Paper Title: Thai Thai Lexicon

Sub-field: Semantic, Corpus Linguistics, Lexicon construction

Key word: reduplication, lexicon construction, computational linguistics

Name of authors:
Thatsanee Charoenporn[1], Sareewan Thoongsup[1], Virach Sornlertlamvanich[1], and Hitoshi Isahara[2]

Affiliations of authors:
[1] Thai Computational Linguistics Laboratory, NICT Asia Research Center, Thailand
[2] National Institute of Information and Communications Technology

Email address:
[1]{thatsanee,sareewan,virach}@tcllab.org
[2] isahara@nict.go.jp

Presented at: The 17th Annual Conference of the Southeast Asian Linguistics Society (SEALS)
University of Maryland, College Park,
August 31st to September 2nd, 2007

ABSTRACT

Reduplication is a significant and conspicuous phenomenon in language studies. Languages around the world, especially in Asian languages, employ reduplication, that a root or stem or part of it is repeated, to express various meanings. For example, in Indonesian, reduplication indicate a plural (*anak* 'child' > *anak-anak* 'children'). In Taiwanese and Japanese, reduplication can convey augmentative degree of the root (Taiwanese: *an* 'red' > *an-an* 'red–red', Japanese: *aka* 'red' > *aka-aka* 'very red') and it can express diminution in Cantonese (*hung4* 'red' > *hung4-hung2 dei2* 'reddish').

Previous studies on Thai full reduplication indicate that reduplication attaches some additional meaning to the root, for example "plural", "not specific", "partitioning", and "emphasis" (Bhandhumedha 1971, Naksakul 1962, Chinachoti 1973). Based on the study on the contemporary corpus, it is found that reduplication serves various functions and conveys several meanings in current Thai language as expressed below. A symbol 'Mai Ya Mok' (ๆ) is always used to produce a reduplicated sound/word of the preceding word.

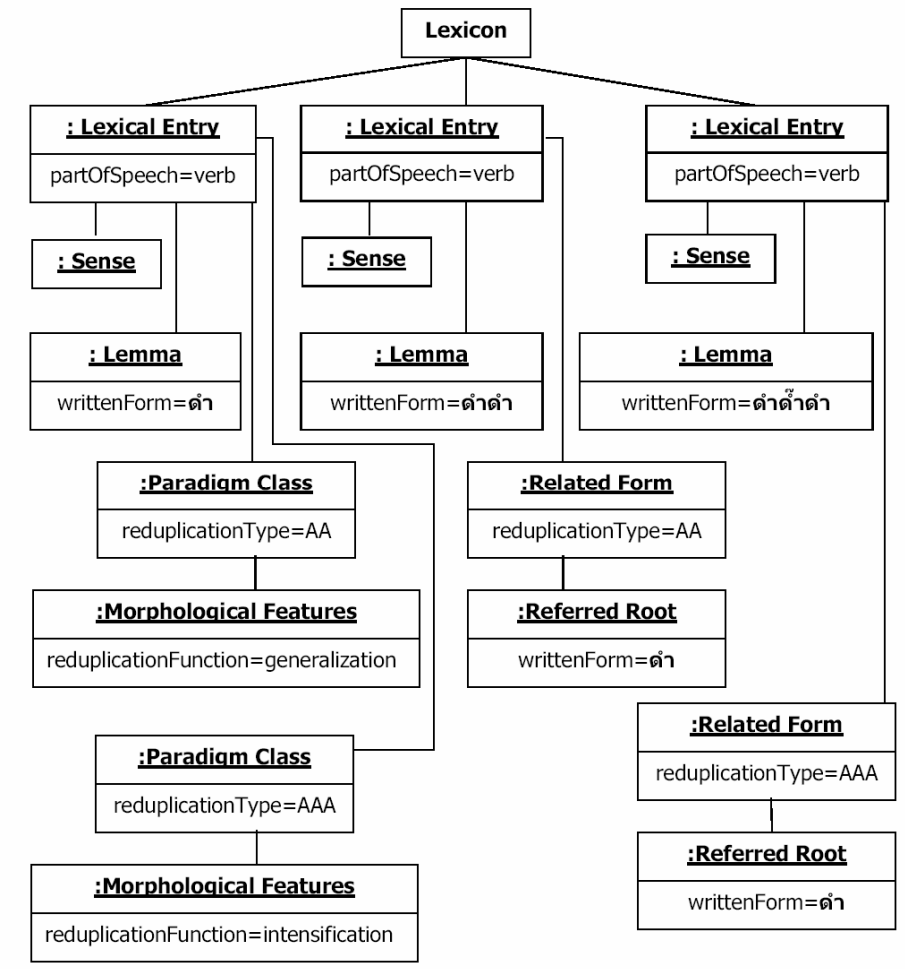
a) Plural/generalize
   เด็กๆ อ่านหนังสืออยู่ในห้อง [*เด็ก* 'child' > *เด็กๆ* 'children']
   Children are reading books in the room.
b) Plural/distribution (to express the meaning of plural and divide among several things)
   ก้าวไปช้าๆ อย่างมั่นคง *ทุก ๆ* ก้าว [*ทุก* 'every' > *ทุก ๆ 'each in every'*]
   Step slowly with steadiness in *every* step.
c) Distribution (to express the meaning of distribution, or one by one)
   เดินเข้ามาในห้องเป็น*คนๆ* [*คน* 'one' > *คนๆ* 'one by one']
   Walk into the room *one by one*.
d) Iterative/continuative (to express the repetition of action)
   แสงไฟ*ติดๆ ดับๆ* ตามข้างทาง [*ติด* 'on' > *'ติดๆ* 'on on', *ดับ* 'off' > *ดับๆ* bff off']
   The light is blinking.
e) Increase degree
   ฉันเป็นคนที่ชอบถ่ายรูป*มากๆ* [*มาก* 'much' > *มากๆ* 'very much']
   I like taking photos *very much*.
f) Decrease degree (show the reduction of meaning of the word)
   เขานั่งอยู่กลางๆ ห้อง [*กลาง* 'in the middle of' > *กลางๆ* 'quite in the middle']
   He is sitting quite in the middle of the room.

It is also found that some reduplication forms convey definitely different meaning of the base in which they are analysised as different words. In addition, the meaning of Thai Reduplication depends on the meaning of the whole token as shown in the following example. *ดำๆ* (*ดำ* 'black' > *ดำๆ* 'black black') in the declarative utterance, expresses the decrease degree or generalization, while it expresses the increase degree in imperative utterance.

ทาสีกำแพงให้*ดำๆ* Paint the wall (some more) black.
เขาซื้อเสื้อตัวสี*ดำๆ* นั้น He bought that blackish shirt.

The representation of the reduplication in the lexical repository is one another focus of our study. We represent the semantics of the term with the Lexical Markup Framework[1]. The study comes up with two ways of representation. 1) Representing as a new term for the reduplication form which has no related meaning to the base. And 2) Indicating the link between a base form and its derived forms for the reduplication which has some related meaning to the base. An operation of deriving a reduplication form and the effect of its reduplication are separating. The former is described as an attribute 'Reduplication_type' and the latter as an attribute 'Reduplication_function' in Morphological Feature Class. And the link between a base form and reduplication form are described by the shared attribute 'Reduplication_type' in Paradigm Class under the base form, and in Related Form class under the reduplication form. The following graph illustrated the reduplication term (*ดำ* 'black' > *ดำดำ* 'blackish').

[1] Lexical Markup Framework is high level model for representing data in lexical databases used with monolingual and multilingual computer applications, proposed to ISO International Standard.



Acknowledgement

This research is a part of the Development of Language Resource Standard for Semantic Web Applications Project, under the financial support of the NEDO International Joint Research Grant Program. Thanks for the members of the project for all valuable discussion, comments and supports.

References

F. Bertagna, A. Lenci, M. Monachini, and N. Calzolari. 2004. The MILE lexical classes: Data categories for content interoperability among lexicons. In T. Declerck et al., editor, A Registry of Linguistic Data Categories within an Integrated Language Resources Repository Area, LREC2004 Satellite Workshop, Lisbon, Portugal.

ISO2006. Language resource management – Lexical Markup Framework (LMF), ISO/TC 37/SC4N130 Rev. 13, ISO CD24613:2006.

K. Chinachoti, 1973. "A Comparison on One type of Reduplication in Thai and Cambodian" Master's Thesis, Chulalongkorn University. (in Thai)

K. Naksakul "A Study of Cognate Words in Thai and Cambodia" Master's Thesis, University of London, London, 1962.

N. Bhandhumedha, 1971. Thai Usage. Bangkok. pp. 62-67. (in Thai)

S. F. Chung, K. Hasan, T.J. Jiang, S. Lee, I-L. Su, L. Prevot and C.R. Huang. 2006. Extending an international lexical framework for Asian languages, the case of Mandarin, Taiwanese, Cantonese, Bangla and Malay. Taipei: Academia Sinica. pp. 87-94.

T. Charoenporn, V. Sornlertlamvanich, and H. Isahara. 1997. Building a Large Tai Text Corpus – Part of Speech Tagged Corpus: ORCHID. In proceedings of the Natural Language Processing Pacific Rim Symposium, pp. 509-512.

T. Tokunaga, V. Sornlertlamvanich, T. Charoenporn, N. Calzolari, M. Monachini, C. Sonia, C.R. Huang, Y. J. Xia. 2006. Infrastructure for Standardization of Asian Language Resources. Proceedings of the COLING/ACL 2006, pp. 827-834.