Bulletin of Toyohashi Sozo University
2011, No. 15, 25–36

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［第３報・完］

舟久保　　登

紀要における前回の2報までで，つづき文字部分中の第2文字パターン103個について文字名毎の標準文字パターンを1個当て用意した最短距離法により，81個を正しく識別可能（認識率78.6％）なことを得ていた．本報告では残った誤り文字パターン22個の認識に対して検討する．文字をつづけて書くことによる第2文字の縦方向位置ずれに対処するよう，最短距離法の識別関数を第1文字との分割位置が＋9画素から－9画素まで移動した文字パターンに適用し，その最大値から決まる識別文字名を求めた．この結果新たに17個の正解が得られ，認識が相当に良い95.2％を達成できた．さらに残り5個に関してその大部分の場合に‘と’の標準パターンで代理されたつづきパターンの出現を示し，これを考慮した認識過程を導入することによりこれら文字パターンも正解にする可能性を述べる．

Key Words：continuous characters parts discrimination, minimum distance method, characters segmentation, standard character pattern dictionary, almost complete recognition rate

Ⅰ　まえがき

この論文は一昨年と昨年の紀要に報告した内容［1］［2］の続きで，かつ今回で一応の終了とするものである．ここでの研究課題は最初，後の謝辞の中でその名前を記載した3人の同僚の教員と共に，科学研究費「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」［3］によって始められた．そしてその科研費期間自体は平成20年度末までであったが，この後も執筆者は研究を継続して結果を報告してきた次第である．

Ⅱ　前回までの内容の概要

（より一層の詳細は，一昨年と昨年の紀要を参照されるようお願いする．）

対象として取り上げたのは，江戸版本「女五経」の見開き2頁（図1）である［5］．そこには孤立した個別文字と共に，多数のつづき文字部分が存在している．

ところで先行した同僚の三好らの研究において，図1にある文字パターンを人手により切り出し，縦50画素×横50画素の正方形領域の左上隅にその位置を寄せたデータが用意されていた．そこでここでの文字認識に必須な標準文字パターンには，上記切り出されたデータ

よりの文字名毎のパターンの平均を求め採用した．この結果の標準文字パターン辞書を図2に示す．図の下方に挙げた名前の一覧から分るように，データの非常に少数な場合や重なっているとき（‘い’や‘え’など）のものは省いている．また‘ゝ’（同じ文字の繰り返し記号）を第10番目に加えてある．

一方研究の課題であるつづき文字部分の認識を明確に追究するために，その対象については次のように考えた．まずつづき文字部分としては，切り出された個別文字パターンの状態では正しく識別されるものから成るそれだけを採った．具体的に対象としたこのつづき文字部分は，図3のようである．そして実際に認識したい対象文字パターンは，これらつづき文字部分中の第2（番目）文字とした．何故ならこのつづき文字部分は個別には正識別できる文字パターンより構成されているゆえ，文字のつづいて書かれた効果は当然第2文字以後に，特にどのような短いつづき文字部分にも必ず存在する第2文字に現われている筈と考えられるからである．さらに以上の状況をより一層明確で純粋なものとするように，第1（番目）文字を正しく識別する文字枠左上横と縦座標位置，また直接の認識対象である第2文字のその左横座標位置についても，人間（筆者）がそれらの値を設定することとした．

図3のつづき文字部分における第2文字の個数は103文字あり，このような条件下で得られた一昨年の正しい認識文字数は56個で，その識別率は54.4％と決して良くない．この場合先行する第1文字は上述したように常に正しく認識されるようになっており，また第2文字の文字枠左横座標位置も正認識をもたらす値に設定してあるので，誤り原因はつづいて書かれている第1文字から第2文字を切り出すための分割位置決定に使う第1文字名に対応した標準文字パターンの縦方向大きさの不適当さにあるといえる．そこで前号である昨年の紀要において，いくつかの条件の下でこの縦方向大きさを変更（改変）した．その結果正認識文字数は81個に増え，識別率も78.6％と大分向上した．しかしながらそこには依然として，22個の認識誤り第2文字パターンが残っていたわけである．

Ⅲ　認識誤り第2文字の識別関数値変化

以前の紀要に述べたように，ここでの文字認識にはもっとも基本的で単純な最短距離識別法を使用している．すなわち具体的に記すと，まず対象とする文字パターンについて各標準文字パターンに対する識別関数値を次の式によって算出し，

識別関数：$g^{(\ell)}(\mathbf{x}) = 2(\mathbf{x}, \mathbf{x}^{(\ell)}) - \|\mathbf{x}^{(\ell)}\|^2$

$$= 2 \times \sum_{i=1}^{I} x_i x_i^{(\ell)} - \sum_{i=1}^{I} x_i^{(\ell)2}$$

ここで　$\mathbf{x}$は対象文字パターン（ベクトル）

$\mathbf{x}^{(\ell)}$は名前‘ℓ’の標準文字パターン（〃）

x_i，$x_i^{(\ell)}$は各々上記$\mathbf{x}$，$\mathbf{x}^{(\ell)}$のi番目の要素で，画素が真白で0，真黒で255の値をとる

またIはベクトルの要素（次元）数で，ここでの場合縦画素数35×横画素数

$25 = 875$

しかる後，二つの文字パターン間の類似さを表す識別関数値の最大なものを見出して，下のように入力した対象文字パターンの文字名を決める方法である．

$\mathrm{Max}_{\ell}\, g^{(\ell)}(\mathbf{x}) = g^{(\ell_0)}(\mathbf{x})$ ならば，

$\mathbf{x}$の名前は‘ℓ_0’である

さていま検討対象である認識誤り第2文字パターンは第1文字のそれにつづけて書かれたもので，その認識に際して縦方向切り出し（分割）位置が不適切だったゆえに，別の文字名に誤った識別結果をもたらしたわけである．そこでこの分割位置を，その位置が短かかった場合については順に9画素まで下方の長い方にずらし，また長過ぎたとき（「過分割位置」）は逆方向に同様に9画素短くして，各々の識別関数値を計算してみた．表1は前回出された認識誤り第2文字22個についてのその結果である．以下この表内容を参照しながら，それに検討を加えていく．（なお上において第1文字名の標準パターンから決まる分割位置が短かったか長過ぎたかの判断は，一先ず別途決められると想定している．*）

上に述べたように識別関数値は二つの文字パターン間の類似さを表し，したがってその最大値をもたらす標準パターンの名前が分割された第2文字の識別結果を与えるのであった．この事実に鑑み，分割位置を移動した10個（移動量0を含む）の第2文字パターンに対して，最大識別値による認識結果を求め記したのが表の「最大識別結果」の行にある文字名である．これを見るとその下方に「正解」と書いたもののように，この当然な方法によって誤り22個中の17個が正しい識別となっていて，対象としている図3の第2文字が全部で103文字であったので，その識別率は $\{103-(22-17)\}/103 = 95.2\%$ と以前の78.6%から上昇し，相当良い認識が達成できている．（ただし表1の最後にある「誤りつづき対象部」‘と›’についてはその「注釈」欄に記した如く，正しい識別をもたらすよう元来人間の設定した第2文字左横座標画素位置65の値が不適当と分ったので，1画素だけ左側にずらした64に変更している．よってこれは本来は前回の結果で正解に含めるべきものであった．）

Ⅳ　5個の認識誤り第2文字についての検討

前節でその結果を示した表1を参照すると，依然としてそこにはつづき文字部分中における5個の認識を誤る第2（番目）文字パターンが残っている．そこでこれらを正しく認識できないかという観点から，その識別特性に対する検討をすることにしよう．

表1でこの場合の「つづき分割位置」を変えた際における識別結果の文字名を観察すると，非常に‘と’の出力が多く，しかもそれが分割位置変化に対し連続して現われ，さらにその後すぐに続いて正解となるべき文字名の出現している事実が分る．ここでは文字パターンがつづけて書かれたことによって生ずる認識誤りを検討しているので，上記の事柄は第1と第2文字間のつづきパターンがこの識別結果をもたらしたのだろうと示唆される．そこで念のため文字名‘と’の標準文字パターンを，確認が容易なようにいくつかの倍率で拡大し図4に挙げてみた．これを見ると，図1の対象資料に出現した‘と’のパターンの平均で決めた

この文字名の標準パターンはずい分不明瞭な形状を呈している感がするが，またしたがって別に二つの文字間のつづきパターン（右上方から左下への斜線，上方から下への縦線など）の形に割合と似ていると眺められなくもない．

そこで以上の事柄を踏まえて，四つの認識誤り第2文字パターンについては，次の認識過程を提案できる（表1も参照のこと）．

‘なり’，‘みか’，‘そら’，‘さて’：識別関数値最大に基づいて決まった識別結果‘と’は全て無視して除く．そしてこの後に残ったものの中で最大識別関数値を与える文字名を，ここでの認識結果として出力する

表1から判明するように，上の処理操作は認識に対して正解を与える．ただしこの過程の中で‘と’を無視し得るかの根拠は，該個所でつづき文字部分の第2文字に‘と’が存在しないという事実が予め分っている必要がある．

もう一つの誤りとして‘うら’が残った．そこでこれについても表1を調べると，その誤り出力‘る’は「つづき分割位置」の最後の移動量9画素において起っていることが判明する．つまり移動量が大き過ぎたのである．したがってこの誤りを正すためには，

‘うら’：「つづき分割位置」を0～±9画素でなく，0～±8画素に狭くする

ぎりぎりのもの（‘かき’や‘さて’）もあるが，こう縮めても他の正解の状態は変わらず，‘うら’を正解にできるのである．すなわち一般的には，この報告で適当に設定した±9画素の値を，つづきパターンの縦方向大きさなどをより精細に考慮して適切に決定する必要があるわけである．*

以上いずれにしてもこの節で述べた方法によって，対象つづき第2文字103個全てを認識可能な結果が得られた次第である．

V　むすびと今後の課題

伝統的な縦書きされた日本語文章に生ずるつづき文字パターンの認識について研究してきた．対象として図1の江戸版本を取り上げ，そこに存在する図3に示したつづき文字部分の第2（番目）文字パターンを認識するよう試みた．方法には問題点の明確化を最重要視して，識別のため用意する標準文字パターンは文字名毎に1個だけを当て，また識別関数には基本的でかつ単純な最短距離法を使用した．この結果全部で103対象文字に対し98文字が正しく認識でき，その認識率は相当良い95.2%となった．さらに誤った残り5文字パターンについても，その誤り原因であるつづいて書かれたことによる影響内容を探究し，これに対する簡単な処置を行えばほとんど問題なく正解可能となる事柄を述べた．以上これらを通して，執筆者は本研究の目標とした課題について，非常にすっきりし一貫した最終的な解答を与え得たと満足している．

しかしながら上述のような性質を有する答えを求めるために，この研究における認識過程にいくつかの実用上からは問題とせねばならない事項を導入したことも否めない．その最大なものは，個別な状態では正しい識別のできる文字パターンのみから成るつづき文字部分を

対象としたゆえ，つづいて書くことの影響はもっぱら第2文字の縦方向位置に現われると単純化した事実である．これを踏まえここでは，つづき部分の第1文字パターンに対してそれが正認識されるようその文字枠の左横上画素座標位置と，また第2文字パターンついてもそうなることに必要な左横画素座標位置を人間が設定して与えた．けれどもこの事柄は実際の場合には当然期待できないわけで，この情報は典型的には科研費の研究報告の図3．9，図3．10に挙げたような文字行についての縦と横の射影像から取り出さねばならない．この点に関してそこでは文字パターン自体の識別に比して精度が得られていると記したが，その後の実験で識別結果が予期以上に文字パターンの存在位置に微妙に左右されるという感触を持ったので，さらなる検討が欠かせないと考えられる．

また本報告の図3に示したつづき文字部分に見られる如く，つづいて書かれた文字パターンには第3（番目）以後の文字パターンも存在している．もちろん基本的にはこれらの認識についてもこれまでの認識方法が同様に適用できようが，後のものになるほど誤差の積み重ねが増えると予想されるゆえ，問題に難しさはより生ずるであろう．

最後の課題として，ここで取り扱った文字パターン数が103個ではずいぶん少ないとする批判があろう．問題点を明確にする上では対象文字個数のこの少なさは役立ったけれども，実用的な観点からはこれも説得力を弱める．そこで次の段階では，文字数を1000個以上の規模に拡大した研究が必要であろう．そしてその際には一つの文字名に対する標準文字パターンの複数化，さらにこれに対応するためのパターンの特徴抽出化もなさねばならないかもしれない．加えて‘と’で代用されることがここでたまたま示されたつづきパターン自体に対する標準パターンを，積極的に用意することも必須になるのではないかと思う．

以上のような課題について，今後も取り組みたい所存である．

謝辞

本研究がその一環であった科学研究費の研究を一緒に実施され，その中でデータの提供，いろいろな機会における議論などをして頂いた，同僚の教員である島田大助教授，三好哲也教授，三輪多恵子准教授に，心から感謝を申し上げます．

【参考文献】

［1］舟久保登「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」『豊橋創造大学紀要』第13号（2009）pp.59–72.
［2］舟久保登「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［第2報］」『豊橋創造大学紀要』第14号（2010）pp.49–59.
［3］舟久保登，三好哲也，島田大助，三輪多恵子『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（17500165）報告書（2008）.
［4］舟久保登「江戸版本のつづき文字部分に対する識別の試み」情報処理学会第71回全国大会（2009）6C–2.
［5］霞亭文庫，東京大学附属図書館・情報基盤センター（http://133.11.199.8/cgibin/KatelIndex）.

＊［補足］

原稿提出と初校の間に，第1文字との分割位置を±9より縮め＋4～－4画素移動した際における認識状態を求めてみた．その結果は

① 標準文字パターンの縦方向大きさで分割した場合の正認識文字81個
→正解が74個，‘と’を無視での正解が2文字，誤りが5文字

② 標準文字パターンの縦方向大きさで分割した場合の誤り文字22（＝103－81）個
→正解が11個，‘と’を無視での正解が4文字，誤りが7文字

となり，認識率は｛103－(5＋7)｝／103＝88.4％であった．またその識別法の内容から予期されるように，分割位置画素移動範囲を狭めるほど①における誤りは少なくなり，②の誤り文字は増加してしまう傾向がある．

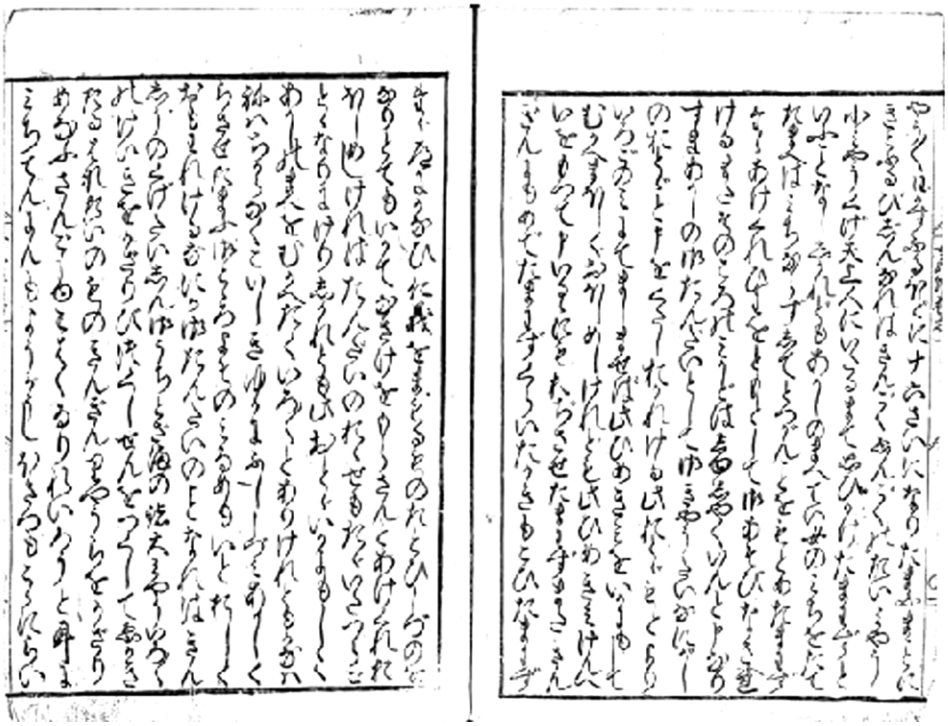

図1–a 「女五経」・グループ1

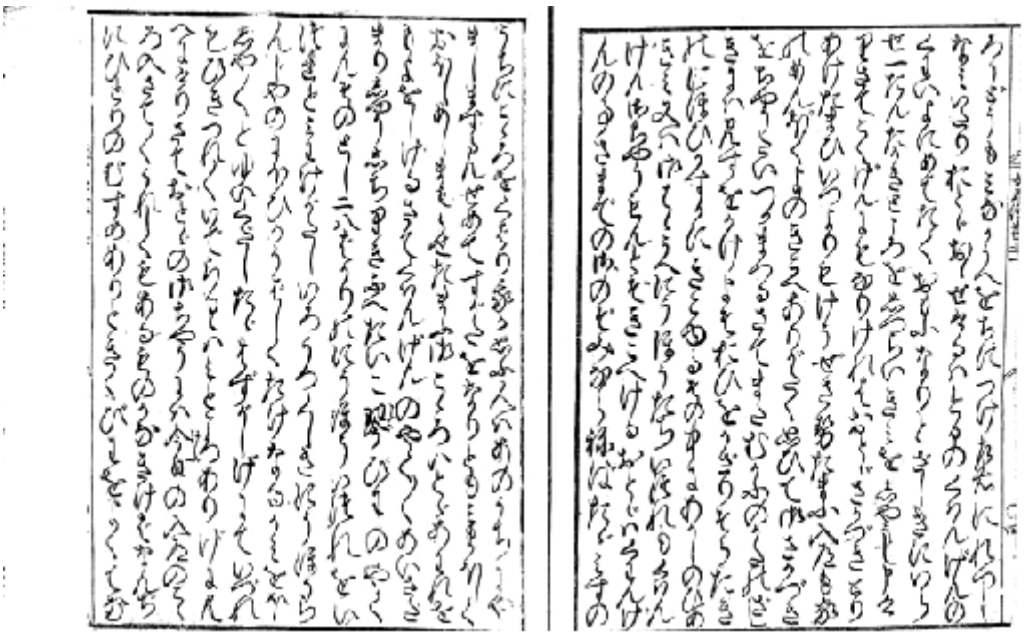

図1–b 「女五経」・グループ2

あ	さ	な	ま	ら
い	し	に	み	り
う	す	(ぬ)	む	る
お	せ	(ね)	め	れ
か	そ	の	も	ろ
き	た	は	や	わ
く	ち	ひ	(い)	を
け	つ	ふ	ゆ	ん
こ	て	へ	(え)	
ゝ	と	ほ	よ	

図2　標準文字パターン辞書

図3　具体的な対象とするつづき文字部分・グループ1（上半部）と2（下半部）

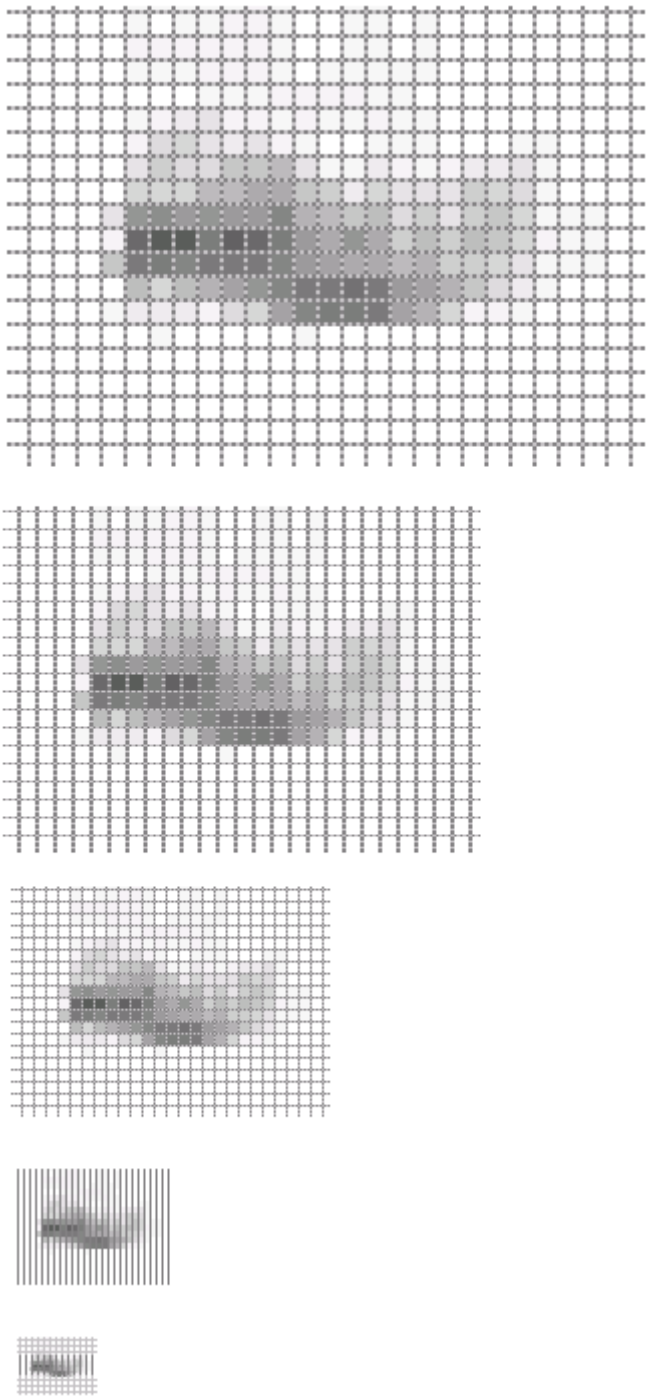

図4　いくつかの倍率で拡大した文字名‘と’の標準文字パターン（横向き）

表1　誤りつづき文字部分の認識内容データ

グループ1・右半部										存在場所	
い		み		こ		し		な		誤りつづき	
ろ		か		ろ		て		り		対象部	
211868	ん	254670	と	467188	ん	475034	く	812478	と	0 −9	つづ分割位置（画素）
471406	ろ	232086	か	577242	ん	472424	く	780936	と	1 −8	
822686	ろ	221422	か	656854	ろ	384163	て	699706	と	2ま−7	
249661	つ	106988	か	1047118	ろ	597305	て	665723	り	3 −6	
520347	つ	−47844	か	840126	ろ	543523	て	670753	り	4た−5	
785935	つ	−134550	か	401281	つ	264809	て	533821	り	5 −4	
706837	つ	−153782	か	484827	つ	257138	か	441127	こ	6は−3	
302395	つ	−166522	か	558213	つ	204114	か	220462	↘	7 −2	
−29669	へ	−192374	か	513583	つ	71368	か	88866	い	8 −1	
−191604	↘	−235990	か	245391	へ	−139162	か	11864	い	9 0	
ろ		と		ろ		て		と		最大識別結果	
ー		か		ー		ー		り		変更認識結果	
正解		'と'を無視で		正解		正解		'と'を無視で		注釈	
								横574→575			
								過分割位置			

グループ1・左半部					グループ1・右半部						
け		た		な		か		き		ま	
い		ん		り		き		み		せ	
43212	と	−390828	る	813822	と	484971	て	419071	し	998655	ふ
130688	い	−277938	る	788904	と	569536	く	359249	し	1104313	せ
286554	い	−214554	る	793762	と	623960	と	344515	し	1078395	せ
360622	い	6	ん	1014967	り	705118	と	383024	み	963941	つ
221070	い	318482	ん	1104009	り	785686	る	724116	み	823671	つ
58970	く	627266	ん	972599	り	788914	る	552140	み	644719	つ
18437	て	726402	ん	745581	り	932853	き	439404	↘	327375	つ
64759	て	624730	ん	491722	↘	1275479	き	366626	↘	28175	へ
47727	て	355542	ん	259498	↘	1295757	き	321765	し	−276423	つ
−80843	て	276822	く	114536	↘	974610	る	373957	し	−361673	つ
い		ん		り		き		み		せ	
ー		ー		ー		ー		ー		ー	
正解		正解		正解		正解		正解		正解	
				過分割位置							

グループ2・右半部						グループ1・左半部				存在場所	
そ		さ		き		う		た		誤りつづき	
ら		て		み		ら		い		対象部	
462210	と	1026811	さ	215507	し	357775	し	296110	く	0 −9	つづき分割位置（画素）
531522	と	1024901	て	237367	し	345233	し	200610	く	1 −8	
573910	と	1087317	て	229203	し	274387	し	275042	い	2ま−7	
536556	と	962323	て	259328	み	260433	し	320042	い	3 −6	
371044	と	1032167	て	368444	み	383921	ら	466546	い	4た−5	
465633	ら	1280575	て	295832	↘	494933	ら	606564	い	5 −4	
463469	ら	1352133	て	327768	↘	476986	る	488646	い	6は−3	
282377	ら	1007563	て	218976	↘	441336	る	115648	い	7 −2	
371285	う	666294	か	93982	↘	444560	る	110734	か	8 −1	
474763	う	716774	か	70596	↘	505786	る	87762	か	9 0	
と		て		み		る		い		最大識別結果	
ら		ー		ー		ら		ー		変更認識結果	
‘と’を無視で		正解		正解		つづき分割		正解		注釈	
		過分割位置				位置9を無視					

グループ2・左半部											
と		し		し		う		あ		さ	
↘		く		き		ほ		い		て	
252484	↘	163464	ん	373012	く	549900	ん	−93386	と	667574	と
72098	↘	181894	と	445879	て	689572	ん	−139946	と	748920	と
−22134	↘	296090	と	553263	て	863951	す	−140524	と	790504	と
−84172	↘	396226	く	694785	て	1050641	は	−112150	い	734296	と
−133508	↘	551276	く	825051	て	1347555	は	−24960	い	756016	と
−225010	↘	673806	く	909538	ん	2358603	ほ	98372	い	777864	と
−355002	か	596988	く	1200937	き	1738695	ほ	19404	い	738206	と
−415898	か	596345	こ	1508517	き	1642587	け	−152207	つ	669277	て
−600412	か	472923	こ	1426515	き	1608417	け	−164770	か	689009	て
−726828	↘	212054	↘	985974	ん	1602881	に	−178376	か	545996	か
↘		く		き		ほ		い		と	
ー		ー		ー		ー		ー		て	
正解		正解		正解		正解		正解		‘と’を無視で	
横65→64										過分割位置	