様 式 Ｃ－７－１

平成２５年度科学研究費助成事業（科学研究費補助金）実績報告書（研究実績報告書）

1. 機 関 番 号 | 1 | 4 | 6 | 0 | 3 |　　2. 研 究 機 関 名　奈良先端科学技術大学院大学

3. 研 究 種 目 名　特別研究員奨励費　　4. 研 究 期 間　平成 25 年度 ～ 平成 26 年度

5. 課 題 番 号 | 2 | 5 | ・ | 9 | 9 | 3 | 5 |

6. 研 究 課 題 名　統計的機械学習を用いた歴史的資料の校訂の自動化に関する研究と自動校訂ツールの開発

7. 研 究 代 表 者

研 究 者 番 号	研 究 代 表 者 名	所 属 部 局 名	職 名
	オカ 岡 テルアキ 照晃	情報科学研究科	特別研究員(DC2)

8. 研究分担者(所属研究機関名については、研究代表者の所属研究機関と異なる場合のみ記入すること。)

研 究 者 番 号	研 究 分 担 者 名	所属研究機関名・部局名	職 名

9. 研究実績の概要

下欄には、当該年度に実施した研究の成果について、その具体的内容、意義、重要性等を、交付申請書に記載した「研究の目的」、「研究実施計画」に照らし、600 字～800 字で、できるだけ分かりやすく記述すること。また、国立情報学研究所でﾃﾞｰﾀﾍﾞｰｽ化するため、図、ｸﾞﾗﾌ等は記載しないこと。

平安時代や明治時代といった古い時代の文献資料(歴史的資料)のコーパス化作業は，人手の校訂作業がコスト高であるため，現代語に比べて遅れている．そこで本研究では，統計的機械学習の手法を用い，コンピュータによる校訂作業の自動化を目的とする．校訂とは，コーパスユーザの可読性・検索性を向上させるために表記を整える作業であり，本研究では特に次の2つの項目を自動化の対象としている．

○表記の標準化:歴史的資料の中には，「及ひ(オヨビ)」のように濁音が期待されるのに濁点の付いていない文字や，歴史的仮名遣と一致しない仮名遣など，表記のバリエーションが多く含まれる．表記のバリエーションはコーパスを検索する際の障害となるため，コーパス整備時には表記を標準化する作業が必要となる．

○文境界判定:歴史的資料の記述中では句読点を含まず文境界が明確になっていないことが多い．文境界が明確になっていれば，テキストを文単位に解析できるといった利点がある．そのため，文境界判定が重要な作業となる．

本研究では，統計的機械学習を用いた日本語自動形態素解析の枠組みにおいて，表記の標準化と文境界判定を同時に実施することで，高精度な自動校訂の実現を目指す．本年度は，まず比較的リソースの多く確保できる近代文語論説文を対象に，形態素解析と表記の標準化の同時解析に取り組んだ．具体的には，辞書登録や辞書引きの工夫により，表記のバリエーションを含んだ単語も単語ラティスへと追加できるようにした．これにより，形態素解析の結果から標準化された表記を獲得することができる．年度前半には，この辞書引き手法の開発に取り組んだ．また年度後半にかけて，形態素解析との同時解析による表記の自動標準化ツールの開発に取り組んだ． このツールを用いることで，従来の単純な文字ベースの自動標準化に比べて高い精度で標準化を実施することが可能になった．

10. キーワード

(1) 校訂　(2) 歴史的資料　(3) 表記の標準化　(4) 形態素解析

(5) 歴史コーパス　(6)　(7)

（注）・印刷に当たっては、Ａ４判（縦長）・両面印刷し、左端を糊付けすること。

11．現在までの達成度

下欄には、交付申請書に記載した「研究の目的」の達成度について、以下の区分により自己点検による評価を行い、その理由を簡潔に記述すること。また、国立情報学研究所でﾃﾞｰﾀﾍﾞｰｽ化するため、図、ｸﾞﾗﾌ等は記載しないこと。

＜区分＞①当初の計画以上に進展している。 ②おおむね順調に進展している。 ③やや遅れている。 ④遅れている。

（区分）②

（理由）本年度予定していた研究はおおむね計画通りに進展した．形態素解析との同時解析による表記の標準化ツールの開発が順調に進み，従来法よりも高い精度で表記の標準化が実施できることが分かった．また研究の途中成果として得られた，表記のバリエーションを含んだ単語の辞書引き手法に関して，外部発表を行うこともできた．

12．今後の研究の推進方策

本研究課題の今後の推進方策について簡潔に記述すること。研究計画の変更あるいは研究を遂行する上での問題点があれば、その対応策なども記述すること。また、国立情報学研究所でﾃﾞｰﾀﾍﾞｰｽ化するため、図、ｸﾞﾗﾌ等は記載しないこと。

近代文語論説文（明治～昭和初期の文体）以前の文体（e.g., 中古和文）に対する提案手法の性能評価を行う．また，本年度は取り組まなかった文境界判定との同時解析にも取り組んでいく．

13．研究発表（平成２５年度の研究成果）

※ 「13．研究発表」欄及び「14．研究成果による産業財産権の出願・取得状況」欄において記入欄が不足する場合には、適宜記入欄を挿入し、それによりページ数が増加した場合は、左端を糊付けすること。

〔雑誌論文〕 計（ 1 ）件　うち査読付論文 計（ 1 ）件

著者名	論文標題			
岡照晃，小町守，小木曽智信，松本裕治	統計的機械学習を用いた歴史的資料への濁点付与の自動化			
雑誌名	査読の有無	巻	発行年	最初と最後の頁
情報処理学会論文誌	有	Vol.54 No.4	2013	1641～1654
掲載論文の DOI（デジタルオブジェクト識別子）				
なし				

著者名	論文標題			
雑誌名	査読の有無	巻	発行年	最初と最後の頁
掲載論文の DOI（デジタルオブジェクト識別子）				

著者名	論文標題			
雑誌名	査読の有無	巻	発行年	最初と最後の頁
掲載論文の DOI（デジタルオブジェクト識別子）				

（注）・印刷に当たっては、Ａ４判（縦長）・両面印刷し、左端を糊付けすること。

〔学会発表〕計（ 1 ）件　　うち招待講演　計（ 0 ）件

発表者名	発表標題		
岡照晃，小町守，小木曽智信，松本裕治	表記のバリエーションを考慮した近代日本語の形態素解析		
学会等名		発表年月日	発表場所
2013年度人工知能学会全国大会		2013年6月5日	富山国際会議場(富山県　富山市)

〔図　書〕　計（ 0 ）件

著者名	出版社		
書名		発行年	総ページ数

14．研究成果による産業財産権の出願・取得状況

〔出　願〕　　計（ 0 ）件

産業財産権の名称	発明者	権利者	産業財産権の種類、番号	出願年月日	国内・外国の別

〔取　得〕　　計（ 0 ）件

産業財産権の名称	発明者	権利者	産業財産権の種類、番号	取得年月日	国内・外国の別
				出願年月日	

15．備考

※　研究者又は所属研究機関が作成した研究内容又は研究成果に関するｗｅｂページがある場合は、ＵＲＬを記載すること。

（注）・印刷に当たっては、Ａ４判（縦長）・両面印刷し、左端を糊付けすること。