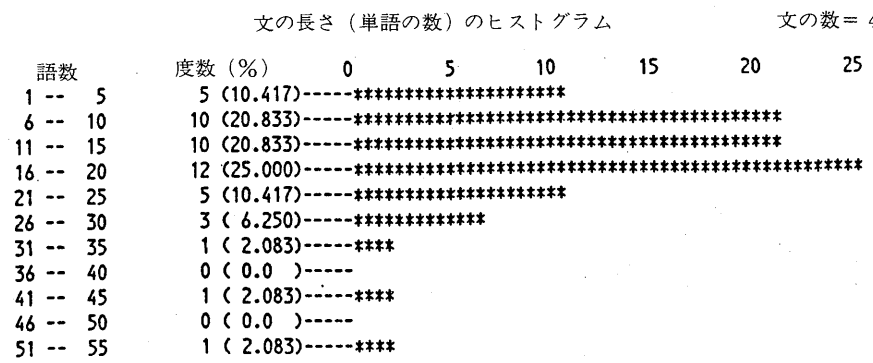

図1. 文の長さの分布.

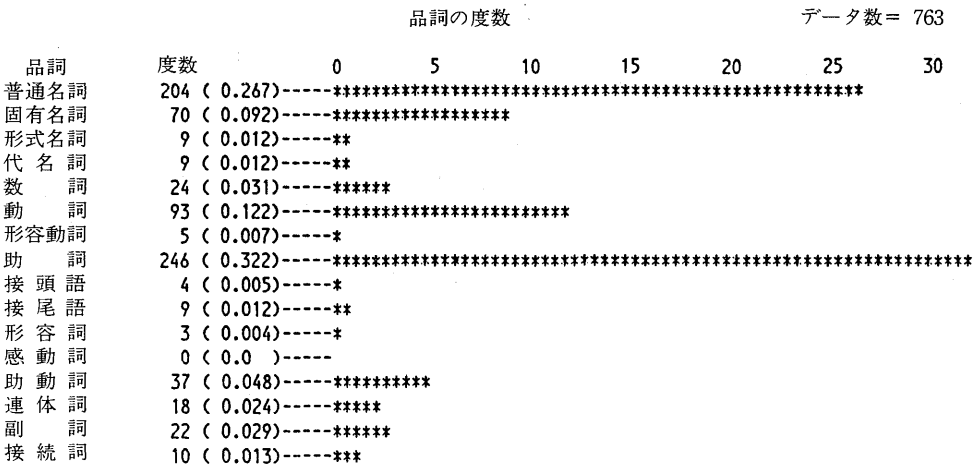

図2. 品詞の出現率.

人文科学とコンピュータ

―― 国立民族学博物館のマルチメディアデータベース ――

国立民族学博物館 杉田繁治

国立民族学博物館（通称 民博，梅棹忠夫館長）は，博物館という名称がついているが，文部省に所属する共同利用機関であり，研究所である．昭和49年6月に設立され，昭和52年11月に展示場が一般に公開された．国立大学と全く同じ身分の専任教官が現在60数名おり，共同研究に参加している研究者は300名に近い．平成元年4月から総合研究大学院大学が発足し，博士課程の学生の研究指導もしている．

民博では民族学研究の情報センターとして，世界の数千の民族についてのデータを集積している．このデータは数値や文字情報のみならず，写真，スライド，フィルム，音楽，音声，物など，その情報形態は多岐にわたっている．この膨大なデータの検索，情報処理を有効に行なうための知的生産の技術として，いろいろな種類の情報処理装置を導入し，新しい研究方法の開発および博物館の情報管理のあり方を追及してきた．

図書，論文，標本資料，映像音響資料，HRAF(人間関係地域ファイル)，などを収集し，研究に活用できるようコンピュータ化を行なっている．現在約130万件の民博独自のデータベースを構築している．これは単に文字表記による書誌的事項のみならず，できるだけ現物に近い情報を提供できるようなマルチメディアのデータベースを目指している．

標本資料の平面，正面，側面，鳥瞰からの映像の蓄積検索システムではすでに3万5千点分の映像データが取れている．1点の標本資料から7メガバイトのデータが発生するので，200ギガバイト以上が光ディスクに蓄積されている．スライドイメージの蓄積・検索システムでは35 mmのスライドをハイビジョンなみの分解能でデジタル化して入力し，各スライドに付けられた自然語の単語によって検索し，表示，プリントができる．また電子ファイルを使って本の頁イメージの検索なども行なっている．そのほか標本画像自動計測装置や，各種画像・音響情報の処理，地図と文化要素との重ね合せ(マッピング)，シミュレーションなど知的生産技術の開発などを行なっている．

人々の生活や自然環境の実態，文明の交流などを多面的に知るためには，単に映画やビデオだけではなく，文章による詳細な記述，写真，スライド，音など種々の情報媒体を総合的に用いる必要がある．しかもそれら異なるメディアの情報が連動して検索されることにより，理解が深まっていく．百科事典の説明と共に，資料の写真，それを使っている場面の映像，あるいはそこに生じる生活音や，現場における会話などが一体となることによって，あたかもフィールドにいるような臨場感が得られる．

このように静止画，動画などの映像情報，百科事典，民族誌などの記述情報，音声，音楽などの音響情報，それらすべてを収納し，検索し，相互に関係づけて提供するための仕掛け，これが民博の目指す総合的データベースである．

民族誌データベースとしてはテキストの全文を入力し，そのインデックスの自動作成や，ある事柄に関する記述がなされている個所を任意の単語で検索できるシステムも現在開発中である．すでに数百冊分のテキストをパンチャーによる手作業で入力しているが，文字自動読取装置を導入しテキスト入力の自動化も試みている．

これらいろいろな形態のデータを蓄積処理するためのコンピュータシステムとしてIBM3090・4341・9375，富士通M340R，VAX11/780・PDP11/60，TOSFILE，池上-RAMTEK，J-STAR，YHP，SONY-TEKTRONIX，その他，日本電気，APPLE，IBMなどのパソコン，画像・音響の入出力装置など各種のコンピュータをLANで結合したシステムを構成している．

著者推定問題の数理統計学的研究

統計数理研究所　村　上　征　勝

文献の数量的な性質，たとえば文献における単語の出現率，単語の長さの分布，文の長さの分布，品詞の出現率，延べ語数に対する異なり語数の比率などを調べ，その統計分析に基づいて文献の真偽判定や著者の推定などを行なう研究は，欧米では一世紀に近い歴史があるが，日本語文献に関しては最近ようやく本格的研究が開始されたというのが現状である．

本報告では文学作品，宗教書，哲学書，新聞記事，書簡などの著者推定を試みたいくつかの代表的研究と，用いられた種々の検定法，判別分析法，クラスター分析法などの統計手法を紹介し，それらの手法の適用妥当性について言及した．また現在進めている日蓮遺文の真偽判定