氏名	金　　明　哲
学位（専攻分野）	博士（学術）
学位記番号	総研大甲第９６号
学位授与の日付	平成 ６ 年 ９ 月２０日
学位授与の要件	数物科学研究科　統計科学専攻 学位規則第４条第１項該当
学位論文題目	自然言語におけるパターンに関する計量的研究
論文審査委員	主　査　教　授　松　縄　　規 教　授　伊　藤　栄　明 教　授　柳　本　武　美 教　授　村　上　征　勝 教　授　樺　島　忠　夫（神戸学院大学） 教　授　安　本　美　典（産能大学）

論文内容の要旨

近年コンピュータを利用した自然言語に関する研究が多くの研究者の注目を集め，さまざまな角度から研究が進めれているが，自然言語に関する工学的な観点からの研究の多くは自然言語におけるパターンに関する研究である．例えば，文章認識（識別），単語の認識，文章の著者の推定，文章（文書）の自動分類などすべてパターン認識やパターン分類などの研究に帰結される．このような自然言語におけるパターンに関する研究は，分析の単位により，文字のパターン，単語のパターン，文のパターン，文章のパターンに大別することができる．本論文では，このうちの単語のパターンと文章のパターンの計量分析が詳細に研究されている．

論文は第Ⅰ部，第Ⅱ部の二つの部分（１０章）から成る．

第Ⅰ部は中国語の高頻度単語のパターン認識に関する計量的研究で，三つの章により構成されている．

まず第１章で研究の現状及び目的を示した後，第２章で中国語をキーボード入力した際の入力誤りを機械的に訂正する際に必要不可欠な中国語単語のローマ字表記パターンの統計的特性を明らかにし，第３章では音声認識の際に重要な役割を演じる単語の音韻パターンの統計的特性を明らかにしている．

第Ⅰ部での研究の背景と得られた主要な結果は以下の通りである．

中国語のピンイン表記の研究，音声・音韻学の研究，及びそのパターン認識などの機械処理に関する基礎研究として，中国語の統計的特性に関する情報はなくてはならぬものであるが，ピンイン表記や中国語の音声・音韻に関する計量的な研究はこれまでない．したがって，単語の長さの分布やローマ字及び音素を単位としたエントロピーなど中国語単語の最も基本的な性質さえも明かにされていなかった．そこで本研究では中国語高頻度単語について計量分析を行い，まず中国語単語（ローマ字，音素の表記列）に関し単語の長さの分布やエントロピーなど，これまで不明であったいくつかの基本的な統計的性質を明かにした．また，中国語単語のパターン認識を行なう上で有益となる多くの知見を得た．

例えば，ローマ字，声母・韻母，音素を単位とした場合，声調と品詞情報を用いても，距離１（声母・韻母の場合はハミング距離，ローマ字，音素を単位とした場合はレーベンシュタイン距離）の単語が，１音節（漢字）では１単語あたりそれぞれ約９語，２１語，１０語もあるため誤り訂正はかなり困難であるが，２音節（漢字）では声調，品詞情報が既知であると距離１の単語は１単語当たり，それぞれ約０．３５語，０．５６語，０．２６語である．したがって，２音節（漢字）以上の単語におけるローマ字，声母・韻母，音素を単位とした距離１の誤り訂正は，記号列の統計的性質や言語情報を有効に利用すると可能となることが示された．声調情報を用いない場合の類似パターンの単語数は，声調が既知である場合の約３～８倍（１音節単語では３～４倍，２音節では５～８倍）であり，品詞情報を用いない場合の類似パターンの単語数は品詞情報が既知である場合の約２～３倍となることなどを明かにした．

これらの結果は，中国語の知的なキーボード入力システムや中国語専用のキーボードの設計，中国語の音声認識の機械処理に関する研究だけではなく，中国語のピンイン表記及び音声・音韻学の研究，比較言語学の研究にも必要不可欠な基本的な情報となると考えら

れる.

第Ⅱ部では文章のパターン分類が扱われており，まず第４章で書き手の文章のパターンを計量的に把握し，それを用いて著者の推定や文章の分類などを行なう研究の現状と第Ⅱ部の研究の目的を示した後，第５章では文章の計量的情報の中で特に単語の長さの分布に関する情報を用いた場合の文章の分類法が，第６章では読点の付け方に関する情報を用いた場合の文章の分類法が示されている．第７章では文章の分類における単語の長さの分布と読点の付け方の情報の有効性を他の文章の計量的情報と比較しその有効性を実証し，第８章では，今後増加するであろうワープロの使用が文章のパターンに与える影響を分析し，第９章では文節間の係り受け距離と書き手の文章のパターンとの関係を分析している．

第Ⅱ部では，文章の著者の推定や文章のパターン分類などに関する研究の鍵であるとも言われている著者の文体の特徴情報に関し，以下の点を明かにしている．

1)これまで日本語の単語は，長さが割に短いため著者の特徴が出にくいと考えられ，著者の推定などの研究にはあまり用いられなかったが，品詞別に分けた場合には単語の長さの分布に著者の特徴が出やすいこと，特に動詞の単語の長さの分布に最も著者の特徴が現れやすく，しかも，漢語・和語の使用率や合成語・非合成語の使用率に著者の特徴が見られない場合でも著者の特徴が明確に現れる．

2）読点の使用法には，はっきりした規則がないため，読点の前の文字の分布，読点の前の品詞の分布，読点を打つ間隔に著者の特徴が現れるが，読点の前の文字に関する分布に著者の特徴が最も出やすく文学作品のみならず研究論文の文章の場合でも著者の特徴が明確に現れる．

3)動詞の長さの分布と読点の前の文字に関する分布について，それらの文体の特徴情報としての有効性を比較するため，従来よく用いられてきた文の長さの分布，品詞の使用率，漢字・仮名の使用率との比較を行った．その結果，著者別に文章を分類するのに最も有効な情報は読点の前の文字に関する分布であり，その次が動詞の単語の長さの分布で，いずれも従来よく用いられてきた文体の特徴情報より著者の特徴が明確に現れる．

また，これらの研究に関連してワードプロセッサ(ワープロ)と手書きの文章の比較研究も行なっている．近年多くの人がワープロを用いて文章を書くようになってきたが．ワープロが文章のパターンに与える影響についてはこれまで明かにされていなかった．そこで，同一人物がワープロと手書きで書いた日記文を用いて，ワープロが文章のパターンに与える影響についても計量分析を行ない，ワープロの方が難しい漢字を多く使用する傾向は見られるものの，漢字の使用率，文の長さ，文章の量などには顕著な差はみられず，ワープロが文章のパターンに与える影響はさほど大きくないことが明かになった．

最後に第１０章でこれまで述べたような本研究で得られた主要な結果がまとめられている．

論文の審査結果の要旨

審査委員会は金明哲君の論文について数物科学研究科における課程博士の授与に係る論文審査等の手続き等に関する規程に基づき、公開の論文発表会を開催し審査を行った結果以下の理由により合格と判断する。

本論文では、単語や文章のパターンに関して種々の統計分析が試みられ、研究の遅れている自然言語の数量的分析に関し数多くの知見を得ているが、それらを要約すると、本論文の貢献は次の２つになると考えられる。

第１の貢献は、まずこれまでほとんど解明されていなかった中国語のローマ字表記及び音韻情報に関し様々の統計的特性を明らかにし、中国語のワープロの入力ミスの機械的誤り訂正や音声による単語認識など、今後の中国語の機械処理に関する基本的情報を与えると同時に、中国語の数量的研究の端緒を拓いた点にある。

第２の貢献は文章を著者別に分類する際に、従来誰も注目していなかった読点の付け方に関する情報と、単語の長さに関する情報の有効性を実証した点にある。特に読点の付け方に書き手の特徴が現れるという分析結果は、国語学会の論文誌「国語学」177集(1994年６月)の中での「平成４年、５年における国語学界の展望」で国語学の専門家からも高く評価されている。また、単語の長さの分布、特に動詞の長さの分布に書き手の特徴が出やすいという結果も、本審査会の委員の国語学の研究者から高い評価を受けている。

加えてこれらの一連の分析を、著者自ら中国語の品詞、声調情報付きデータベース(6321語)、日本文の品詞情報付きのデータベース(104387語)を作成し、それを用いて行なったことも評価される。

なお、本論文の一部はすでに５編の論文として行動計量学、Behaviormetrika、統計数理、計量国語学(２編)に発表されている。