4J-13

タイ語字書の入力法と入力特性

柴山　守
(京都大学東南アジア研究センター)
星野　聰
(京都大学大型計算機センター)

1．はじめに

計算機による自然言語処理、機械翻訳など、言語の機械処理において、辞書は貴重な資料である。その内容、構成が直接その処理系に反映される処理系の重要な一構成要素であり、辞書のデータベース化には膨大な人手と費用が費やされるが、関連の事例報告は少ない。また、最近では研究の進展につれ、母国語でない言語テキストを短期間で扱わねばならない状況もある。我々は、タイ国三印法典(1805年編纂、約1700頁)テキストのセグメンテーションやタイ語構文解析を試みるために、約6週間(3名が担当、延べ約9週間・人)を費やして、タイ語辞書の見出し語32,404語の入力を完了し、現在、品詞の入力及び字書の修正作業を進めている。

本稿では、計算機で扱う字書作成の一事例として、マイコンを用いて独自に開発したタイ語テキスト・エディタ[1]におけるタイ語入力法とタイ語字書作成時の入力速度の測定及び評価について述べる。

2．タイ語入力法

タイ語は、子音字母44字、母音字母32字から成る表音文字で、各々の字体は、82(旧字体を含み、特殊文字を除く)の部分品に分割できる。入力には、2つの方法があり、一つは、部分品の各々がキーボード上のいづれかのキーに一意に対応させることにより、入力を実現する方法である。IBM 電子タイプライタ等、現地タイ国で一般的に用いられる方式(これをダイレクト・マッピング法：DMM と呼ぶことにする)で、この配列を図1に示す。必要とされるキー数は、92(特殊記号を含む)である。一方、タイ文字の発音をローマ字表記で入力し、タイ文字へ変換するローマ字・タイ字変換方式(これをトランスリテレーション法：TMと呼ぶ)である。この方式は、使用されるキー数が49(DMM比 53.3%)であり、例えば、ค、ฆ、ฃ、ฅ はKH、KH1、KH2、KH3 のように入力する。キー・ストローク数で

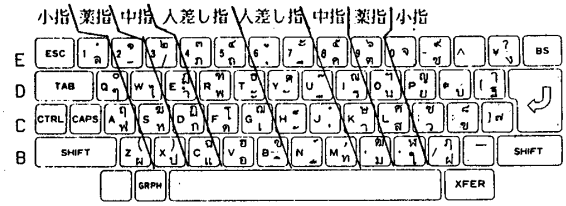

図1　DMMのキーボード配列

は、DMMのキー数 92に対し、1.91倍となるが、タイ文字の入力に通常の英字配列キーボードが使用できること、タイ語を母国語としない者にとって扱い易い利点がある。

3．タイ語字書作成における入力速度の測定と評価

3．1　文字出現頻度

辞書に含ませる見出し語の文字数は、217,929 字で、72字種(数字を除く)のすべてが出現し、子音 63.5%、母音 29.8%、声調 5.5%、その他1.2%の出現率である。これらの文字頻度分布を図2に示す。X軸は、順位を示し、Y軸は出現確率を示す。子音、母音、共に約半数について出現率が 0.005以上である。辞書は、子音に従って順序正しく配列されているため、一般のテキストと比べ、タイピストの習熟度は高いと考えられる。実際の入力作業では、各語の先頭に繰返し同一文字列が現れる場合、特殊記号で置き換えることにより、入力キー・ストローク数を減少させた。また、この見出し語の入力にTMを用いると、DMMより 42.9%キー・ストローク数が増加することが判った。

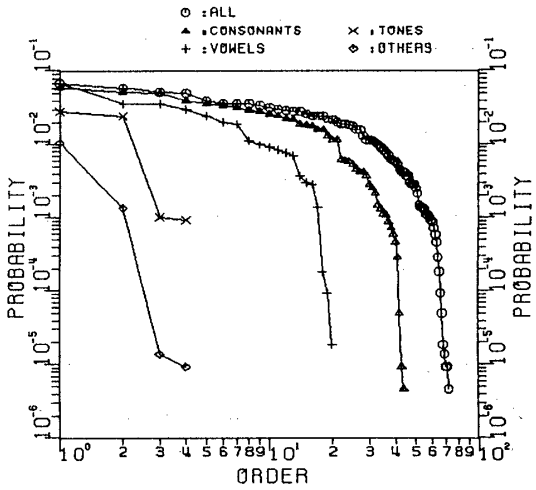

図2　文字出現頻度分布

Methods and Characteristics of Thai Dictionary Inputs

Mamoru SHIBAYAMA, Satoshi HOSHINO
Kyoto University

3．2　入力速度の測定と評価

タイ語字書作成におけるタイピストの動作は、図3のようになる。入力作業には、3名が携わり、測定の被験者は2名である。入力法は、DMM、TMの順に用いた。被験者2名の入力テキスト量は、表1のとおりで、入力語数30,084 語、文字数 191,248字である。また、両被験者ともタイ語に関する予備知識はなく、英文で約 200ストローク／分のタイピストである。

まず、指、手、および、キーボード上の各段の使用率は、ほぼ図1に示す指、手の使用によって、図4に示す結果が得られた。DMM、TM共に利き指度が最も低いとされる小指の使用率が高く（英文標準：左 10.1%、右 13.5%）、また、左手に比べ右手使用率が高すぎる（DMM：24.8%、TM：17.0% ）ことで、入力速度の向上に悪影響を与えている。段使用率は、DMM、TM共にホームポジション段Cの使用率が高く、平均移動距離は、$d_{DMM}=0.185*1+0.349*0+0.295*1+0.171*2=0.882$ 、$d_{TM}=0.63$ である（英文標準 $d_E=0.66$、ﾘｺｰ $d_2=0.6$、JIS $d_J=0.91$）。[2)]

次に、入力作業におけるDMM、TM両方式の測定結果と習熟曲線を被験者(A) についてのみ図5に示す。図のX軸は、入力作業の経過時間を示し、Y軸は、単位時間（分）に入力された文字数である。習熟曲線を表すために、次式を用いる。[2)]動作速度S(t)は

$$S(t)=M(1-e^{-Gt})$$

ここで、Mは限界速度、Gは練習効果指数、t は練習時間である。実測値を上式にあてはめると、DMMではM=37.25 、G=0.0527、TMではM=40.9 、G=0.0797 の結果を得た。特にTMでは、タイ字ローマ字変換テーブルを目視する時間が加算され、キー・ストローク数が1.43倍（入力文字 40字/分で57.2ストローク）になるにもかかわらず、入力速度はDMMを上回っており、タイ語を母国語としない者にとってTMが有効であることが確認された。また、両方式による限界速度の推定値Mは、測定時間が短いため最適な値ではないと考えている。

4．　おわりに

タイ語字書を作成するために、辞書の見出し語に対する文字出現頻度を求め、指、手、段使用率を示して、テキスト入力作業における入力速度の測定と結果を示した。

測定結果は、タイ語を母国語としないタイピストの短時間の作業において、DMMよりTMの入力速度が速く、TMの有効性が確認できた。これは、特殊な文字パターンを扱う場合、ローマ字表記により文字パターンを記号化することで、タイピストの学習効果が高められた結果だと考えられる。また、本報告は、言語の機械処理の研究の進展とともに扱われる言語の種類も増加し、多様の辞書作成が必要となるが、特に、アジア・アフリカ諸言語の辞書作成やテキスト入力の際の事例として有効であると考える。

最後に、タイ語の指導をいただいた京都大学東南アジア研究センター所長　石井米雄教授に謝意を表するとともに、字書作成には、文部省科学研究費補助金の援助を得た。

[参考文献] 1)SHIBAYAMA.M,HOSHINO.S:Implementation of an intelligent THAI computer terminal,JIP,Vol.8,No.4
2)村山登：2ストローク法、情報処理、Vol.23,No.6

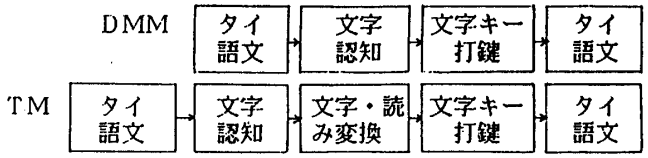

図3　　タイ語テキスト入力プロセス

表1　字書作成における入力テキスト量

入力	被験者(A)		被験者(B)	
方法	語数	文字数	語数	文字数
DMM	12,455	78,340	4,478	29,181
T　M	8,490	54,527	4,661	29,200
合　計	20,945	132,867	9,139	58,381

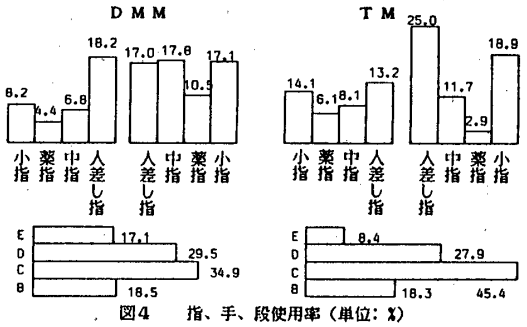

図4　　指、手、段使用率（単位：%）

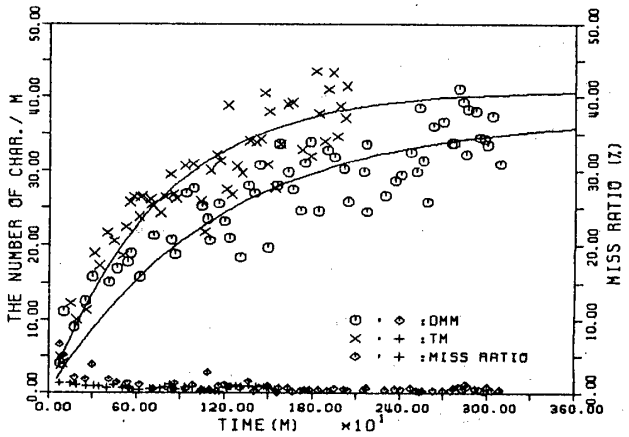

図5　作業時間と習熟曲線［被験者(A)］