

# An Efficient Japanese Parsing Algorithm for Computer-assisted Language Learning

Chi-Hong LEUNG  
Department of Curriculum & Instruction  
Faculty of Education  
The Chinese University of Hong Kong  
Hong Kong  
leungch@cuhk.edu.hk

Wun-Na YUNG  
Japanese Language Program  
International Center  
Keio University  
Japan  
annayung@ezweb.ne.jp

## Abstract

*Instructional grammar is often used in Computer-assisted Language Learning (CALL) and the grammatical error detection is an important feature. However, it is not an easy task in Japanese language. There is no delimiter separating consecutive words in Japanese sentences. Word segmentation is a process in which proper word boundaries are identified. Before syntactic parsing of a Japanese sentence, word segmentation has to be performed. Traditionally, the word segmentation is often followed by the syntactic parsing. An algorithm in which the Japanese word segmentation and syntactic parsing are combined into one process can increase the overall efficiency.*

Computer Assisted Language Learning (CALL) is related to the use of computers for language teaching and learning. Higgins [5] described three different models of grammar teaching: instructional, revelatory and conjectural. Instructional grammar is often used in CALL because of being computerized easily. Japanese characters are often grouped together to form words which are considered the basic syntactic and semantic units in Japanese. But when words are placed together to form a sentence, it is a character string without any indication of which sub-string as a word. In order to achieve the objective of word segmentation, a number of methods have been attempted by researchers [4,7,8,9,11]. To examine the syntactic structure of a sentence, we need the grammar and the parsing technique [6,10]. According to the language classification of Chomsky [1], there are context-free grammar, context sensitive grammar, and unrestricted grammar. The context-free grammar is a very important class of grammars because it is powerful enough to be able to describe most of the structure in natural languages. The Earley parser [3] was designed for the context-free grammar. Some examples can be found in the works of Church [2] and Wang [12]. The parser is divided into three parts: predictor, scanner and completer. In the proposed approach, the scanner is modified to segment a word that can match the terminal symbol. Hence, only segmentations that can lead to correct parses will be generated to reduce the processing time.

An experiment was performed to evaluate the efficiency of the algorithm. The test data used for this experiment were derived from the corpus Nihon Keizai

Shimbu (1993-1994) made available by the Linguistic Data Consortium, University of Pennsylvania. The average number of states generated for a sentence in these two different approaches are calculated. The new approach generated 5,289.45 states while the traditional approach generated 648,592.25 states. It is proved that the algorithm of word segmentation associated with syntactic parsing mentioned in this paper can increase the processing speed significantly.

## References

- [1] Chomsky, N., *Syntactic Structures*, Mouton, The Hague, 1957.
- [2] Church, K., Gale, W., Hanks, P., and Hindle, D., "Parsing, Word Association and Typical Predicate-Argument Relations", *Proceedings of the International Workshop on Parsing Technologies*, Pittsburg, Pennsylvania, August 1989.
- [3] Earley, J., "An Efficient Context-free Parsing Algorithm", *Communications of the Association for Computing Machinery*, 13, 1970, 94-102.
- [4] Fuchi, T., and Shinichiro T., "Japanese Morphological Analyzer Using Word Co-Occurrence -JTAG-", *Proceeding of ACL-COLING 1998*, 409-413.
- [5] Higgins, J., "The Computer and Grammar Teaching", in Leech, G. and Candlin, N.L. (eds), *Computers in English Language Teaching and Research*, Longman, London, 1986.
- [6] King, M., *Parsing Natural Language*, Academic Press, 1983.
- [7] Kurohashi, S., and Makoto, N., "Building a Japanese Parsed Corpus While Improving the Parsing System", *First LREC Proceedings 1998*, 719-724.
- [8] Murakami, J., and Sagayama, S., "Hidden Markov Model Applied to Morphological Analysis", *IPSJ 3*, 161-162, 1992.
- [9] Nagata, M., "A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A\* N-Best Search Algorithm", *Proceedings of COLING 1994*, 201-207.
- [10] Sparck Jones, K., and Wilks, Y.A., *Automatic Natural Language Parsing*, Ellis Horwood, 1983.
- [11] Takeuchi, K., and Matsumoto, Y., "HMM Parameter Learning for Japanese Morphological Analyzer", *IPSJ 1997*: 38-3.
- [12] Wang, L.J., "A Parsing Method for Identifying Words in Mandarin Chinese Sentences", *Proceedings of the Twelfth International Conference on Artificial Intelligence*, Darling Harbour, Sydney, August 1991.