辞典とその電子化── タイ語を例に

AA 研共同プロジェクト「語彙と文法」研究会報告要旨
2010/1/23
東京外国語大学　アジア・アフリカ言語文化研究所
峰岸真琴

概要

以下の 4 点についての報告と考察を行った。

1. **コーパスに基づく言語学教育研究拠点**　グローバル COE プログラム（Corpus-based Linguistics and Language Education），以下 CbLLE プログラムと略称。
2. **タイ日・日タイ辞典の電子化プロジェクト**
3. **電子化辞典の言語研究面での発展的継承**　教育面だけでなく，学術研究においても，今後発音，品詞，文法，借用関係などの情報を付与していくことにより，Machine Readable Dictionary, MRD の開発へ。
4. **語彙研究とタイ語文法研究**

上記のうち，1 ～ 3 までに関しては峰岸・赤木 (2009b) 1 ～ 3 を参照しながら，その概略を紹介した。言語学に関わる問題として，4 を中心に報告した。

1　「コーパスに基づく言語学」拠点の研究戦略

CbLLE プログラムは，先端的な言語学・言語教育学の拠点を形成することを目的に，2007 年度から 2011 年度にわたるグローバル COE 拠点プログラムとして文部科学省により採択された。

`http://cblle.tufs.ac.jp/dic/th/thja/`

コーパス　言語のデータ，それも発音，品詞，文法など，言語の分析に必要な情報を与えられ，構造化された言語データの集合体。

コーパスに基づく言語学　Corpus-based Linguistics とは，大量データの統計処理などの情報工学的研究に代表される，いわゆる「コーパス言語学」だけでなく，言語運用の実態を明らかにすることを目的とした，言語データに基づく実証的な言語研究を包含するもの。

CbLLE プログラムでは，情報基盤および学術の観点から，世界の言語を大きく 3 つの言語グループに分ける。

大言語　電子辞書および大規模言語コーパスの両者が容易に入手できる言語。
中言語　電子辞書あるいは大規模コーパスのいずれかが存在するか，あるいは容易に入手可能な言語。
小言語　電子辞書も大規模コーパスも存在しないような言語。

上記大，中，小の三言語グループに対応して，それぞれ研究上の戦略を立案した。

大言語　分析ツールの開発や，コーパスを利用した言語学的研究を推進。
　　話し言葉コーパスなど，目的別コーパスを構築。
中言語　電子辞書や分析用のツールを開発しつつ，コーパス構築のための基礎的研究を行う。
小言語　一次資料の収集から，言語コーパス構築に至る研究過程を再検討し，基礎研究を充実。

これらは CbLLE プログラムのための研究戦略であるが，アジア・アフリカの多様な言語文化の研究を目的とする AA 研において，諸言語の研究を今後どのように進めるのかとも深く関係するものである。

2　タイ日・日タイ語辞典電子化プロジェクト

タイ国内においても，さまざまなタイ語のテキストが電子化され，電子コーパス化が進行中であるが，多くは著作権上の問題があり，簡単に利用できない状況である。

現在，CbLLE プログラムの下でのタイ日・日タイ辞典の電子化が進行中である。

冨田 (2003)　『タイの人々のための日タイ・タイ日辞典』改訂新版：タイ人が日本語を学ぶための辞典として編纂されたもので，タイ語および日本語の見出し語はそれぞれ約 20,000 語。

日タイ辞典：　ローマ字，漢字かな交じりの日本語見出しについて，タイ語の訳語。

タイ日辞典：　タイ文字で書かれたタイ語に対して，日本語の訳語。発音記号は付けられていない。

冨田 (1997)　『タイ日大辞典』: 約 50000 語の見出し語の他，タイに関する豊富な文化・社会情報を含む，最大のタイ日辞典。

上記両辞典については，CbLLE プログラムの活動により，東京外国語大学が，原編著者の著作権継承者から学術・研究上の使用，電子的複製，記述項目・記述内容の増補改訂に関する許諾を取得した。

ネットワークを通じた辞典の利用に関するプロジェクトの初期段階の作業は一応終了し，その成果は以下のサイトにおいて試験的に公開されている。

`http://cblle.tufs.ac.jp/dic/th/thja/`
`http://cblle.tufs.ac.jp/dic/th/jath/`

3　言語研究のための辞典の発展的継承

3.1　町田式電子辞典開発とタイ語文字処理の現状

アジアの国家レベルの公用言語の電子化辞典としては，既にヒンディー語その他の南アジア諸言語について，インド系文字の使用が可能な e-dictionary が町田和彦氏によって以下のサイトで公開されている。

`http://www3.aa.tufs.ac.jp/~kmach/gicas/LOSAL/platform/hje_pt.htm`

以下に町田式システムと現状のタイ語辞典データベースの特徴を対比して述べる。

町田式 e-dictionary システム　「基底形」（原音韻表記と 1 対 1 に対応するローマ字翻字）→「表層形」（インド系文字）を表示・印字する。

タイ語辞典のデータベース　「表層形」（タイ文字データ）の処理を基本とする。

現実に存在・流通する「タイ語文字データ」の処理系としては，**当面は適切**であるが，今後の言

語処理に関しては，町田式システムへと改良を進める必要がある。。

以下では，タイ語の「文字処理」と言語研究の上で必要になる「言語処理」のために必要な改良作業について述べる。

3.2 東南アジアのインド系文字

以下では東南アジアのインド系文字の特徴をまとめておく。

東南アジアのインド系文字 全てが南インド系文字（母音記号 “e” を音節初頭子音字の左側に置く）である。

cf. 北インド系文字（母音記号“e”を音節初頭子音字の上に置く。）

語の分かち書きをしない 言語処理の際には節から語を抽出するプロセスが必要となる。

タイ語やラオス語 表層形における文字構成素の順に，データストレージが行われている。

cf. カンボジア語（クメール文字）: Unicode の原則に従い，例外的に音韻表記に近い順でのコード付けが行われている。

3.3 タイ文字とその言語データ資産

1. 文字構成とコード，データストレージ，基底から表層を導くことはできるが，逆に表層から基底にさかのぼることはできない。
2. サンスクリット固有の発音を表記するために，最も保守的な子音文字体系を保つ。(←→ラオス文字)
3. 一方で，声調表記，母音表記などに極端な革新性を見せる。
4. 一部の例外的表記は碑文に遡る。rr=/aM/ eg. ダルマ dhrrm=/thamma/
5. タイは東南アジア大陸部で最も情報化が進展している国である。かつてのタイプライターの普及度は高く，現在は電子化テキストが豊富，web も充実している。

3.4 冨田タイ語辞典の発展的継承

電子辞書は人文科学における言語，社会の教育と研究に有用なだけでなく，さらに自然言語処理に必須の機械辞書の開発へと発展させることで，自動翻訳や多言語検索などの産業分野への応用も期待される。

3.4.1 日本語教育への応用

1. 社会変動に伴い，新概念を表す語彙は常に生じているため，辞典は常に増補・改訂が必要である。そのためには，言語教育・研究の専門家，人文社会系，理工学系の研究者が共同して改訂，改良にあたる仕組みを作っていかなければならない。
2. CbLLE プログラムの活動により，2009 年 3 月に東京外国語大学・泰日工業大学 (Thai-Nichi Institute of Technology) 間の学術協定を締結した。

3. タイは高等教育機関だけでなく，中等教育，民間レベルでも日本語教育が盛んである。タイ人の日本語学習者数は６万人を超える。

3.4.2　タイ語の学術的研究の基盤データとしての発展

学術的研究の基盤としての辞典データを用いることにより，以下のような研究が進展することが期待される。

タイ語研究　発音表記，品詞（語類）情報を加える必要がある。タイ語の綴り字から発音が一義的に定まらない場合があるため，タイ文字データから機械的に発音を導くことはできない。

音声学・音韻論的な研究　現在行えば膨大な手作業が必要である。

形態論　複合語形成がどのような語類の組み合わせにおいて可能かと言った研究が進む。

孤立語の品詞分類の問題　暫定的に，名詞，動詞，数詞といった品詞を付与した MRD を開発し，それをコーパスデータの統語論的研究に使いながら拡張し，各語について n-gram 法などにより共起制限を分析，語類を細分化することができる。

語彙論研究　語源および借用語に関する情報を付与していく必要がある。

3.4.3　多言語情報基盤の一環としての発展

人文科学的研究だけでなく，情報工学についても，以下の各分野に研究が発展する可能性がある。

1. 音声コーパス構築，音声認識システムなどの開発。
2. 文字処理，自動翻訳などの開発。
3. 語源・借用語情報の付加によるアジア諸言語の串刺し曖昧検索。
4. 冨田大辞典のタイ文化，タイ文学の情報は，維持・継承すべき貴重な遺産である。
5. 品詞分類から，非西洋的オントロジー構築へと結びつく可能性がある。

タイ語について本報告で指摘してきた辞典編纂の重要性を考えると，今後アジア・アフリカ諸言語の研究および研究資産の継承・発展について，どのような戦略を立てていくのかが課題である。

4　語彙研究とタイ語文法研究

以下は辞典の編纂と関わってくる，タイ語の語彙研究および言語学的な問題の例である。定性的分析にも、「数」が問題になるはずである。

1. 音韻，特に２音節以上の多音節語における軽声化の問題
2. 品詞（語類）の分類、さらには動詞の下位分類の問題

以下は，峰岸 (2009) の口頭発表の発表資料からの抜粋である。

4.1 タイ語の境界指示機能の例

２音節語　タイ固有の語（およびクメール語からの借用語）は，常に {副音節 (minor syllable) + 主音節 (major syllable)} で構成されている。それぞれの音節に現れうる母音と声調は，副音節では /a, i, u/の３種, 低平調あるいは高平調の２種に限られているが，主音節では，母音９種，５声調。　→一定の軽重，弱強のリズムパターンを形成する。cf. 中国語の軽声, r 化

多音節語　パーリ語，サンスクリット語からの借用語は多音節が基調であるが，上記の２音節語に準じるか？

近年の英語からの借用語はどうか？例えば，最終音節が第５声（下降調）となる傾向： miitôə < E. meter

軽音節の声調は？　音声的観察・記述が必要 例 : rawaŋ《注意する》 vs. ráwàaŋ《間》 **声調言語のアクセント記述は未開拓の研究分野である。**

4.2 他動性

他動性 (transitivity) については，角田他編 (2007: 3-11) を参照のこと。

以下の考察は角田他編 (2007) 所収の峰岸 (2007b) からの抜粋と再検討である。以下のような語類の分類と分析（動詞の下位分類など）を行う際には，語彙集合全体のどのくらいのメンバーについて考察，検証するかを念頭に置くべきであろう。

4.2.1 二項述語の分類

表１は，角田他編 (2007）の「二項述語階層」（他動性は，１から７に向かって低下する，とされる）に従って，いくつかのタイ語の動詞の意味的な分類を試みたものである。

表 1　タイ語の二項動詞の分類

類	意味	タイ語の二項動詞の例
1 類	直接影響	
1A 類 1B 類	変化 無変化	khâa 「殺す」, ʔùn 「温める」 tii 「たたく」, tèʔ 「蹴る」chon 「ぶつかる」
2 類	知覚	
2A 類 2B 類		hěn 「見える」, dâyyin 「聞こえる」, cəə 「見つかる」 duu 「見る，眺める」, faŋ 「聴く」
3 類	追求	rɔɔ 「待つ」, hăa 「探す」
4 類	知識	rúu 「知る」, khâwcay 「理解する」, cam 「覚える」, luɯum 「忘れる」
5 類	感情	rák 「愛する」, chôɔp 「好む」, klìat 「嫌う」, kròot 「怒る」, klua 「恐れる」
6 類	関係	mii (持つ，ある」, mŭan 「似る」, khláay 「似る」, khàat 「欠ける」, pen 「成る」, ruam 「含む」
7 類	能力	V dây 「V が可能である」, V pen 「V が可能である（習得)」, kèŋ 「上手である」, thon 「耐える，強い」

以下，表１についての注意点を挙げる。

1. 他動性と動詞の結合価とは同値ではない。二項動詞であっても，P pùat C「P は C が痛い」などは他動性は低い。
2. タイ語（その他東南アジア大陸部の孤立語的な諸言語）では，意味的な特性である**随意性 voluntariness**（以下，**「する動詞」**と呼ぶことあり）と**不随意性「なる動詞」**と呼ぶことあり)」の方が，結合価や他動性よりも重要ではないかと考える。一項動詞であっても wîŋ「走る」のような動詞は「する」類動詞であるし，二項動詞であっても，上記 pùat C「C が痛い」のように，他動性・随意性が低い「なる」類動詞もある[*1]。
3. 「随意性と不随意性」は，コントロールに関する「意志性 volitionality と無意志性」や「他動性と自動性」，「動作性と状態性」の概念と一部重なりつつも，完全に一致はしない。

   随意性　有情物である動作主が，ある動作をコントロールできるのは，その動作を行おうと企てることまでである場合，これを随意的動作と呼ぶ。動作あるいは動作の結果が実現されるかまでは含意されない。

   不随意性　事態（有情物の行為の結果，無情物の状態，状態の変化といった出来事全般）に関して，その事態の成立そのものに着目し，特定の関与者によるコントロールに着目する必要のない事態を不随意的事態と呼ぶ。

   定義により，意志性の含意する意味内容からコントロール可能性を分離した上で，より狭く定義したものが随意性である。
4. **（今後，さらに検討を要する課題）**「随意性と不随意性」は語および文全体の意味に関わるだけでなく，タイ語（その他東南アジア大陸部の孤立語的な諸言語）におけるヴォイス，モダリティー，動詞連続といった構文上の問題の理解と深く結びついている。

4.2.2　動詞の随意性と不随意性

表２は「する」「なる」両類の動詞を区別する基準を示したものである。

表 2　　タイ語動詞の「する」「なる」の類別

	動詞の類別	する類	なる類
(1-1)	事態の実現を「随意」に企てられるか	＋	−
(1-2)	否定の場合，行為の意志まで否定されるか	＋	(−)
(1-3)	行為として命令できるか	＋	(−)
(1-4)	事態の実現を目的化できるか	＋	(−)
(2)	結果や状態が含意されるか	−	＋

ただし，記号 (−) は，そのような特徴を持たないのが一般的だが，例外的に持つ場合があることを示す。

表３は，表１の内容を，タイ語における統語現象のうち，「受け身文化」「意志の否定」「命令化」「結果の含意」の４つを基準として並べ替えたものである。ただし，表１の 1A 類と 1B 類の区別は，タイ語では無関係であるため，一つの類にまとめてある。

[*1] 「する」「なる」の概念については峰岸 (1986) を参照のこと。

表 3　「する」「なる」の類別と，角田の二項述語階層との対応

随意性の類別	する			なる				
受け身文化	＋	(−)	−	−	(−)	(−)	−	−
意志の否定	＋	＋	＋	−	(−)	(−)	(−)	−
命令化	＋	＋	＋	−	(−)	(−)	(−)	(−)
結果の含意	−	−	−	＋	＋	＋	＋	＋
角田の類	1 (1A+1B) 直接影響	3 追求	2B 知覚	2A 知覚	4 知識	5 感情	6 関係	7 能力

ただし，記号 (−) は，そのような特徴を持たないのが一般的だが，例外的に持つ場合があることを示す。

表 3 を見ると，タイ語では，被動作性および二項述語の分類よりも，随意性および不随意性という意味上の対立が，統語現象との関わりにおいて重要であると考えられる。

改めて考えてみると，**表 3 では，＋/− について例外のない 2A こそ，典型的な「なる」動詞**であって，表の右端に置くべきであった。

さらに対照言語学的観点から興味深いのは，2A には日本語「見ゆ，聞こゆ，見える，聞こえる」などの**自発の動詞**にあたるタイ語の動詞が含まれていることである。このような理由から，筆者は，日本語を含めた対照研究において，能動・受動，自・他の区別だけではなく，自発性，不随意性といった意味・統語上の特徴にも着目する必要があると考える。例えば，日本に隣接する東南アジア島嶼部のタガログ語などのヴォイスは，対格型でも能格型でもないフィリピン型の言語である。(角田他編 (2007) 参照。) Agent, Patient 以外のどのような指示対象の意味役割に注目するかは，「主題化」，focusing といった構文上の問題を考える上で重要である。

4.3　更に検討すべき事

以下の 2 点について問題提起を行った。

1. 孤立語の「形態論と統語論」cf. 峰岸 (2006b)
2. 動詞連続は統語論であつかうべきか？ cf. 峰岸 (2006a)

参考文献

Minegishi, Makoto (2004) 'Southeast Asian Languages: A Case for the Caseless?', *Non-nominative Subjects*, Peri Bhaskararao and K.V. Subharao (eds), pp. 301–317, John Benjamins. Amsterdam/Philadelphia.

Makoto MINEGISHI and Osamu AKAGI (2009) "Development of Electronic Dictionary for Analyzing Linguistic Data", Proceedings of Chula-Japan Linguistics Conference 2008, pp.11-16, 東京外国語大学.

峰岸真琴 (1986)「クメール語の動詞連続における/baan/の意味について」,『東京大学言語学論集'86』pp.45–57. 東京大学文学部

峰岸真琴 (2006a)「動詞連続と言語理論の諸前提」, pp.191-211, 東ユーラシア言語研究会編『東ユーラシア言語研究』, 第 1 集, 2006.3.21. 好文出版.

峰岸真琴 (2006b)「形態論と統語論」『言語基礎論の構築に向けて』pp.109-128, アジア・アフリカ言語文化研究所.

峰岸真琴 (2007)「孤立語の他動詞性と随意性：タイ語を例に」角田光枝，佐々木冠，塩谷亨編『他動性の通言語

的研究』pp.205-216, くろしお出版.
峰岸真琴・赤木攻 (2009a) 「『コーパスに基づく言語学』プロジェクトにおける電子辞典開発 — タイ語を例に —」峰岸真琴，川口裕司（編）『コーパスに基づく言語学教育研究報告 3 – フィールド調査，言語コーパス，言語情報学』, pp.183-194, 東京外国語大学 2009.5.15.
峰岸真琴・赤木攻 (2009b) 「タイ語辞典の電子化から言語研究へ」人工知能学会第 2 種研究会 第 32 回ことば工学研究会資料 SIG-LSE-A901, pp.21–25, 2009.7.10.
冨田竹二郎（編著）(2003)『タイの人々のための日タイ・タイ日辞典』改訂新版，367pp+270pp. バンコク：タイ教育文化振興協会.
冨田竹二郎（編著）(1997)『タイ日大辞典』，2336pp. 日本タイクラブ：めこん.
角田光枝，佐々木冠，塩谷亨編 (2007)『他動性の通言語的研究』pp.205-216, くろしお出版.
http://unicode.org/

[口頭発表]

峰岸真琴 (2009) 「対照研究から言語類型論へ」，国際日本研究センター研究会 (2009.12.19) : 東京外国語大学語学研究所.