An Evaluation of the Multi-Engine MT Architecture

Christopher Hogan and Robert E. Frederking

Language Technologies Institute
Carnegie Mellon University
4910 Forbes Avenue
Pittsburgh, PA 15213 USA
Phone: (412) 268-6593 or (412) 268-6656
FAX: (412) 268-6298
Email: chogan@cs.cmu.edu, ref+@cs.cmu.edu

Abstract. The Multi-Engine MT (MEMT) architecture combines the outputs of multiple MT engines using a statistical language model of the target language. It has been used successfully in a number of MT research systems, for both text and speech translation. Despite its perceived benefits, there has never been a rigorous, published, double-blind evaluation of the claim that the combined output of a MEMT system is in fact better than that of any one of the component MT engines. We report here the results of such an evaluation. The combined MEMT output is shown to indeed be better overall than the output of the component engines in a Croatian $\leftrightarrow$ English MT system. This result is consistent in both translation directions, and between different raters.

The Multi-Engine Machine Translation (MEMT) architecture [9] has been used successfully in a number of MT research systems, for both text [11, 24] and speech translation [12, 26]. As described in the next section, these researchers believe that the MEMT architecture allows one to combine the strengths of different MT technologies while ameliorating their weaknesses.

Up to now, this belief has been justified by argumentation, but not empirical evidence. While at least one of these MEMT-architecture systems was the subject of independent evaluation,[1] this was an overall system evaluation, and thus did not distinguish between the quality of the component engines and any benefit (or detriment) caused by the MEMT architecture. The lack of any rigorous, double-blind evaluation of the MEMT architecture itself was the motivation for our current effort.

We first describe the MEMT architecture and its presumed benefits in general. We then describe the specific translation sources used in this experiment, and discuss the design of our evaluation. We present a detailed statistical analysis of the results. Finally we conclude with the observation that the MEMT system has indeed been shown to produce better output than its component engines.

[1] Pangloss participated in the DARPA MT evaluations [11, 28]

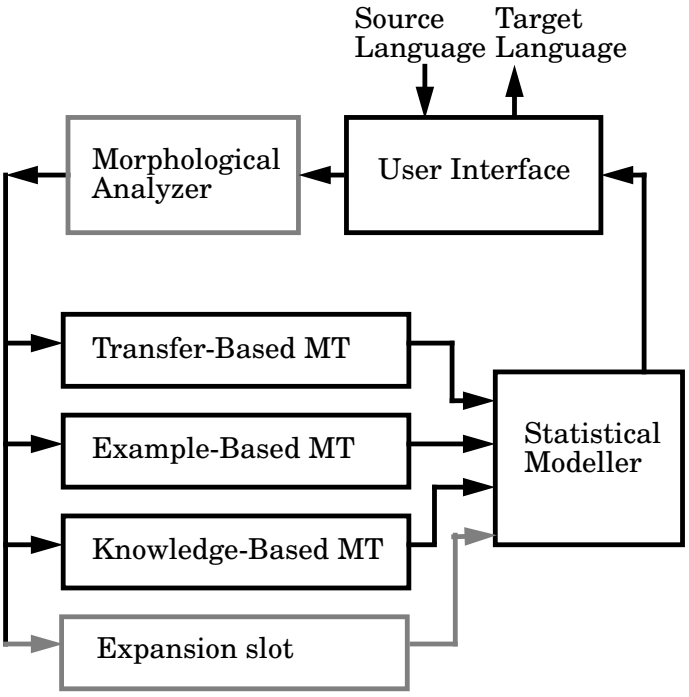

Fig. 1. Structure of MEMT architecture

1 The Multi-Engine MT Architecture

Different MT technologies exhibit different strengths and weaknesses. Technologies such as Knowledge-Based MT (KBMT) can provide high-quality, fully-automated translations in narrow, well-defined domains [21, 7]. Other technologies such as lexical-transfer MT [23, 8, 19], and Example-Based MT (EBMT) [4, 22, 27] provide lower-quality general-purpose translations, unless they are incorporated into human-assisted MT systems [10, 20], but can be used in non-domain-restricted translation applications. Moreover, these technologies differ not just in the quality of their translations, and level of domain-dependence, but also along other dimensions, such as types of errors they make, real-time translation time [26], required development time [12], cost of development, and ability to easily make use of any available on-line corpora, such as electronic dictionaries or online bilingual parallel texts.

The Multi-Engine Machine Translation (MEMT) architecture [9] makes it possible to exploit the differences between MT technologies. As shown in Fig. 1, MEMT feeds an input text to several MT engines in parallel, with each engine

employing a different MT technology[2]. Each engine attempts to translate the entire input text, segmenting each sentence in whatever manner is most appropriate for its technology, and putting the resulting translated output segments into a shared chart data structure [16, 29] after giving each segment a score indicating the engine's internal assessment of the quality of the output segment. These output (*target language*) segments are indexed in the chart based on the positions of the corresponding input (*source language*) segments. Thus the chart contains multiple, possibly overlapping, alternative translations. Since the scores produced by the engines are estimates of variable accuracy, statistical language modelling techniques adapted from speech recognition research are used to select the best overall set of outputs [3, 13]. These selection techniques attempt to produce the best overall result, taking the probability of transitions between segments into account as well as modifying the quality scores of individual segments.

Among the differences that one frequently wishes to exploit using the MEMT architecture is the differing trade-off between coverage and quality that exists between different technologies: one would like to cover as much of the input as possible, but still get the highest possible quality for any given segment of the input. While there has been a strong perception that MEMT does provide this benefit, up to now there has never been a rigorous, double-blind evaluation that empirically verified this claim.

2 Experiment Design: Translations

In this experiment, we seek to evaluate the MEMT architecture separately from its implementation with specific translation engines. However, because MEMT performs no translations of its own, only combining and choosing from among the translations provided by other translation engines, it is necessary to include several translation engines in order to perform the evaluation. For this reason, we will simultaneously evaluate the MEMT system as a whole as well as each component translation engine. By comparing the results, we hope to shed light on the contribution of the MEMT architecture to the overall translation process.

The translations to be evaluated will therefore come from three sources. For every source language sentence to be translated, we will first translate it using the entire MEMT system as it is currently designed. Then, we will translate the sentence again using each of the separate translation engines that are part of the MEMT system. The system used for this experiment (the translation component of the DIPLOMAT Croatian $\leftrightarrow$ English system [12]) employs two kinds of **translation engines**:

Lexical-Transfer Simple dictionary (word-for-word translation) and glossary (phrase-for-phrase) translation.

[2] Morphological analysis, part-of-speech tagging, and possibly other text enhancements can be shared by the engines.

Example-Based MT Translation via partial matching of the input against a parallel corpus (example-base).

These translation engines are described in more detail elsewhere [8, 4].

In the MEMT architecture, each translation engine is permitted to hypothesize multiple translations for any sequence of words in the source sentence. The multiple, overlapping translations are sorted out by the language model, which performs a search over the set of available translations to find the subset that exactly covers the input and yields the best combination. When testing the translation engines separately from the MEMT architecture, other means must be employed for sorting out the conflicting translations. For each of the three **translation sources**, the following describes the technique used to obtain a workable translation from among the conflicting possibilities output by the engine.

LEX Using only translations from the Lexical-Transfer engine, perform a random search in the chart of candidate translations. Avoid translations that do not result in a completely translated sentence. There will be more than one possible translation per sentence.

EBMT Using the Example-Based engine, randomly select as many non-overlapping translations as possible. Because Example-Based MT is not able to translate all parts of all sentences[3], there may be parts of the source sentence that are untranslated. Fill gaps in the translation with appropriate selections from the Lexical-Transfer engine. There may be more than one possible translation per sentence.

MEMT The optimal translation as selected by the MEMT architecture. There will be only one translation possible per sentence.

3 Experiment Design: Evaluation

The actual evaluation took the form of a questionnaire, evaluated by native speakers of the target language. Two questionnaires were designed, one for each translation direction: English $\rightarrow$ Croatian and Croatian $\rightarrow$ English. The questionnaires had the following format: A series of source-language sentences were presented to the evaluators, each accompanied by four target-language translations. Evaluators were asked to qualitatively evaluate each of the translations based on their knowledge of the target language and the meaning of the source language sentence.[4]

[3] The reason for this is that the EBMT engine operates by performing partial matches against its parallel corpus. If no match is found between the input and corpus, or between the source and target sides of the corpus, no translation is produced. On the other hand, the algorithm increases the likelihood that any sentences that do match will be accorded a fairly high quality translation.

[4] Because the native English-speaking evaluators were not bilingual in Croatian, a high-quality human translation of the Croatian sentence into English was provided for them in addition to the original sentence.

3.1 Selection of Translations

In order to reinforce the goal of double-blind evaluation as well as to deal with certain difficulties posed by the translation engines, we used the following algorithm to generate the four translations of each sentence:

```
1. Randomly place the MEMT translation in
   one of the four slots.

2. If at least one EBMT translation is
   available, randomly select one and
   randomly place it in one of the
   available three slots.

3. For each of the two or three slots still
   empty, randomly generate a LEX
   translation and place it there.
```

The identity of the source of each of the translations is recorded but hidden from both the evaluator and the researcher.

3.2 Method of Evaluation

Evaluators were asked to evaluate the quality of each of the four translations using a scale from 1 to 5. Due to the often inconsistent intra-sentence quality of MEMT translations, additional scale points are used to distinguish translations that are partly correct from those that are uniformly bad. The scale is given below exactly as it was presented to the evaluators.

5. perfect
4. one or two errors, but otherwise perfect
3. several errors but understandable
2. some parts correct, but cannot understand
1. totally incomprehensible

Because some MT researchers, *e.g.* [14], have attempted to establish a simplified scale of `GOOD`, `ACCEPTABLE`, `UNACCEPTABLE` (or `BAD`) in order to encourage comparisons between evaluations of different systems, we would like to suggest that our scale be mapped into the three point scale in the following way: `GOOD` = 5, `ACCEPTABLE` = { 4, 3 }, `UNACCEPTABLE` = { 2, 1 }.

3.3 Evaluators

The evaluators were native speakers of the target languages they were asked to evaluate, and included two speakers of Croatian and two speakers of English. We

will denote the evaluators with the labels *cro1, cro2, eng1* and *eng2*. Although one of the evaluators was knowledgeable about the workings of the translator, none had information about the sources of the translations they were asked to evaluate.

The questionnaires were presented to the evaluators as printed versions of HTML pages. The Croatian evaluators were given 500 sentences each, the English evaluators 161 sentences each.[5] Evaluators received only clarification of the meaning of the source language from the researcher.

3.4 Domain

The source language sentences for translation were drawn from the domain of travellers' phrasebooks, a domain closely related to that of the actual system. Phrases to be translated were drawn from available English and Croatian phrasebooks, neither of which had been used in the development of the system.

4 Results

In this section, we present the results of the evaluation. First, we present simple statistics comparing the evaluators and translation engines. We will argue that the simple statistics are not sufficient. We then present a somewhat different approach to measuring the data, and make clear why we believe that it is a superior measure, more accurately reflecting the quantities that we wish to assess.

4.1 Initial Statistics

First let's look at the simple descriptive statistics. We compute averages and standard deviations of the scores $[1 \ldots 5]$ assigned by each evaluator {*cro1, cro2, eng1, eng2*} to each translation source {EBMT, LEX, MEMT}.

		cro1	*cro2*	*eng1*	*eng2*	Total
EBMT	mean	2.08	1.98	2.59	2.99	2.20
	stddev	1.13	1.19	1.29	1.20	1.23
LEX	mean	1.66	1.53	2.30	2.68	1.82
	stddev	0.93	0.91	1.32	1.29	1.10
MEMT	mean	1.92	1.86	2.58	2.96	2.10
	stddev	1.12	1.20	1.42	1.29	1.27
Total	mean	1.80	1.70	2.42	2.80	1.96
	stddev	1.03	1.06	1.34	1.28	1.18

Looking at this data, there are two kinds of comparison which are simple to make: comparisons between translation sources and comparisons between evaluators. We would like to raise several issues regarding both of these comparisons.

[5] The smaller size of the English evaluation was due to the necessity of providing the English evaluators English translations of the source-language sentences as well as the original Croatian ones.

4.2 Inter-source Comparison

A superficial overview of the data suggests that insofar as quality is concerned, EBMT > MEMT > LEX. Taken at face value, this would seem to suggest that MEMT is contributing negatively to the translation process, that EBMT alone would be superior to the entire system. However, we can statistically test the hypothesis that the means are the same with Student's t test [25, 2]. Doing so reveals that while both EBMT and MEMT are significantly different from LEX[6], they are not distinguishable from one another. Thus, we are prevented from establishing the relative ranking of MEMT against its component engines.

Note that this difficulty is not due simply to the large standard deviations:

> Our first thought is to ask "how many standard deviations" one sample mean is from the other. That number may in fact be a useful thing to know. It does relate to the strength or "importance" of a difference of means *if that difference is genuine*. However, by itself, it says nothing about whether the difference *is* genuine, that is, statistically significant. A difference of means can be very small compared to the standard deviation, and yet very significant, if the number of data points is large. Conversely, a difference may be moderately large but not significant, if the data are sparse. [...] (emphasis original)
>
> [25, pp. 464–5]

Further investigation reveals that the higher quality of EBMT is a result of the way the averages were computed. As stated earlier, this EBMT system cannot provide a translation of every sentence. The averages listed above for EBMT were computed only over those sentences for which an EBMT translation was available. This necessarily results in skewed statistics, since the other translation sources (MEMT, LEX) are forced to provide a translation for every sentence, with no option to "give-up".[7]

One obvious way of dealing with the problem of EBMT is to artificially skew the results back toward what is expected. Assuming that empty translations are evaluated as "totally incomprehensible", we can insert a translation with a score of 1 every time a sentence fails to include an EBMT translation. Doing this produces the following scores for EBMT:

		cro1	*cro2*	*eng1*	*eng2*	Total
EBMT	mean	1.79	1.72	2.07	2.35	1.87
	stddev	1.08	1.11	1.30	1.36	1.17

The data now show that EBMT is performing about as expected as compared to the other translation engines: the average scores are 2.10 for MEMT, 1.87

[6] With significance better than $p < 0.001$, which is highly significant.

[7] Not that this would be easy to implement for the other translation sources. The problem of self-evaluation for machine translation is rather difficult, and even EBMT (which does better than most) can only reliably return "yes" or "no".

for EBMT and 1.82 for LEX. The statistical tests now show that MEMT is significantly better than EBMT and LEX, which are not distinguishable from one another. While this establishes that MEMT is useful, we are in the unfortunate situation of being unable to provide a simple, statistically sound ranking of the engines relative to the overall architecture. In addition, this method of "artificial" evaluation seems to us to be an inelegant hack for a problem which may have a better solution.

4.3 Inter-evaluator Comparison

The second interesting aspect of the data presented earlier is that of inter-evaluator agreement. Agreement between human subjects is a well-investigated area in Psychology [18], Sociology [17], Medicine [1, 6] as well as Translation [15]. Without delving into the area of agreement measures, it seems to us that it is desirable to have a performance metric for which different evaluations of the same material produce similar scores. A comparison of our evaluators using the mean as a performance metric suggests that, at least quantitatively, the mean does not have the desired property (compare *cro1* 1.80 to *cro2* 1.70 or *eng1* 2.80 to *eng2* 1.96). This highlights a common problem in evaluating translations (*cf., e.g.* [5]), namely that there is no way to ensure that evaluators will agree on what constitutes a certain score. On the other hand, the mean does contain some information about the relative ranking of the sources insofar as all evaluators agree that the ranking of the sources is EBMT $\approx$ MEMT $>$ LEX (or MEMT $>$ EBMT $\approx$ LEX, depending on which set of EBMT scores one believes). The matter at hand is how to reconcile these qualitative agreements into a statistic that also agrees quantitatively.

4.4 A More Informative Statistic

In this section we will present a different kind of statistic, one based on comparisons between translation sources, and comment on the degree to which it solves the problems raised in the previous sections.

Consider carefully the way in which MEMT translates a sentence. The sentence is first translated by all of the translation engines. The candidate translations are scored and placed into a chart structure. The language model then selects the set of translations from the chart that best form a sentence in the target language. Clearly, given this scheme, the MEMT translation can be no better than the best of its engines on a given sentence, for MEMT does no actual translation itself, only using those provided by its engines. We must therefore include in any statistic that measures the effectiveness of MEMT some measure of the translation engines' performance on the same sentence. The metric we suggest is the following: how often does MEMT achieve the best that it is capable of, namely: the best that is available from the translation engines?

We will therefore calculate for each translation source the following statistic: for what percentage of the sentences does this translation source receive a score that is equal to or greater than that of all the other translation sources.

More formally, let $s_1, \ldots, s_N$ be the sentences in the source language to be translated, and $\tau_1, \ldots, \tau_M$ be the available translation sources. Now let $T_i = (t_{i1}, \ldots, t_{iM})$ be the translations into the target language of s_i by each of the translation sources. Finally, let $\Sigma_i = (\sigma_{i1}, \ldots, \sigma_{iM})$ be the scores assigned by the evaluators to each of the translations of sentence s_i. For each translation source $\tau_m \in \{\tau_1, \ldots, \tau_M\}$, we define the following metric:

$$d(\tau_m) = \frac{\sum_{i=1}^{N} \delta(\sigma_{im}, \max\{\sigma_{i1}, \ldots, \sigma_{iM}\})}{N}$$

Where $\delta(i,j)$, the Kronecker delta function, is given by:

$$\delta(i,j) = \begin{cases} 1 & for \quad i = j \\ 0 & for \quad i \neq j \end{cases}$$

This defines a measure in the range [0,1]. In the following table, we present the values as percentages.

	cro1	*cro2*	*eng1*	*eng2*	Total[8]
EBMT	55.80%	55.60%	49.69%	46.58%	53.86%
LEX	72.80	69.00	69.57	66.46	70.20
MEMT	73.20	73.00	70.19	70.81	72.47

This statistic appears to be superior to the mean of the score in several ways. Fundamentally, this statistic provides a measure of the degree to which MEMT is doing the job it was designed for: picking the best possible translation, and provides a clear goal (100%) to aim for. This measure is also independent of the actual translation engines used. If more translation engines were used in an evaluation, or different ones, we would expect to be able to compare the results with the current evaluation. In this sense, the statistic is a measure of the MEMT architecture rather than of a particular MEMT system with specific translation engines.

Secondly, this statistic implicitly deals with the problems that arise when translation sources cannot always produce translations, such as is the case for EBMT in our evaluation. Such translation sources are penalized rather severely as the final measures for EBMT (46% – 56%) indicate.

Thirdly, this statistic shows significantly better inter-evaluator agreement than the mean. For those pairs of evaluators that worked on the same material, the maximum difference appears to be about 3%. For the MEMT scores, which are of greatest interest to us, there is remarkable agreement, with less than 1% difference.

These statistics clearly indicate that MEMT is doing its job: it is selecting the best translation available 72.47% of the time.

[8] We only report totals across all evaluators. Since our measure is a comparison between translation sources, totals across all translation sources do not make sense.

5 Conclusion

In this evaluation, a Croatian ↔ English MT system using the MEMT architecture has been shown to produce better output than its component engines. Careful statistical analysis of the results showed that the best translation available was selected 72.47% of the time. This result was consistent in both translation directions, and between different raters. This is the first empirical evidence substantiating the claims previously made for the MEMT architecture.

References

1. J. J. Bartko and W. T. Carpenter. 1976. On the Methods and Theory of Reliability. *The Journal of Nervous and Mental Disease*, 163(5):307–317.
2. I. N. Bronstein and K. A. Semendyayev. 1985. *Handbook of Mathematics*. Verlag Harri Deutsch, Thun and Frankfurt/Main, Frankfurt/Main, third edition.
3. Ralf Brown and Robert Frederking. 1995. Applying Statistical English Language Modeling to Symbolic Machine Translation. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95)*, pages 221–239.
4. Ralf Brown. 1996. Example-based Machine Translation in the Pangloss System. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*.
5. David Carter *et al.* 1997. Translation Methodology in the Spoken Language Translator: An Evaluation. In Steven Krauwer *et al.*, editors, *Spoken Language Translation: Proceedings of a Workshop*, pages 73–81, Madrid, Spain, July. Association of Computational Linguistics and European Network in Language and Speech.
6. G. Dunn. 1989. *Design and Analysis of Reliability Studies: The Statistical Evaluation of Measurement Errors*. Oxford University Press, New York.
7. David Farwell and Yorick Wilks. 1991. Ultra: A Multi-Lingual Machine Translator. In *Proceedings of Machine Translation Summit III*, Washington D. C., July.
8. Robert Frederking and Ralf Brown. 1996. The Pangloss-Lite Machine Translation System. In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 268–272, Montreal, Quebec, Canada, October.
9. Robert Frederking and Sergei Nirenburg. 1994. Three Heads are Better than One. In *Proceedings of the Fourth Conference on Applied Natural Language Processing (ANLP-94)*, Stuttgart, Germany.
10. Robert Frederking *et al.* 1993. An MAT Tool and its Effectiveness. In *Proceedings of the DARPA Human Language Technology Workshop*, Princeton, New Jersey.
11. Robert Frederking *et al.* 1994. Integrating Translations from Multiple Sources with the Pangloss Mark III Machine Translation System. In *Proceedings of the First Conference for Machine Translation in the Americas (AMTA)*, Columbia, Maryland, October.
12. Robert Frederking, Alexander Rudnicky, and Christopher Hogan. 1997. Interactive Speech Translation in the DIPLOMAT Project. In Steven Krauwer *et al.*, editors, *Spoken Language Translation: Proceedings of a Workshop*, pages 61–66, Madrid, Spain, July. Association of Computational Linguistics and European Network in Language and Speech.

13. Robert Frederking. 1994. Statistical Language Models for Symbolic MT. In *Language Engineering on the Information Highway Workshop*, Santorini, Greece, September.
14. Donna Gates *et al.* 1996. End-to-End Evaluation in Janus: A Speech-to-Speech Translation System. In Elisabeth Maier *et al.*, editors, *Dialogue Processing in Spoken Language Systems*, volume 1236 of *Lecture Notes in Artificial Intelligence*, pages 195–206. Springer-Verlag, Berlin.
15. Juliane House. 1981. *A Model for Translation Quality Assessment.* Gunter Narr Verlag.
16. Martin Kay. 1967. Experiments with a Powerful Parser. In *Proceedings of the 2nd International COLING*, August.
17. K. Krippendorf. 1970. Bivariate Agreement Coefficients for Reliability of Data. In E. F. Borgatta and G. W. Bohrnstedt, editors, *Sociological Methodology.* Jossey-Bass, San Francisco.
18. R. J. Light. 1971. Measures of Response Agreement for Qualitative Data: Some Generalizations and Alternatives. *Psychological Bulletin*, 76:365–377.
19. R. R. MacDonald. 1963. General Report 1952–1963. Number 30 in Georgetown University Occasional Papers in Machine Translation. Washington, D. C.
20. A. K. Melby. 1983. Computer-Assisted Translation Systems: The Standard Design and a Multi-Level Design. In *Conference on Applied Natural Language Processing*, Santa Monica, California, February.
21. Teruko Mitamura, Eric Nyberg, and Jaime Carbonell. 1991. Interlingua Translation System for Multi-Lingual Document Processing. In *Proceedings of Machine Translation Summit III*, Washington, D. C., July.
22. M. Nagao. 1984. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In A. Elithorn and R. Banerji, editors, *Artificial and Human Intelligence.* NATO Publications.
23. Sergei Nirenburg *et al.* 1995. The Pangloss Mark III Machine Translation System. Technical Report Issued as CMU Technical Report CMU-CMT-95-145, Joint Technical Report, Computing Research Laboratory (New Mexico State University), Center for Machine Translation (Carnegie Mellon University), Information Sciences Institute (University of Southern California).
24. Sergei Nirenburg *et al.* 1996. Two Principles and Six Techniques for Rapid MT Development. In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, pages 96–105, Montreal, Quebec, Canada, October.
25. William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. 1986. *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, Cambridge, England.
26. Manny Rayner and David Carter. 1997. Hybrid Processing in the Spoken Language Translator. In *Proceedings of ICASSP-97*, Munich, Germany.
27. S. Sato and M. Nagao. 1990. Towards Memory Based Translation. In *Proceedings of COLING-90*, Helsinki, Finland.
28. J. S. White and T. A. O'Connell. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA).*
29. Terry Winograd. 1983. *Language as a Cognitive Process. Volume 1: Syntax.* Addison-Wesley.