) were correct. This is needed, because the targeted usage for this alignment procedure is to model controlled language, where duplicates are not needed. The actual results of the alignment can be found in the Table 1.

The result table contains 8 values for each test set. "Sw" and "Fi" refer to the number of sentences given by sentence segmentation algorithm. "Total"-value is the total number of alignment pairs found by the alignment method and "non-faulty" is the number of correctly aligned

sentence pairs. This value includes both unnecessary complex and directly correct pairs. "Uniq. pairs" refer to the number of unique sentences that were found. These are separated because the description format has repeating information (like washing instruction and colour sentences), and duplicates are not needed in language modelling.
"Good pairs" means the sentence pairs that are direct sentence correspondents. "Comp." here means the unnecessary complex pairs. This means 2-2 sentence pairs that could have been aligned as two 1-1 pairs. Though they are not necessary faulty, especially as most of these pairs were retrieved from headers of descriptions, which typically contain one-word classification sentence and header-phrase. Such sentence-pairs could be used in language modelling to indicate special structure and requirements of the headers whereas in other parts of the description these one-word sentences could be prohibited. "Faulty pairs" are the sentence pairs that are aligned incorrectly.

Table 1 Sentence alignment test results

	Set 1	Set 2	Set 3	Set 4	Set 5	Total
Sw	276	272	267	298	338	1451
Fi	265	273	297	300	339	1474
Total	226	178	207	255	292	1156
Non-faulty	225	175	194	254	288	1136
(%)	99,6 %	98,3%	93,7%	99,6%	98,6%	98,1%
Uniq. pairs	138	142	137	159	176	
Good pairs	135	139	120	158	166	
(%)	97,8 %	97,9%	87,6%	99,3%	94,3%	
Comp	2	0	4	0	6	
(%)	1,4 %	0,0%	2,9%	0,0%	3,4%	
Faulty pairs	1	3	13	1	4	
(%)	0,7 %	2,1%	9,5%	0,6%	2,3%	

The results confirm that the alignment method of Webtran works well with the overall level of 98,1 %. This is a good result, as the material is not properly punctuated. This can be noted, when the faulty alignments are examined. The most common reason for failing is sentence segmentation, as can be seen from Table 2.

Table 2 Alignment errors

Reason	Count (%)
Incorrect sentence segmentation	12 (54,5)
Expressions	3 (13,6)
Other	7 (31,8)

Expressions are difficult cases for alignment, as their length and words can change totally, but they do not belong to the language specifications. Other errors were mostly cases where two similar sentences were one after each other and they were melted as 2-2 alignment, whereas they should have been two 1-1 alignments.

Split Alignment

Split alignment operated on the rules retrieved by sentence alignment. In the tests, the first task was to separate rules, which had same amount of split points in the source and target language. Otherwise, test sets are the ones created by sentence alignment. The results are shown in Table 3.

Table 3 Split alignment results

Name	Orig. numb.	Split parts	Uniq. parts	Corr. parts (%)	Incor. split (%)	Faulty orig. (%)
Test Set 1	23	46	33	33 (100)	0	0
Test Set 2	30	60	37	33 (89,2)	2 (5,4)	2 (5,4)
Test Set 3	35	100	54	50 (92,6)	4 (7,4)	0
Test Set 4	61	158	66	60 (90,9)	5 (7,6)	0
Test Set 5	60	151	69	63 (91,3)	6 (8,7)	0

The table shows the number of entries that were split "Orig. numb." is the number of entries that were processed, "Split parts" is the number of new entries that were generated, "Uniq. parts" is the number of unique parts. The results include "Correct parts", that is, the number of correctly split part, "Incorrect split" referring to method failure and "Fault. orig." is the number of incorrect splits, because split was tried on a rule incorrectly aligned.

The table shows that this kind of split operation can be used for creating smaller rules. These rules then enable easier adaptation of language definition to new materials.

Semantic Classification Tests

Semantic classification was tested with material that has not been previously used in language definition. All three test sets were loaded and non-translated parts were extracted. These parts were then compared with the language definition. The results are shown in Table 4.

Table 4 Semantic classification results

	Amount
Classifications found (incl. ambiguous)	261
Classifications suggested	90

Correct classification	60

Table includes "classifications found", which means all the words that could be part of the language definition just by adding the semantic class. Not all of these are relevant though, because some words match multiple existing sentence structures and thus there are ambiguous suggestions for some words. "Classifications suggested" is the subset of "classifications found", that are shown to the user for evaluation. Only those classifications were suggested where the score is over the threshold. Correct classification shows that 2/3 suggestions made by the system for the user are correct. The results indicate that the semantic classification method is useful, but the part used for eliminating ambiguous suggestions should be enhanced. The problems arise with very general entries that create ambiguous suggestions, which in turn make the system to dismiss suggestions.

Generalisation Tests

Tree generalisation method was tested with the entries from alignment and split methods. In the tests, automatically generated generalised entries translated new material that had not been used either in manual processing or in learning phase, and translation results were evaluated. New material described products of the same type as used in earlier processing, that is, women clothes. Table 5 shows the overall results.

Table 5 Translation results with automatically generated entries

Classification	Sentence count (%)
Correct	133 (76,9)
Inflection error(s)	13 (7,5)
Faulty/partially translated	27 (15,6)
Total	173

The "co-operation" of entries seemed to work very well with this new material as it reveals that three out of four sentences can be translated with automatically retrieved entries after the work with lexicon. This actually helps the language definer at the beginning of the language definition work. Later, when working with an existing language definition, these methods can be used to expand the language definition by applying methods to parts, which do not yet belong to the definition.

When the very first set of the material had been processed, there were over 700 generalised and surface form rules. After processing the second set of the material in about couple of days, the amount of rules decreased rapidly and after a few sets of new material the were a bit over 300 (mostly generalised) rules. Also the time used in the process became shorter after each new set of texts. As the rules become more and more generalised during the process and at the same time also more suitable for the new material, only the actual definition of the new words is needed. This does not take more than a couple of workdays of a professional translator (in a case of product descriptions).

CONCLUSIONS

In this article four methods for human-assisted controlled language definition have been presented. The methods have been tested with real-world text material of an online mail-order catalogue.

The three learning methods for creating rules, alignment, split and generalisation showed that they can be used together in order to create an initial language definition with relatively high translation quality.

The fourth method, monolingual semantic classification extends the coverage of the lexicon. Even though the recall of the method was low, the precision was satisfactory. In the future, the fourth method could be enhanced to work multilingually together with alignment. This could also help to raise the precision. After this combination, the method could be applied for retrieval of dictionary entries using the same basic idea.

The final test was to check how well the methods work together, i.e., how much automatic learning can happen. It was shown that the methods are very capable together. After processing women clothes descriptions of only one catalogue, the system was able to translate over three out of four sentences in a text excerpt of women clothes of a totally new catalogue. This is a very good result as the methods were intended to be interactive and human-assisted, but the results indicate that this manual work is needed only once after all the methods have been used, instead of three times. Thus after the initial manual work with lexicon, it is possible to let the machine work automatically throughout all three methods and use human-assistance only for making the final corrections.

The approach described in this paper can speed up the translation rule discovery process, especially in its beginning. It can lower the knowledge acquisition costs and reduce the time needed for adapting our machine translation system to a new controlled language. This way it promotes the use of machine translation technology in new domains and languages.

Acknowledgements

We would like to thank Eeva Palosuo and Maaret Virtanen from Tieto, Antti Mälkönen, René Holm and Irma Sinerkari from Ellos Oy for supporting our work. The work has been funded by Tekes Technology Development Centre in Finland.

REFERENCES

[1] G. Adriaens and L. Macken, "Technological Evaluation of a Controlled Language Application: Precision, Recall and Convergence Tests for SECC", *The Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 123-141, 1995.

[2] I. Almqvist and A. Sågvall-Hein, "Defining ScaniaSwedish - A Controlled Language for Truck Maintenance", *Proceedings of the 1st Int. Workshop on Controlled Language Applications, CLAW'96*, pp. 159-164, KU Leuven, Belgium, 1996.

[3] C. Carter, H. Hamilton, "Efficient Attribute-Oriented Generalization for Knowledge Discovery from Large Databases", *IEEE Transactions on knowledge and data engineering*, Vol. 10, No. 2, March/April 1998, pp.193-208.

[4] S. F. Chen, "Aligning Sentences in Bilingual Corpora Using Lexical Information", *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, 1993, pp. 9-16.

[5] S. Douglas, M. Hurs, "Controlled Language Support for Perkins Approved Clear English (PACE)", *Proceedings of the 1st Int. Workshop on Controlled Language Applications, CLAW'96*, pp. 93-105, 1996.

[6] P. van der Eijk, "Controlled Languages and Technical Documentation", *Report of Cap Gemini ATS*, Utrecht, 1998.

[7] W. Gale, K. Church, "A Program for Aligning Sentences in Bilingual Corpora", *Proceedings of 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA, 1991, pp.177-184.

[8] J. Han, Y. Cai, N. Cercone, "Data-Driven Discovery of Quantitative Rules in Relational Databases", *IEEE Transactions on knowledge and data engineering*, Vol. 5, No. 1, February 1993, pp.29-40.

[9] W. J. Hutchins and H. Somers, *An Introduction to Machine Translation*, Academic Press, 1992.

[10] R. Kittredge, "The Significance of Sublanguage for Automatic Translation", S. Nirenburg, editor, *Machine translation: Theoretical and methodological issues, Studies in Natural Language Processing*, pp. 59-67, Cambridge University Press, 1987.

[11] A. Lehtola, J. Tenni, and C. Bounsaythip, "Definition of a Controlled Language Based on Augmented Lexical Entries", *Proceedings of the Second International Workshop on Controlled Language Applications (CLAW'98)*, pp. 16-29, Pittsburg, Pennsylvania, USA, 21-22 May 1998, Language Technologies Institute, Carnegie Mellon University.

[12] T. Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997, 414 p.

[13] R. Schwitter and N. E. Fucchs, "Attempto Controlled English – A Seemingly Informal Bridgehead in Formal Territory", *Proc. poster session of JICSLP'96*, Bonn, Germany, September 1996.

[14] M. Simard, G. F. Foster, P. Isabelle, "Using Cognates to Align Sentences in Bilingual Corpora", *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal/Canada, 1992, pp.67-81.

[15] A. Sågvall-Hein, "Language Control and Machine Translation", *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-97)*, Santa Fe, 1997

[16] J. Tenni, *Methods and a Tool for controlled language definition*, Master Thesis, University of Helsinki, Finland, 1999.

[17] P. Whitelock and K. Kilby, *Linguistic and Computational Techniques in Machine Translation System Design*, UCL Press Limited, second edition, 1995.