江戸版本のつづき文字部分に対する識別の試み

舟久保登†

豊橋創造大学　メディア・ネットワーク学科††

1　まえおき

講演者は平成１９年度までの３年間，「江戸版本の読解を支援する運筆特徴を考慮したつづき文字の認識に関する研究」なる科研費補助金による研究を，後の謝辞中に記す同僚３人と共同で実施した．その結果は文献[１]，[２]として発表しているが，本報告は講演者の分担範囲に関して，これらに引き続く内容を披露するものである．

2　対象つづき文字とその理想的識別

ここでの対象であるつづき文字を図１に示す．これは元とした江戸版本の見開き２頁の文章について，そこにある文字パターンを人手により個別に分離した場合，図２の標準文字パターン辞書に基づいて正しく識別された文字パターンだけから成るつづき文字部分である．なおこの際に用いた識別法は，もっとも単純で基本的な最短距離法を使った．

さてこのつづき文字に対して，その各文字をなるべく正しく識別するように，人間（講演者）が分割箇所（具体的には仮想的な文字枠の左上隅座標）を指定したときの識別状況を調べた．この結果はいわば，この場合における理想的な最高識別さを達成したものであると解釈される．表１にこれの整理した結果を挙げる．その識別率は93.3%（＝293/(342-28)×100）というところである．

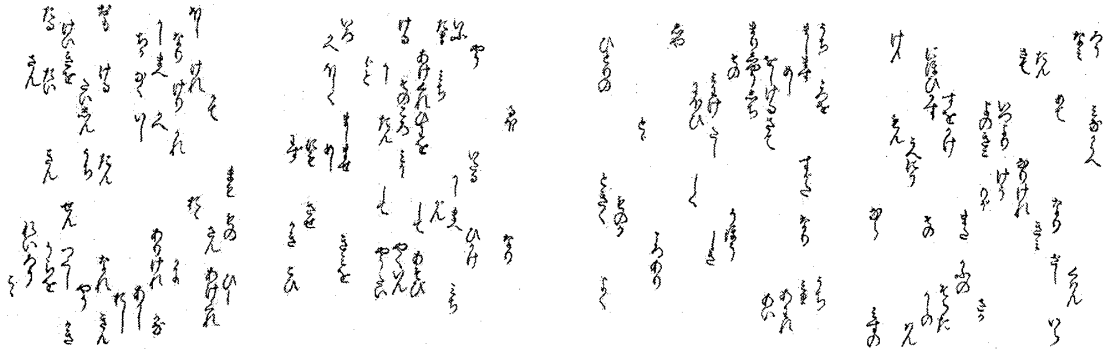

図１　対象つづき文字（左からグループ１左半部，右半部，グループ２左半部，右半部）

表1　人間により分割した場合の識別結果

対象文字総数	正識別文字数	辞書外文字数	誤識別	文字数
			先頭文字数	非先頭文字数
342個	293個	28個	13個	8個
			(「か」×8	(「か」×4
		([志」など)	「そ」×2	「こ」、「そ」、
			「ま」、「み」、	「ま」、「わ」)
			「め」)	

表2　つづき文字識別の試み結果

対象文字個数	正識別文字数	正識別率	誤識別文字内わけ
104個	56個	53.85%	「か」×5
			「い」、「う」、「き」、「ろ」×3
			「く」、「せ」、「ち」、「の」、「ほ」
			「ま」、「み」、「も」、「ら」、「り」
			「を」×2。その他9文字×1

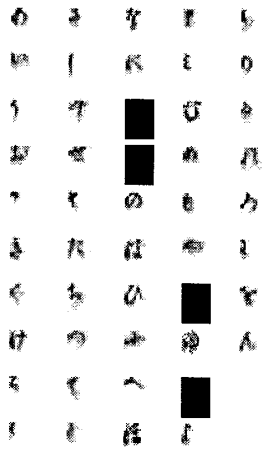

図２　標準文字パターン辞書

Trials on discrimination of continuous character parts on an engraved printing book in the Edo period

† Noboru FUNAKUBO

†† Toyohashi Sozo University, Department of Media Information and Networks

3　つづき文字に対する識別の試みとその結果

頁上にある元の文章からつづき文字部分の各個別文字パターンを識別するためには，大別して２段階の処理過程が必要である．その第１は図１に示したようなつづき文字部分を取り出すことで，この処理は文献[２]の４で検討した如く，通常良く行われる文字線についての縦，横方向濃度ヒストグラムの閾値化に基づいて割合容易に達成できる．そこで以下では第２の処理過程である抽出された１つのつづき文字部分に対し，これを構成している個別文字を識別する試みについて述べる．

ここでのその方法の骨子は，もっぱら標準文字パターン辞書（図２）に依拠するものである．まず１つのつづき文字部分の抽出について，以下の試みでは上記濃度ヒストグラムの使用に代えて，２において述べた人間による左上隅座標箇所の指定をもってした．この結果そこでの条件から，つづき文字部分の先頭文字パターンは必ず正しく識別されていることになる．次に２番目の文字の識別を行うために，あらかじめ標準パターンについて設定しておいた縦方向大きさ（高さ）に従って，識別の対象部分を下方に移動する．またこのとき同時に，標準パターンの横方向大きさ（幅）も設定してあるので，この値を頼りに横方向ずらしもする．その具体的な数値はここでは示していないが，図２から分るように標準文字パターン毎に横方向大きさは結構変化しているので，このような処理操作も必須である．そしてもし第３字以後も存在すれば，この識別のために同様な手続きが再帰的に適用される．

さてこのようにして試みたつづき文字に対する識別結果を，前頁の表２に示した．この場合の対象104 文字はほとんどについて，第１文字が正しく識別されたつづき文字部分中の第２文字となっているものである．その結果は正識別の文字数 56 個（率にして 53.85%）で，決して良くない．この原因はここでの場合，その文字分割箇所が各標準パターン毎に唯一に設定した縦方向大きさに依っており，しかし実際のつづき文字における第１文字はいろいろな縦方向大きさを持っているためである．

この辺の状況を調べる目的で，表１にある正識別文字の縦方向大きさがどの程度ばらついているかを，図３にグラフ化してみた．図における縦棒の範囲がこの高さ変化を表しており，また短い横棒は平均値である．そして比較のために，この図に標準パターンに対し設定した縦方向大きさも実線と点線で描いてある（今回の試みでは点線を採用）．これから分るように，文字パターンの高さは文字名を固定してもずいぶん変化し，またその範囲が標準パターンの設定値と重ならないこともある．これでは識別率が半分位になるのは当然予測され，この事態についてより積極的な対策処理が必要である．

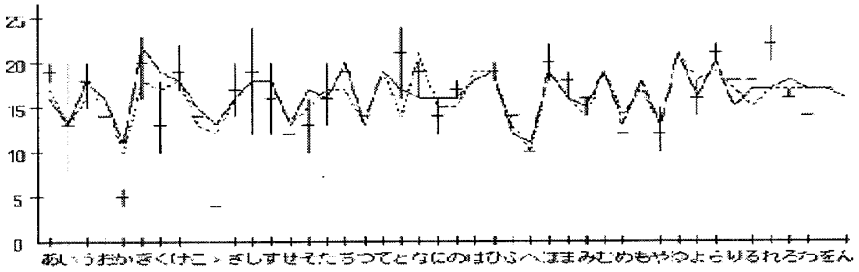

図３　文字の縦方向大きさ[画素数]（実・点線：標準，縦棒：人間による分割文字）

4　あとがき

標準的な伝統型の方法を用いつづき文字の識別を試みたが，予想以上に相当難しいことが分った．科研費研究全体の中での講演者の立場はその脇を詰めるものであるとしたので，この事実はそれなりの目的達成と考えている．なおここでの研究課題を真正面から扱った内容としては文献[３]と文献[１]の該当部分があるゆえ，念のためその事柄を最後に付け加え述べておく．

謝辞：本研究がその一環であった科研費研究を共に行い，その中でデータの提供，各種機会における議論などをして頂いた島田大助，三好哲也，三輪多恵子の３氏に対して，心から感謝を申し上げます．

参照文献

[1]江戸版本の読解を支援する運筆特徴を考慮したつづき文字の認識に関する研究，科学研究費（課題番号 17509165）報告書，平成 20 年 3 月

[2] 江戸版本におけるつづき文字部分の識別についての検討，舟久保登，豊橋創造大学紀要，平成 21 年 2 月，印刷中

[3] 運筆情報を利用した古文書におけるつづき文字認識，三輪多恵子他，電子情報通信学会ソサイエティ大会，平成 18 年 10 月