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

(4) 線条性を持つまとまりであること

が必要になる.

この文章について，構造を捉えようとすると，

A 意図のあり方
B 意味内容
C 文脈の切れ続き
D 表記

の4点の総合として行うことになる.

この中で，AとCとを特に取り出して，次のような文章の構造を作る.

新製品の紹介

1. 生活や仕事の中での必要
   生活や仕事の中で，こんな必要がある，こんなことが実現できたらよいということを述べる.
2. 現在はどうなっているか
   2-1 現状の報告
   その必要を満たすために，今はどのような器具や機械があるかを説明する.
   2-2 それに見られる限界，不十分
   しかし，現在の器具や機械には，性能に不十分な点があったり，機能に限界がある，ということを述べる.
3. 新製品についての解説
   3-1 こんな製品が発明/製作/発売されたと紹介する.
   3-2 その製品の機能，効果を解説し，それが，役に立つものであることを解説する.
   3-3 新製品の構造，形態，発売予定，予価などについて解説する.

このような文章の構造の各項目に，その中で書くことが考えられる内容の概略，思いつきを書き込んだリストを，アウトラインということにする.

種々の目的の文章を書くために役立つ，よく出来た文章構造を数多く集め，文章構造のデータベースを作って，文章作成を支援するシステムを作ることができる.

関係構造分析法を適用した学術文献間の特徴抽出

北海道大学 工学部 斉 藤 たつき

研究論文を主体とする学術文献情報間の構造を解明し，その特徴を明確にする目的には，研究の促進，研究動向の把握，あるいは研究そのものの本質の探求等がある．本研究では，研究者ならびに学生のためのアドバイスシステムの開発の側面からその方法論を議論する.

研究や教育をする場合，ある程度成熟した分野であればサーベイ的論文が存在するためそれを手掛りにして研究・教育を進めたり，その分野の専門家に助言を求めたり，あるいはその専門家が学生に適切なアドバイスをすることができる．また，未成熟な分野あるいは新しい分野の場合は，その周辺の関連分野の状況を把握する必要があろう．本研究はこうしたいずれの目