様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 １５ 日現在

機関番号：１３３０１
研究種目：挑戦的萌芽研究
研究期間：2013～2015
課題番号：２５５８０１０９

研究課題名（和文）日本語学習者データマイニングのための既存ツール評価とマニュアル設計

研究課題名（英文）Datamining Tool evaluation and the Manual Development for Japanese language teachers

研究代表者
松田　真希子（Matsuda, Makiko）

金沢大学・国際機構・准教授

研究者番号：１０３６１９３２

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：本研究では日本語教育のためのデータマイニング技術開発を目的に、[1]既存データマイニング技術の日本語教育研究への応用可能性の検討、[2]既存ツールのカスタマイズ、[3]ツール開発、[4]マニュアル開発を行った。その結果、[1]ではKh-Coder,SVtoolsなどを用いた日本語教育研究への応用研究を行い、有効性を示した。[2]では日本語学習者誤用換言対データを約3000対開発した。[3]では日本語学習者アクセントの自動評定技術開発を行った。[4]ではマニュアルを一部Web公開した。学術的成果としては、6件の学術論文の発表、11件の学会発表等を行った。

研究成果の概要（英文）：In this project, several datamining tools such as Kh-Coder and SV-tools were 1) applied in the field of Japanese language education, 2) customized, 3) developed to a new tool and 4) documented for using these tools. In 2), we got a dataset of about 3,000 pairs of mistake-correction phrases by Japanese learners. In 3), an automatic evaluation system for accentuation of Japanese read speech was developed. In 4), we published a part of the user's manual for Japanese language teachers on the web. In summary, six articles and eleven presentations were done for this project.

研究分野：コーパス言語学

キーワード：データマイニング　コーパス　統計　テキストマイニング　日本語教育　アンケート調査　アクセント評価

１．研究開始当初の背景

近年,コーパスに基づいた日本語・日本語教育研究 が盛んに行われている。BCCWJ 等の日本語母語話者コーパスと共に,C-JAS 等の日本語学習者コーパスも ここ数年でかなり整備された。今後はそれらの構築さ れたコーパスをどのような情報財に加工するかが重要 になる。

日本語教育では数年前からテキストマイニングツールの KH-Coder を用いた研究が見られるようになったが,それ以外のツールを用いた研究や,音声や評価法等,日本語教育への応用可能性について幅広く検討した研究は管見の限りない。

２．研究の目的

そこで本研究は，日本語教育向けデータマイニングツールの開発を将来構想に据えつつ,データマイニングの日本語教育への応用方法について幅広く検討すると共に,既存ツールのカスタマイズと活用マニュアルの設計を行う。

具体的には（１）過去 のデータマイニングの応用研究の調査と既存のマイニングツールの比較・研究（２）マイニング ツールのカスタマイズ（中間言語,音声対応） （３）データマイニング技術の日本語教育への新たな活用法の提案 （４）（３）を含むデータマイニング活用マニュアルの公開（５）日本語教育 研究向けデータマイニングツールの開発,である。

３．研究の方法

まず既存データマイニングツールについて調査するとともに,日本語教育研究への応用を試みた。その試みによって今後必要とされる技術を明らかにした。同時にすでに技術開発が望まれている音声データマイニングに関する基礎技術開発を行った。最後に日本語教師向けツールマニュアルについて検討を行った。

４．研究成果

（１） 既存ツールの日本語教育への応用とマニュアル化

まずフリーのテキストマイニングツール Kh-Coder によってどのような日本語教育研究が可能かについて検討を行った(業績 2,5,6 他多数)。その結果,形態素解析器に機能語を追加することで,かなりの程度非母語話者と母語話者の産出する日本語との比較分析ができることが明らかになった。その一方,日本語学習者の中間言語は形態素解析器で抽出できないため,中間言語を辞書登録することが必要であることが明らかになった。そのため学習者の誤用換言対コーパスを約 3000 対開発した（業績①)。

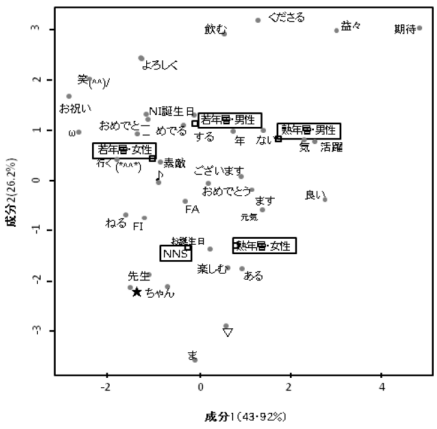

図１　Facebook ポストに関する対応分析結果（業績 2）

また Kh-Coder では文字列や単語 N-gram の情報が出せないため,研究協力者の山本和英氏の協力によって N-gram リスト抽出ツールを開発した(この研究成果は本研究の成果ではない http://snowman.jnlp.org/n-gram_tool)。本研究ではこのツールと Kh-Coder の情報を合わせて理工系テキストに出現する漢字語彙の分析を行った。また,他にも日本語教育に有用な分析ツールとして Antconk, Elan, Langtest, JS-Star2012, SV-Tools などがあり,いくつかについて日本語教育研究に応用した（業績１他)。当初は機械学習を用いた研究（業績⑧）や Deep Learning，LDR などに適用可能性の検討も行ったが，日本語教師が活用するには難度が高いという結論を得た。

マニュアル化には既に着手しているが,完成はしていない。これらのツールのさらなる有効な活用方法の検討及び日本語教師に向けたマニュアル完成が今後の課題である。

（２）音声データマイニングの基礎技術開発

本研究では音響特徴に基づいてアクセント型を自動判定する方法を開発し,その判定方法と人手評価との一致率について検討した（業績③,⑥)。分析データには「日本語学習者による日本語発話と,母語話者との対照データベース」（国立国語研究所）に収録された朗読音声を用いた。アクセント型の判定は日本語母語話者 1 名と中国語母語話者 1 名が行った。

自動判定技術については,まず音声認識システム Julius を用いて音素セグメンテーションを行い,母音と有声子音部分から F0 を抽出した。アクセント判定に用いる音響パラメータは,セグメント内の F0 平均値のほか,中央値と外挿値（セグメント内の F0 価から求めた回帰直線の終端値,F0-Target）を用いた。結果は以下のとおりである。

・自動判定と人判定の一致率は 70～80%である。

・中国語母語話者による朗読音声の方が,日本語母語話者によるものよりも,自動判

定と人判定の一致率が高い。
・音響パラメータで最も一致率が高いのは中央値だが,文節内のモーラ数やアクセント位置によって最適なパラメータは異なる。
以上から,本研究の提案方法は学習者音声の発音評価ツールとして一定の精度を有するといえる。
今後はこの技術を Web 等に実装し,学習者が自身の音声を読み込むと,アクセント精度の判定ができるように開発を行う予定である。

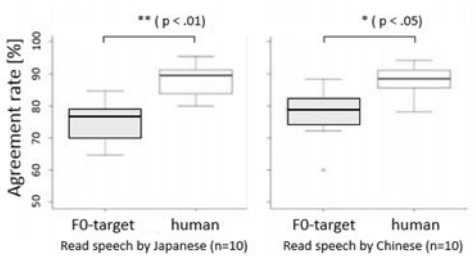

図２　アクセントの人手評価と自動評価の一致度（業績③）

５．主な発表論文等

〔雑誌論文〕（計 6 件）

(1) 呉麗楠, 波多野博顕, 金村久美, 磯村一弘, 松田真希子「JFL 環境下での発音学習ストラテジー使用と発音習得－中国の大学で学ぶ日本語学習者を対象に―」『音声研究』,査読有, 20(1), 2016, pp.6-15.
(2) 松田真希子, 小林ミナ「「私らしい」日本語の産出支援のための一考察―Facebook 誕生日メッセージ調査に基づく分析―」『統計数理研究所共同研究レポート 358 言語テキストと学習者特性の量的分析』, 査読なし, 2016,pp.42-56.
(3) 松田真希子, 嶋崎明美「日本の「見えない」文化の継承教育を考える ―メキシコ日系人・日本人アンケート調査に基づく分析―」『国際語としての日本語教育のための国際シンポジウム論文集』査読有, （本田弘之,松田真希子編『複言語・複文化時代の日本語教育』（凡人社,印刷中）
(4) 坂野永理, 渡部倫子「ラッシュモデルによるプレースメントテスト改訂版の検証『日本テスト学会誌』査読有,Vol.1, 2015, pp.99-110.
(5) 宮永愛子, 松田真希子「聞き手配慮要素からみた超級日本語話者の発話の特徴」『日本語／日本語教育研究』査読有, 5, 2014, pp.1-17
(6) 松田真希子「アメリカの大学生のための短期文化教育プログラム―ことばと文化をつなぐ教育の実践」In Proceeding of the 21st Princeton Japanese Pedagogy Forum, 査読なし, 2014, pp.242-258.

〔学会発表〕（計 11 件）

① 松田真希子「日本語学習者誤用換言対コーパスに見られる表記エラーについて」統計数理研究所言語系共同研究グループ合同発表会 言語研究と統計 2016, 統計数理研究所, 2016 年 3 月 15 日
②Matsuda Makiko, Yuka Ishikawa (2015) Basic Kanji characters and Kanji word extraction for the scientific and technological field using the Log-likelihood Ratio, In Proceeding of the Construction of digital resources for learning Japanese, UNIBO, Forli, Italy, Oct23, 2015
③Hiroaki Hatano , Carlos Toshinori Ishii & Makiko Matsuda Automatic evaluation of accentuation of Japanese read speech, In Proceeding of the Construction of digital resources for learning Japanese, UNIBO, Forli, Italy, Oct24, 2015
④松田真希子(2015)「Web 日本語 N グラムコーパス分析に基づく深層格の偏りの検証」計量国語学会第 59 回大会, 神戸大学, 2015 年 9 月 26 日
⑤松田真希子「パラ言語としての文字音声コミュニケーション」,研究集会「日本語音声コミュニケーション研究のこれまでとこれから」, 神戸大学, 2014 年 3 月 21 日
⑥波多野博顕, 石井カルロス寿憲, 松田真希子「日本語朗読音声を対象にしたアクセント型自動判定方法の検討」日本音響学会 2014 年秋季研究発表会, 北海学園大学, 2014 年 9 月 4 日
⑦松田真希子, 森篤嗣, 川村よし子, 庵功雄, 山本和英, 山口昌也「二格深層格の定量的分析」言語処理学会第 20 回年次大会, 北海道大学, 2014 年 3 月 19 日
⑧竹野峻輔, 松田真希子, 梶原智之, 山本和英 「機械学習を用いたニ格深層格の自動付与の検討」言語処理学会第 20 回年次大会, 北海道大学, 2014 年 3 月 19 日
⑨松田真希子「書き手と読み手の属性,関係性の抽出ーFacebook に誕生日メッセージを書き込むー」平成 26 年度日本語教育学会秋季大会, 富山国際会議場 2014 年 10 月 11 日
⑩林 良子, 張 亜明, 松田 真希子,金田 純平 緊張下における日本語学習者音声の特徴 日本音響学会 2014 年秋季研究発表会, 北海学園大学, 2014 年 9 月 5 日
⑪松田真希子,本田弘之「学校配布物における地域差と外国人支援 ―金沢・神戸の小学校一年生の配布物の比較分析―」日本語教育学会北陸地区研究集会, 福井大学 2014 年 6 月 21 日

〔図書〕（計 1 件）

(1)松田真希子『ベトナム語母語話者のための日本語教育』春風社, 2016, 299

〔産業財産権〕
○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://nihongokyo4.wiki.fc2.com/wiki/%E7%9B%AE%E6%AC%A1

６．研究組織

(1)研究代表者
松田　真希子（MATSUDA MAKIKO）
金沢大学・国際機構・准教授
研究者番号：10361932

(2)研究分担者
林　良子（HAYASHI RYOKO）
神戸大学・国際文化学研究科・教授
研究者番号：20347785

渡部 倫子（WATANABE TOMOKO）
広島大学・教育学研究科・准教授
研究者番号：30379870

金田 純平（KANEDA JUNPEI）
国立民族学大学・大学共同利用機関等の部局等・研究員
研究者番号：10511975