タミール言語の統計的性質

3Q－5

武富　敬、スハルナン・シバスンタラン、松永一一美
佐世保工業高等専門学校

1.はじめに

タミール語は、南インド、スリランカ、マレーシアなどで使用されている言語であるが、文字種の多様さや複雑さ、文字長の不均一性などのためにコンピュータ化が進んでいない。そこで昨年、それらを解決するための基礎技術（フォント生成、コード割当て、入力方式）を開発し［1］、それらを基にタミール語ワードプロセッサを開発した［2］。これにより、タミール語、英数字、日本語の混在した文書を容易に作成できるようになった。

今回、こうして蓄積した機械可読な文書（テキストデータベース）を統計処理するためのシステムを開発し、タミール語の持つ統計的性質について調べた。その結果について報告する。

2.入力方式の改良

タミール文字は247種あり、それらを全部キーボードに割り振るのは実用的でない。そこで、昨年のタミール語ワープロでは、子音18個と母音12個をそれぞれ左右に配置したキー配列を採用した。しかし、この方式では、その新しいキー配列に習熟する必要があった。今回この点を改良し、タミール文字の母音と子音の読みをアルファベットに対応させる方式を考案した。その入力対応表を図１に示す。この改良により、入力時の手間が格段に軽減された。

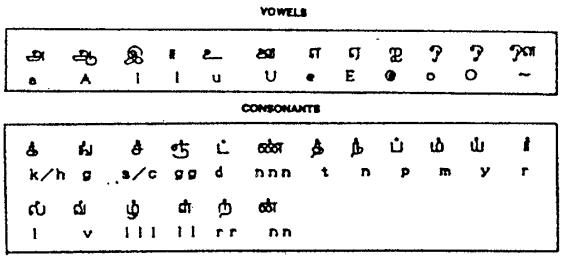

図１.タミール文字入力対応表

3.統計処理システム

これまでに蓄積したテキストデータを表１に示す。タミール文書では、図２に示すように一種の「分ち書き」が行われるので、”空白”を単語取り出しなどの処理の区切り記号として取り扱うことができる。ここでは、まず基礎データを収集する目的で、表２に示す統計処理を行うシステムを開発した。ここで、”単語”とは、空白で区切られた文字列を意味する。

（１）文字の度数・・・文字の出現頻度の計数
（２）文字数別の”単語”の度数・・・文字数別の”単語”の出現頻度の計数
（３）文字相関・・・”単語”の中で隣接するｎ文字（ｎ＝2～5）の相関

இலக்கணநூலாவது உயிர்தோ் வழங்குமுறையும் மெய்யுள் வழக்குமுறையும் அறிந்து விதிப்படி எழுதுதற்கும் பேசுதற்கும் கருவிகள் ஆகின்றன அந்நூல் எழுத்ததிகாரம் சொல்லதிகாரம் பொருளதிகாரம் என மூன்று வகைப்படும் இலக்கண நூல்லாவது யாது அந்நூல் எத்தனை வகைப்படும் எழுத்துகளின் வடிவங்கள் எழுதுவது சொல்லிந்த முதற்காரணமான ஒலியாகும் அவ் வழுத்து உயிரெழுத்து மெய்யெழுத்து ஆயுத எழுத்து என மூன்று வகைப்படும் உயிரெழுத்துகள்

図２.タミール文字テキストデータの例

表１　タミール語テキストデータ

タイトル	読　み	分　野	文字数
[illegible]	pukull	小　説	58321
[illegible]	llakanasurukam	文法書	41496
[illegible]	arulury	その他	5017
計			104834

表２.タミール語統計処理システム

	name of the program
frequency of one character	tmlchar
frequency of a word length	tmlwlen
correlation of successive n characters in a word　(n = 2 ～ 5)	tml2cor tml3cor tml4cor tml5cor

'Statistical Properties of Tamil Language
Hiroshi Taketomi, Suharnan Sivasundaram
and Kazumi Matunaga
Sasebo College of Technology, Sasebo, Nagasaki
857-11 Japan

4.統計的性質

1） 文字の出現頻度

タミール文字は247種あるが、その使用には著しい偏りがある。図3に、その様子を立体棒グラフで示した。これから全く出現しない文字がかなり存在することが分る。また、その使われ方には、分野間（小説と文法書）では大きな差異は見られなかった。図４に、横軸に出現頻度の順に並べた文字を、縦軸にその文字の度数をとった場合の累積度数分布を示す。これから、タミール語全文字種のうち、１０５種で全体の使用の９５％を占めることが分る。

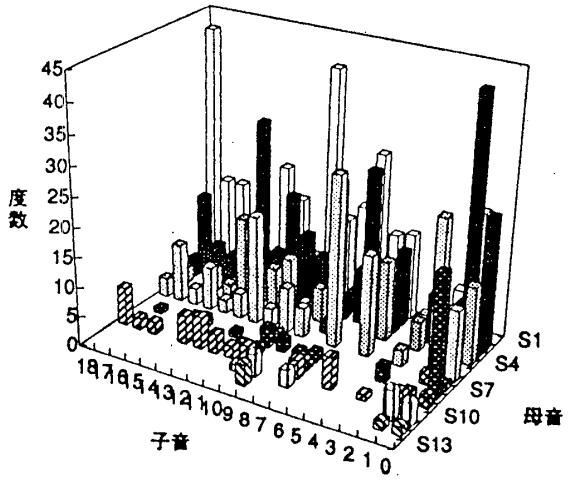

図３.タミール文字の出現頻度の度数分布

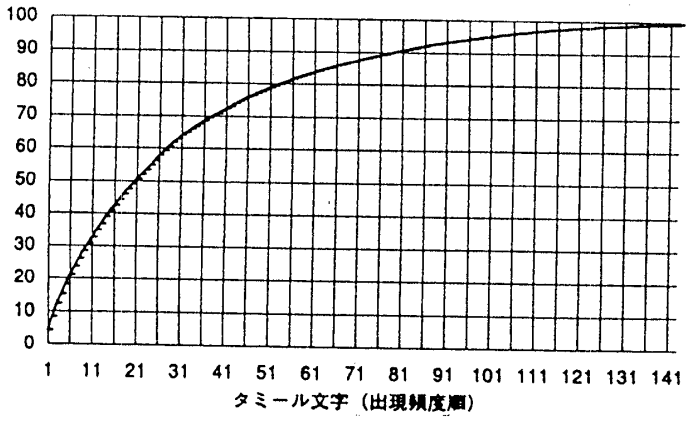

図４.タミール文字の出現頻度の累積度数分布

2） 文字数別の”単語”の度数

図５に、文字数別の”単語”の度数分布を示す。これから、”単語”は主として２文字から６文字で構成され、平均値は４.３７文字であることがる。

3） 文字相関

２文字の相関で意味のある単語が現れるのは、頻度順に [illegible] 、であり、３文字では、 [illegible] 、４文字では [illegible] 等である。

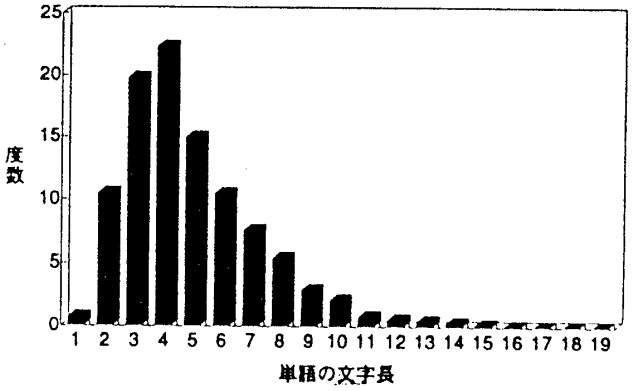

図５.文字数別の”単語”の度数分布

5.おわりに

タミール語では分ち書きが行われるが、その文字列から単語そのものを取り出すことは日本語と同様に難しい。そのため、ここでは基礎データ収集として簡単な解析にとどめた。これ以上の解析にはどうしても辞書が必要になる。現在、タミール語・日本語電子化辞書の設計を進めている。

参考文献

[１] 武富、横山、スハルナン、松瀬：タミール語ワードプロセッサ開発における基礎技術、情報処理学会九州支部研究会報告、ｐｐ.７－１０（１９９３）

[２] 武富、横山、スハルナン、松瀬：簡易タミール語ワードプロセッサ、情報処理学会第４６回全国大会講演論文集（３）、ｐｐ.２４３－２４４（１９９３）