機械処理を考慮した中国語最小音素のいくつかの統計性質

長春郵電学院　金　明哲
宇都宮大学・教養部　徳田尚之
統計数理研究所　村上征勝

中国語は世界で最多の使用人口をもつにも関わらず，機械処理を目的とした言語研究は欧米や日本などの先進国より遅れている．しかし近年，中国語機械処理の研究が盛んになり，ようやく本格的な研究が開始されたというのが現状である．今回の報告は筆者の「中国語の計量分析と機械処理に関する研究」の一環として既に発表した「機械処理のための中国語拼音の調査」，「中国語高頻度単語の品詞と近距離単語について」，「中国語高頻度単語の拼音対」の継続である．

本調査に使用した中国語高頻度単語は新聞，雑誌，学校のテキスト，校外読み物など879,300語を含むサンプルの90%をカバーするものを抽出した異なる高頻度単語であり，そのサンプルの大きさから中国語語彙の特徴をある程度反映しているものであると考えられる．これらの言語に関する統計的性質は中国語機械処理の基礎資料になると考える．

本報告では中国語高頻度単語の

(1) 最小音素の出現頻度
(2) 最小音素を単位とした単語長
(3) 最小音素を単位としたエントロピー，それの1次条件エントロピー，2次条件エントロピー，単語単位のエントロピー
(4) 最小音素を単位としたLevenshtein DistanceとHamming Distanceの近距離単語数，声調，品詞が近距離単語数に与える影響
(5) 最小音素における置換対

などについて行った統計分析の結果を述べた．今回の研究では今後中国語の音声機械処理を行ううえでいくつかの有益な結果が得られた．

今後の研究課題として中国語ローマ字列，音素列のマルコフ性の及ぶ範囲，遷位確率などの統計的性質を更に調査し続けると同時に，統計データを利用し人工知能の分野で使われ始めている確率推論法を使った訂正率の高い単語単位の誤り訂正方法について検討するつもりである．

文章作成アウトライン・システム

大阪府立大学 総合科学部　樺島忠夫

時枝誠記は，文章をそれ自体まとまりを形作っている一つの統一体と定義して，文と区別した．では「まとまり」を作るものは何か．それには，

(1) 意図のまとまり
(2) 意味のまとまり

の二つが考えられる．また，文章を特性づけるためには，

(3) 言語表現であること