近代語文献を電子化するための文字セット

近代語文献を電子化するための文字セット

高田 智和　（国立国語研究所 理論・構造研究系）[1]

1．はじめに

本稿では、近代語文献を電子テキスト化する際に準拠する文字セットについて検討する。

一般的に、紙媒体の文書を電子テキストに写し取る際には、標準化された符号化文字集合に準拠することが多い。かつて、JIS X 0208 文字セットに拠って、『大正新脩大蔵経テキストデータベース』（http://21dzk.l.u-tokyo.ac.jp/SAT/）の前身、『青空文庫』（http://www.aozora.gr.jp/）、『太陽コーパス』など、学術的価値を有する電子テキスト群が作成されたが、その都度、JIS X 0208 文字セットでは表現できない文字（外字）への対処が問題となっていた（安永（1998）、下田・師（1999）、富田（2000）、池田・白井・高田（2002）、高田（2002）、田中（2005）、當山（2009）、須永・堤・高田（2011）など）。

2000 年に制定された JIS X 0213 は、外字の問題を解消し、「現代日本語を符号化するために十分な文字集合」を目指して開発された規格である。JIS X 0208 が、第 1 水準漢字、第 2 水準漢字、非漢字（X0208 非漢字と呼ぶ）の約 6,800 字であるのに対して、JIS X 0213 は、JIS X 0208 文字セットを拡張し、第 3 水準漢字、第 4 水準漢字、非漢字（X0213 非漢字）を加え、約 11,000 字の文字セットとなっている。

JIS X 0213 文字セットに準拠して作られた『現代日本語書き言葉均衡コーパス』では、のべ 99.96%の文字を符号化できることが確認されている（高田ほか（2009））。『現代日本語書き言葉均衡コーパス』での運用実験により、JIS X 0213 は「現代日本語を符号化するために十分な文字集合」であることが実証されたと言えよう。

しかし、現代から時代を遡って、歴史的な日本語文献に対して、JIS X 0213 文字セットがどの程度の有効性を持ち得るのかは、まだ試みられていない。以下に、『太陽コーパス』の JIS X 0208 外字を対象として、JIS X 0213 による再符号化結果を報告し、近代語文献の電子テキスト化における JIS X 0213 文字セットの有効性について見通しを述べる。

2．『太陽コーパス』の文字処理

『太陽コーパス』は 2005 年に公表された近代語研究向けのコーパスである。1895(明治 28)年から 1928(昭和 3)年まで発行された総合雑誌『太陽』のうち、1895 年、1901 年、1909 年、1917 年、1925 年の各 12 冊、全部で 60 冊分の記事を電子テキスト化し、引用、文、振り仮名などのタグを付けた電子化コーパスである。

『太陽コーパス』の文字処理では、JIS X 0208 文字セット（第 1 水準漢字・第 2 水準漢

[1] ttakada@ninjal.ac.jp

字・非漢字、約 6,800 字）に準拠して、『太陽』の記事の入力処理が行われている。『太陽コーパス』の開発は、JIS X 0213 が制定される以前から着手されており、開発当時の処理水準を考慮すれば、妥当な選択であったと思われる。

雑誌『太陽』では主に旧字体が使われているが、「羽」のように新字体「羽」との字体差がわずかであるものは、新字体で表現することを許容している。字体粒度の許容範囲は、原則として JIS 規格の包摂規準にしたがっている。また、JIS 規格の包摂規準を拡張させて、独自の包摂も行っている（田中（2005））。語彙研究や文法研究での活用が主に期待されているコーパスであるから、字体の忠実な再現は行っていない。

『太陽コーパス』における JIS X 0208 文字セットの使用状況をまとめたものが表 1 である。表の各セルで、上段がのべ字数、下段が異なり字数である。『現代日本語書き言葉均衡コーパス』では、第 2 水準漢字は異なり 1,311 字の使用にとどまっていることから、近代日本語書き言葉では多種多様の漢字を用いていることが裏付けられる。

表 1：『太陽コーパス』の JIS X 0208 による符号化の内訳

水準	1895	1901	1909	1917	1925	計
第 1 水準漢字	1,395,861 [2,687]	1,285,938 [2653]	1,107,698 [2,648]	1,012,692 [2,623]	848,430 [2,633]	5,650,619 [2,721]
第 2 水準漢字	221,940 [2,558]	200,767 [2,275]	172,761 [2,057]	168,536 [1,922]	124,880 [1,757]	888,884 [2,864]
非漢字	1,688,600 [291]	1,651,017 [279]	1,556,071 [278]	1,454,064 [268]	1,447,583 [287]	7,797,335 [318]
計	3,306,401 [5,536]	3,137,722 [5,207]	2,836,530 [4,983]	2,635,292 [4,813]	2,420,893 [4,677]	14,336,838 [5,903]

『太陽コーパス』では、JIS X 0208 文字セットで表現できない文字を、外字、踊字、合字、小書の 4 種のタグを用いて表現している。JIS X 0213 は JIS X 0208 を拡張したものであるから、外字、踊字、合字、小書の各タグで表現された文字類を、JIS X 0213 の拡張領域（第 3 水準漢字・第 4 水準漢字・X0213 非漢字、約 4,000 字）と突き合わせることで、『太陽コーパス』の JIS X 0213 文字セットによるカバー率をおおよそ求めることができるであろう。

JIS X 0213 の特長の一つに、JIS X 0208 では符号位置を区別せず包摂されていた字体を、符号位置を区別して表現できる点があげられる。例えば、「徳」の旧字体「德」や、「鴎」の康熙字典体「鷗」などの異体漢字が該当する。厳密さを追求すれば、JIS X 0208 では包摂するとされていた異体漢字や、『太陽コーパス』で独自に包摂した異体漢字についても、雑誌『太陽』で実際に使われている字体を確認し、JIS X 0213 の拡張領域の符号位置で表現できるかどうかを検討しなくては、JIS X 0213 文字セットによる真のカバー率を求める

ことはできない。しかし、今回の調査目的は、近代語文献の電子テキスト化における JIS X 0213 文字セットの有効性について見通しを得ることであるから、『太陽コーパス』の開発において、JIS X 0208 文字セットで表現できないとされた文字だけを調査対象としても、調査目的が達せられるものと判断した。

さて、外字、踊字、合字、小書、各タグの使用状況は表 2 のとおりである。次節以降、各タグについて、JIS X 0213 による再符号化結果を示す。

表 2：『太陽コーパス』の JIS X 0208 外字の内訳

タグ名	のべタグ数
外字	5,507
踊字	18,019
合字	8,554
小書	187
計	32,267

3. 『太陽コーパス』の JIS X 0213 による再符号化

3.1 外字タグ

『太陽コーパス』の外字タグの運用方法は 2 種類ある。一つは代用で、もう一つは本当の外字である。代用は JIS 規格の包摂規準では処理できないが、JIS X 0208 に採録された文字の異体漢字である場合に用いる。外字は JIS X 0208 の文字と異体関係が認められず、全くの別字の場合に用いる。それぞれタグの使用例を示す。なお、タグの属性「文字番号」は文字鏡番号である。

（例 1）【代用】 **凞**

<外字 文字番号="001721">凞</外字>々たる明治二十八年の新旭光は〔t189501〕

（例 2）【外字】 **磅**

内には浩然たる正氣の磅<外字 文字番号="024597">〓</外字>するところ禁ぜんと欲して能はざるあり。〔t189501〕

外字タグの再処理結果をまとめたものが表 3 である。JIS X 0208 を用いたときに外字タ

グで表現したものの 77.7%が、JIS X 0213 を用いると符号化することができる。

表 3：外字タグの再処理結果

	のべ字数	異なり字数
第 2 水準漢字	2	2
第 3 水準漢字	2,685	446
第 4 水準漢字	1,371	426
X0213 非漢字	220	38
（小計）	4,278	912
X0213 外字	1,229	579
計	5,507	1,491

第 2 水準漢字で表現できる文字は、次の 2 字（表 4 参照）である。これらはコーパス開発時の単純なバグであろう。

表 4：第 2 水準漢字で表現できるもの

文字番号	字形	面区点	度数
037880	蹴	1-77-12	1
047275	鷺	1-83-25	1

第 3 水準漢字で表現できる文字は、次の 446 字（表 5 参照）である。『太陽コーパス』での使用度数順に示す。

表 5：第 3 水準漢字で表現できるもの

文字番号	字形	面区点	度数
048824	麁	1-94-86	242
016085	欸	1-86-31	239
042124	雞	1-93-66	138
001721	濕	1-14-55	90
051106	开	1-84-17	71
056009	厲	1-14-84	64
035458	詹	1-92-08	54
050005	咤	1-14-88	37
023439	睜	1-88-85	34
032700	虛	1-91-45	29
053267	噶	1-15-20	25
040346	鉸	1-93-13	24
001244	儞	1-14-45	20
019395	燄	1-87-64	20
041315	閩	1-93-49	19
004276	嘻	1-15-18	17
019890	牖	1-87-69	17
016408	殂	1-86-38	16
024597	礪	1-89-18	16
053201	墓	1-84-88	16
011865	拘	1-84-72	15
012081	拐	1-84-77	15
011617	戝	1-84-66	14

017165	汴	1-86-52	14
027733	縈	1-90-16	14
038785	迤	1-92-52	14
038791	迨	1-92-53	14
000076	丰	1-14-06	13
005113	埈	1-15-47	13
010174	徧	1-84-34	13
015155	楣	1-85-86	13
040272	鈹	1-93-07	13
045469	鬓	1-94-27	13
000774	倘	1-14-30	12
009808	弹	1-84-22	12
018251	潢	1-87-13	12
050268	燮	1-87-67	12
008295	凿	1-47-85	11
010149	徜	1-84-33	11
015163	楨	1-85-88	11
000597	侔	1-14-22	10
001115	傲	1-14-40	10
009610	弈	1-84-19	10
010094	徉	1-84-32	10
027069	糕	1-89-86	10
028039	纍	1-90-24	10
037868	躎	1-92-39	10
045430	鬃	1-94-26	10
059830	笳	1-89-60	10
000471	你	1-14-13	9
002930	匡	1-14-82	9
014552	枘	1-85-54	9
015843	櫧	1-86-25	9
016752	毗	1-86-44	9
019012	烘	1-87-42	9
024232	碘	1-89-06	9
025635	窻	1-89-54	9
028810	翱	1-90-35	9
035370	詎	1-92-04	9
039198	邈	1-92-58	9

003528	咀	1-14-94	8
004394	噯	1-15-23	8
012181	捥	1-84-80	8
014005	晷	1-85-32	8
017526	涇	1-86-75	8
020916	珉	1-87-89	8
021253	璣	1-88-28	8
022383	癙	1-88-53	8
027247	納	1-89-90	8
034513	褰	1-91-84	8
038930	逭	1-92-56	8
003523	咖	1-14-93	7
004407	嗆	1-15-25	7
010496	怳	1-84-45	7
010771	惋	1-84-51	7
010815	惝	1-84-54	7
010949	愜	1-84-56	7
011833	扯	1-84-71	7
012105	挹	1-84-78	7
021062	琦	1-88-06	7
021270	璨	1-88-31	7
026734	籙	1-89-79	7
028952	耦	1-90-38	7
031565	蒞	1-91-13	7
032805	虬	1-91-50	7
040223	鈐	1-93-05	7
059130	溮	1-87-58	7
000418	份	1-14-09	6
004631	囉	1-15-31	6
006297	娓	1-15-81	6
007559	厊	1-47-62	6
010679	悞	1-84-50	6
018164	漪	1-87-06	6
018965	灶	1-87-40	6
024580	礹	1-89-16	6
025458	窅	1-89-50	6
026077	笄	1-89-66	6

027750	縑	1-90-17	6
043909	飂	1-94-07	6
043965	飋	1-94-08	6
001375	兗	1-14-50	5
003341	吧	1-14-86	5
003770	唵	1-15-06	5
003898	喁	1-15-09	5
007047	孽	1-47-55	5
008028	峋	1-47-74	5
010803	惕	1-84-53	5
011015	愷	1-84-59	5
011961	拖	1-84-74	5
012808	擎	1-84-92	5
015076	椵	1-85-83	5
015272	榭	1-85-92	5
017132	汙	1-86-49	5
017186	沅	1-86-54	5
017699	淼	1-86-86	5
018010	滇	1-87-01	5
018948	炫	1-87-39	5
019137	煞	1-87-48	5
021065	琨	1-88-07	5
023689	瞪	1-88-91	5
024175	硎	1-89-02	5
025556	窠	1-89-51	5
026844	敉	1-89-81	5
029700	腭	1-90-51	5
033077	蛺	1-91-54	5
036819	賡	1-92-25	5
039974	醞	1-92-88	5
040280	鉀	1-93-10	5
041383	閻	1-93-52	5
041446	閾	1-93-54	5
042940	鞮	1-93-79	5
042997	鞹	1-93-80	5
043716	顙	1-94-06	5
046597	鱷	1-94-55	5

056060	煒	1-87-54	5
000572	佾	1-14-20	4
000601	侗	1-14-23	4
004403	嘍	1-15-24	4
005179	埭	1-15-50	4
005558	墝	1-15-67	4
006002	奭	1-15-74	4
008106	峴	1-47-77	4
008488	嶠	1-47-89	4
009398	庾	1-84-13	4
012778	撾	1-84-93	4
014224	曛	1-85-42	4
015679	櫼	1-86-22	4
017319	泫	1-86-62	4
017609	涿	1-86-80	4
017639	淖	1-86-82	4
018328	澥	1-87-16	4
018833	灤	1-87-35	4
019883	牏	1-87-68	4
021865	畯	1-88-42	4
023307	眴	1-88-80	4
023310	眶	1-88-81	4
024110	砭	1-88-93	4
026313	篙	1-89-70	4
027246	紇	1-89-89	4
027371	絇	1-90-01	4
028727	翟	1-90-32	4
037452	跎	1-92-33	4
039301	邢	1-92-63	4
040365	銈	1-93-14	4
040857	鏽	1-93-39	4
041740	隄	1-93-60	4
047012	鵑	1-94-62	4
056422	鍥	1-93-25	4
056428	鏤	1-93-41	4
057094	氳	1-86-48	4
000420	仿	1-14-10	3

001895	刓	1-14-60	3
001999	剉	1-14-62	3
003422	呿	1-14-87	3
004797	圄	1-15-33	3
004889	圯	1-15-36	3
005167	埤	1-15-49	3
008548	嶸	1-47-92	3
008622	巂	1-47-93	3
009844	彀	1-84-25	3
010354	忡	1-84-40	3
011201	憍	1-84-61	3
012624	摭	1-84-91	3
013862	昱	1-85-21	3
013939	晙	1-85-27	3
013991	晢	1-85-31	3
015094	楂	1-85-84	3
017168	汶	1-86-53	3
017323	泮	1-86-63	3
017408	洮	1-86-67	3
017809	湄	1-86-89	3
018026	滎	1-87-02	3
018174	漳	1-87-08	3
018319	澈	1-87-15	3
018814	灞	1-87-33	3
020845	玕	1-87-83	3
021324	璨	1-88-35	3
021717	甯	1-88-41	3
022979	盌	1-88-72	3
026032	筠	1-89-63	3
026592	簇	1-89-75	3
027087	糙	1-89-87	3
027757	緼	1-90-18	3
028663	翎	1-90-30	3
029614	腊	1-90-47	3
030606	艷	1-90-60	3
034220	袪	1-91-73	3
034285	裊	1-91-74	3
034499	褧	1-91-83	3
037646	踠	1-92-36	3
038412	輗	1-92-46	3
038542	轔	1-92-48	3
038700	辵	1-92-51	3
040475	鋠	1-93-19	3
040640	鍪	1-93-30	3
042575	靚	1-93-75	3
042728	靳	1-93-77	3
043047	韁	1-93-81	3
045985	魦	1-94-34	3
046105	魬	1-94-40	3
047714	鱻	1-94-76	3
048480	鼳	1-94-84	3
048886	鯀	1-94-89	3
053113	儆	1-14-42	3
056019	壏	1-15-65	3
000497	佈	1-14-14	2
002497	勻	1-14-75	2
003590	哆	1-15-02	2
003908	喆	1-15-10	2
004175	嘈	1-15-16	2
004299	噉	1-15-19	2
004562	嚪	1-15-29	2
005190	埵	1-15-51	2
005592	壟	1-15-69	2
006206	姝	1-15-80	2
006469	媮	1-15-86	2
006533	媧	1-15-89	2
006941	孖	1-47-54	2
008846	帗	1-84-09	2
008851	帝	1-84-10	2
009058	帾	1-84-11	2
009079	幙	1-84-12	2
010529	恇	1-84-47	2
011940	拄	1-84-73	2
012125	掅	1-84-79	2

012494	搢	1-84-87	2
012912	擻	1-85-05	2
013040	攇	1-85-07	2
013796	昉	1-85-13	2
013903	晌	1-85-25	2
013952	晡	1-85-29	2
016275	歧	1-86-36	2
017369	泊	1-86-66	2
017412	洱	1-86-68	2
018416	澶	1-87-21	2
018939	炤	1-87-38	2
021018	琊	1-88-02	2
021067	琪	1-88-08	2
021102	琦	1-88-16	2
021361	瓚	1-88-37	2
022395	痤	1-88-54	2
022529	癌	1-88-58	2
022601	癎	1-88-59	2
022630	癭	1-88-61	2
024586	磐	1-89-17	2
027276	紓	1-89-91	2
027485	紲	1-90-06	2
027856	絲	1-90-20	2
032418	蘄	1-91-38	2
033079	蜳	1-91-55	2
033372	螈	1-91-60	2
033578	蠥	1-91-65	2
034523	褲	1-91-85	2
034544	褖	1-91-87	2
035069	觔	1-91-91	2
039658	鄱	1-92-82	2
040173	鈁	1-92-94	2
040446	鋋	1-93-16	2
040808	鏽	1-93-36	2
041052	鑩	1-93-44	2
041451	閬	1-93-55	2
044912	騽	1-94-16	2

045657	鬮	1-94-31	2
046171	鯁	1-94-42	2
046803	鶚	1-94-57	2
047034	鷂	1-94-63	2
050918	丰	1-14-05	2
053621	邝	1-92-61	2
056156	茁	1-90-76	2
056180	荇	1-90-82	2
056254	蒇	1-91-11	2
056374	藿	1-91-37	2
056390	蘸	1-91-44	2
057018	櫻	1-86-19	2
058024	檠	1-86-03	2
000382	仡	1-14-08	1
000499	佔	1-14-15	1
000789	倥	1-14-31	1
001846	刁	1-14-58	1
001899	刢	1-14-61	1
002781	卡	1-14-79	1
003302	吒	1-14-85	1
003446	呫	1-14-89	1
003540	咩	1-15-01	1
003755	唅	1-15-05	1
005419	墉	1-15-60	1
005470	墩	1-15-63	1
005747	夔	1-15-72	1
006329	娣	1-15-82	1
007098	宓	1-47-56	1
007257	寘	1-47-57	1
008010	岑	1-47-73	1
008209	崧	1-47-81	1
008502	嶧	1-47-91	1
009972	彤	1-84-29	1
009983	彧	1-84-30	1
010498	怵	1-84-46	1
012041	抃	1-84-76	1
012596	揔	1-84-90	1

012787	擄	1-85-01	1
012900	擷	1-85-04	1
013836	昺	1-85-16	1
013852	昀	1-85-19	1
013860	昱	1-85-20	1
014089	暠	1-85-36	1
014451	杈	1-85-50	1
014805	柠	1-85-67	1
015502	槼	1-86-11	1
015556	槪	1-86-15	1
016061	欻	1-86-30	1
016139	歂	1-86-32	1
016750	毖	1-86-43	1
017027	氐	1-86-47	1
017136	汜	1-86-50	1
017289	汿	1-86-60	1
017353	洄	1-86-65	1
017421	洹	1-86-69	1
017578	淬	1-86-79	1
017646	淛	1-86-84	1
017978	湊	1-86-93	1
018139	溫	1-87-04	1
018169	潔	1-87-07	1
018338	澔	1-87-18	1
018811	瀕	1-87-32	1
019174	燀	1-87-51	1
019291	熇	1-87-60	1
019304	燓	1-87-61	1
019519	燾	1-87-65	1
019521	爣	1-87-66	1
020431	狻	1-87-75	1
020643	獐	1-87-80	1
020861	玲	1-87-85	1
020869	玦	1-87-87	1
020874	玫	1-87-88	1
020962	珣	1-87-93	1
021015	琇	1-88-01	1

021071	琮	1-88-11	1
021073	琰	1-88-13	1
021122	瑗	1-88-18	1
021242	璜	1-88-26	1
022152	痎	1-88-46	1
022297	痙	1-88-49	1
022317	瘷	1-88-50	1
022634	癯	1-88-62	1
023167	盼	1-88-77	1
023392	睨	1-88-83	1
023523	睺	1-88-88	1
023541	瞀	1-88-89	1
024147	硃	1-89-01	1
024342	碭	1-89-10	1
024409	磯	1-89-11	1
024545	礨	1-89-15	1
025068	稃	1-89-44	1
025601	竊	1-89-53	1
025998	筎	1-89-62	1
026071	筋	1-89-65	1
026889	粔	1-89-83	1
027101	糝	1-89-88	1
027345	紱	1-89-94	1
027555	綮	1-90-09	1
028454	羌	1-90-28	1
029670	腠	1-90-48	1
029929	膻	1-90-52	1
032512	蘩	1-91-43	1
032820	虵	1-91-51	1
033494	螭	1-91-62	1
034292	裎	1-91-75	1
034457	褚	1-91-82	1
034816	覓	1-91-88	1
035968	譔	1-92-18	1
035976	譙	1-92-19	1
036404	豨	1-92-23	1
036878	賸	1-92-27	1

037473	跗	1-92-35	1
037887	蹐	1-92-40	1
039413	郤	1-92-69	1
039476	郷	1-92-72	1
039630	鄧	1-92-80	1
039684	鄴	1-92-83	1
040447	鋌	1-93-17	1
040940	鐳	1-93-40	1
041022	鑢	1-93-42	1
041053	鑲	1-93-45	1
041283	閧	1-93-47	1
041330	閥	1-93-50	1
041367	閶	1-93-51	1
041430	関	1-93-53	1
041467	闥	1-93-57	1
041650	陘	1-93-59	1
041849	隕	1-93-63	1
042230	雯	1-93-69	1
043191	韲	1-93-83	1
043269	師	1-93-85	1
043357	頊	1-93-87	1
043463	類	1-93-89	1
043599	顯	1-93-92	1
043600	顳	1-93-93	1
043614	顼	1-94-01	1
044779	駸	1-94-12	1
044780	駿	1-94-13	1

045597	鬻	1-94-29	1
045969	魛	1-94-32	1
046046	鯱	1-94-37	1
046071	鮭	1-94-39	1
046249	鰈	1-94-44	1
047204	鷁	1-94-66	1
048837	龔	1-94-87	1
050021	墜	1-87-25	1
053024	劑	1-91-14	1
053212	异	1-84-18	1
056028	廋	1-84-15	1
056061	燁	1-87-62	1
056068	璘	1-88-25	1
056083	磷	1-89-14	1
056138	芾	1-90-69	1
056142	苕	1-90-72	1
056206	荸	1-90-90	1
056209	莉	1-90-91	1
056220	若	1-91-03	1
056272	萌	1-91-15	1
056273	蒺	1-91-16	1
056380	蔑	1-91-40	1
056382	蘘	1-91-42	1
057035	機	1-86-21	1
058230	鯽	1-94-46	1
065501	籥	1-89-78	1

第 4 水準漢字で表現できる文字は、次の 426 字（表 6 参照）である。『太陽コーパス』での使用度数順に示す。

表 6：第 4 水準漢字で表現できるもの

文字番号	字形	面区点	度数
018019	瀛	2-79-06	252
034969	覿	2-88-42	21
025263	穌	2-83-03	17
056318	壹	2-86-81	17

057871	熳	2-80-01	13
028121	欽	2-84-66	12
056006	僊	2-03-04	12
006573	媳	2-05-70	11
013431	斃	2-13-72	11

000763	倏	2-01-57	10
008848	帕	2-08-83	10
032958	蛢	2-87-41	10
039805	酧	2-90-33	10
025324	穠	2-83-08	9
026062	筩	2-83-48	9
033703	蠲	2-87-92	9
039044	遝	2-89-93	9
007253	寖	2-08-07	8
012586	撜	2-13-41	8
012932	攇	2-13-58	8
019727	爹	2-80-13	8
025582	窯	2-83-17	8
041329	闆	2-91-56	8
050013	健	2-12-24	8
056286	蓯	2-86-65	8
056355	蘼	2-87-04	8
005805	貪	2-05-29	7
011826	扭	2-12-93	7
012694	撑	2-13-47	7
013406	斀	2-13-70	7
018777	灂	2-79-53	7
028367	屬	2-84-80	7
038473	轀	2-89-67	7
041601	陋	2-91-67	7
059756	焰	2-79-88	7
001798	凳	2-03-20	6
002689	匾	2-03-48	6
004502	嚕	2-04-45	6
011562	戕	2-12-83	6
016647	殺	2-78-04	6
018092	滹	2-79-10	6
024043	砐	2-82-28	6
028971	耬	2-85-09	6
029758	瞍	2-85-45	6
036786	賒	2-89-12	6
037149	趕	2-89-23	6

040542	錕	2-91-07	6
041362	閬	2-91-57	6
056323	蕞	2-86-82	6
000964	傖	2-01-77	5
003874	啡	2-04-08	5
004350	唒	2-04-39	5
008458	嶒	2-08-63	5
008477	嶙	2-08-66	5
011216	慦	2-12-72	5
015945	欄	2-15-85	5
018062	滙	2-79-07	5
019179	煊	2-79-84	5
023532	睽	2-82-11	5
028167	罄	2-84-70	5
032882	蚜	2-87-34	5
033873	鐲	2-88-02	5
033968	邮	2-88-04	5
037648	踢	2-89-38	5
041425	鬧	2-91-59	5
044024	飣	2-92-46	5
044179	餇	2-92-55	5
045359	髡	2-93-19	5
051023	壳	2-05-22	5
002281	劇	2-03-32	4
005746	夔	2-05-28	4
006527	媢	2-05-68	4
007811	屧	2-08-20	4
011960	挖	2-13-03	4
014483	杓	2-14-34	4
015347	橐	2-15-30	4
018239	澅	2-79-21	4
020774	獷	2-80-55	4
021777	吠	2-81-26	4
027368	紟	2-84-17	4
027532	網	2-84-33	4
027775	縝	2-84-51	4
032833	虺	2-87-29	4

034257	裀	2-88-18	4
034299	裒	2-88-19	4
035873	謗	2-88-73	4
035934	謹	2-88-74	4
037750	躅	2-89-44	4
039260	邐	2-90-04	4
040001	醨	2-90-40	4
040957	鑲	2-91-44	4
042577	靛	2-91-94	4
042957	鞘	2-92-08	4
043881	颼	2-92-35	4
043921	颶	2-92-38	4
044278	餂	2-92-63	4
044453	饜	2-92-73	4
048477	饑	2-94-69	4
053339	瞢	2-82-16	4
058529	蠒	2-87-81	4
065717	社	2-88-09	4
065718	衧	2-88-10	4
000680	俏	2-01-52	3
003579	咿	2-03-85	3
003808	啊	2-04-05	3
004974	坨	2-04-68	3
006214	姣	2-05-49	3
006215	姤	2-05-50	3
006720	嫣	2-05-81	3
006776	嬴	2-05-84	3
006932	孑	2-05-87	3
007058	宄	2-05-93	3
007804	屧	2-08-19	3
008172	崐	2-08-45	3
009125	幫	2-08-92	3
011158	慼	2-12-68	3
012096	挲	2-13-11	3
014580	枰	2-14-44	3
014759	桅	2-14-64	3
015449	椹	2-15-45	3

015660	檑	2-15-62	3
016856	搓	2-78-12	3
020286	犹	2-80-30	3
020895	玷	2-80-66	3
022453	瘳	2-81-69	3
025126	稞	2-82-92	3
025561	窣	2-83-16	3
029175	瞶	2-85-14	3
030449	艋	2-85-75	3
032583	蘼	2-87-21	3
033147	蝑	2-87-54	3
033202	蜺	2-87-58	3
033320	蝌	2-87-66	3
035660	詮	2-88-65	3
037865	蹯	2-89-49	3
044316	餼	2-92-67	3
045653	鬮	2-93-28	3
045681	虙	2-93-29	3
045861	魋	2-93-32	3
046890	鵅	2-94-14	3
047195	鸙	2-94-32	3
048261	龕	2-94-62	3
057479	鼯	2-94-68	3
057773	艉	2-85-72	3
078863	蹂	2-89-31	3
002190	劂	2-03-30	2
003701	唉	2-03-94	2
003910	喈	2-04-13	2
004069	嗒	2-04-16	2
005253	堠	2-05-01	2
006232	姮	2-05-52	2
006568	嫮	2-05-69	2
006701	嫵	2-05-80	2
007051	孿	2-05-91	2
009127	幬	2-08-93	2
010319	忒	2-12-30	2
011106	憿	2-12-67	2

011218	憘	2-12-73	2
011970	挐	2-13-04	2
012359	揜	2-13-31	2
013286	敫	2-13-68	2
014758	桄	2-14-63	2
015295	榼	2-15-27	2
017283	泱	2-78-33	2
017786	淔	2-78-82	2
018920	茨	2-79-63	2
019337	煡	2-79-92	2
019935	物	2-80-18	2
020730	獫	2-80-49	2
022631	癮	2-81-77	2
023466	罼	2-82-07	2
024392	磈	2-82-48	2
024495	磾	2-82-58	2
024713	祋	2-82-70	2
026299	篔	2-83-63	2
026314	篚	2-83-65	2
026458	簏	2-83-71	2
026801	籩	2-83-79	2
027489	綆	2-84-30	2
028006	繾	2-84-58	2
028143	缾	2-84-68	2
030107	臬	2-85-57	2
030564	艣	2-85-82	2
034945	覷	2-88-41	2
035344	訾	2-88-57	2
036168	讜	2-88-84	2
036207	谿	2-88-88	2
038099	躶	2-89-55	2
039794	酗	2-90-32	2
040296	鉎	2-90-60	2
040937	鐲	2-91-42	2
041945	雋	2-91-79	2
043812	飇	2-92-33	2
044199	餚	2-92-58	2
044480	饢	2-92-74	2
044967	驄	2-93-03	2
045535	髺	2-93-24	2
045543	鬒	2-93-25	2
047507	鸝	2-94-49	2
048274	鼁	2-94-63	2
048361	鼙	2-94-67	2
053340	慒	2-12-81	2
053365	潛	2-79-24	2
056074	瘼	2-81-68	2
056234	萹	2-86-38	2
056311	蘐	2-86-74	2
056387	藱	2-87-18	2
056394	虌	2-87-23	2
056433	轐	2-92-10	2
056462	棻	2-14-86	2
057354	箐	2-83-52	2
057486	酁	2-90-37	2
066039	鮅	2-93-77	2
000107	乀	2-01-08	1
000510	佔	2-01-36	1
000725	傛	2-01-55	1
000843	偎	2-01-65	1
001051	僇	2-01-85	1
001189	儈	2-03-01	1
001797	兢	2-03-19	1
002307	劦	2-03-35	1
002380	勅	2-03-41	1
003254	匝	2-03-67	1
003387	呃	2-03-72	1
003437	呦	2-03-74	1
003535	咦	2-03-80	1
003585	哃	2-03-86	1
003790	啁	2-04-02	1
003804	啇	2-04-04	1
003889	噂	2-04-11	1
004138	噆	2-04-27	1

004194	嘐	2-04-31	1
004625	瞲	2-04-51	1
004926	垒	2-04-65	1
004988	坳	2-04-70	1
005248	堞	2-04-94	1
005433	墔	2-05-14	1
005698	夆	2-05-25	1
006461	婺	2-05-63	1
006575	媵	2-05-71	1
006600	媿	2-05-73	1
006655	嫈	2-05-76	1
006734	嬈	2-05-82	1
006912	孌	2-05-86	1
007433	尃	2-08-13	1
008186	崛	2-08-46	1
008267	崔	2-08-53	1
008532	嵩	2-08-68	1
008549	嶂	2-08-71	1
008702	巋	2-08-77	1
008843	帑	2-08-82	1
009134	幮	2-12-01	1
009311	庶	2-12-03	1
010320	忏	2-12-31	1
010403	忼	2-12-37	1
010661	悕	2-12-46	1
010718	悰	2-12-49	1
010933	愒	2-12-59	1
011021	惮	2-12-63	1
011244	憨	2-12-75	1
012013	拽	2-13-05	1
012050	挍	2-13-07	1
012054	挐	2-13-08	1
012101	挹	2-13-12	1
012127	捄	2-13-16	1
012148	捎	2-13-17	1
012238	掄	2-13-23	1
012265	挣	2-13-24	1
012510	掐	2-13-40	1
012587	摛	2-13-42	1
012678	撇	2-13-46	1
012802	擋	2-13-50	1
012857	擃	2-13-55	1
014013	暔	2-14-09	1
014266	曬	2-14-21	1
014495	柿	2-14-36	1
014936	棖	2-14-94	1
014990	概	2-15-05	1
015378	槳	2-15-39	1
015851	櫬	2-15-75	1
015868	櫳	2-15-78	1
016517	殛	2-78-01	1
016983	穖	2-78-15	1
017085	氵	2-78-17	1
017157	汭	2-78-24	1
017320	沆	2-78-39	1
017397	泊	2-78-45	1
017398	洧	2-78-46	1
017546	涘	2-78-59	1
017593	涴	2-78-67	1
017827	湏	2-78-88	1
018177	滤	2-79-16	1
018489	澡	2-79-36	1
018709	瀹	2-79-46	1
018881	灾	2-79-59	1
019018	烜	2-79-73	1
019069	焄	2-79-75	1
019210	煞	2-79-86	1
019410	燋	2-80-03	1
019774	牉	2-80-15	1
020430	狟	2-80-36	1
020495	狰	2-80-40	1
020646	獘	2-80-47	1
020754	獯	2-80-53	1
020804	獵	2-80-56	1

020857	玦	2-80-62	1
021060	玲	2-80-78	1
021248	琉	2-81-02	1
022154	痄	2-81-46	1
023224	眴	2-81-91	1
023867	稍	2-82-20	1
024396	碘	2-82-49	1
024399	磎	2-82-50	1
024643	祊	2-82-65	1
024693	祧	2-82-69	1
025407	穸	2-83-11	1
025440	牢	2-83-13	1
026057	笂	2-83-47	1
026136	箑	2-83-53	1
027370	絀	2-84-18	1
027377	絇	2-84-19	1
027438	綑	2-84-25	1
027636	綀	2-84-41	1
027660	緦	2-84-43	1
027854	繅	2-84-55	1
027960	繳	2-84-56	1
028027	纇	2-84-60	1
028370	罋	2-84-81	1
028639	翍	2-84-90	1
028860	耄	2-85-03	1
028880	耑	2-85-06	1
029508	脞	2-85-33	1
030206	舄	2-85-62	1
030438	舿	2-85-73	1
030477	艏	2-85-77	1
032852	蚍	2-87-32	1
033137	蜙	2-87-53	1
033208	蜾	2-87-59	1
033213	蝀	2-87-60	1
033268	蝘	2-87-63	1
033570	蠴	2-87-80	1
033682	蠻	2-87-90	1

033745	蠓	2-87-93	1
034106	衱	2-88-11	1
034353	裱	2-88-25	1
035750	諼	2-88-66	1
036120	譻	2-88-81	1
036435	豭	2-89-03	1
037475	跙	2-89-28	1
037617	踔	2-89-35	1
038190	軔	2-89-59	1
038721	迁	2-89-76	1
038988	遘	2-89-92	1
039822	酡	2-90-34	1
039926	酟	2-90-39	1
040031	醮	2-90-41	1
040106	釃	2-90-44	1
040184	釨	2-90-50	1
040205	鈁	2-90-51	1
040216	鈇	2-90-52	1
040291	鉊	2-90-59	1
040351	鉼	2-90-71	1
040506	鋹	2-91-03	1
040556	錝	2-91-08	1
040642	鏊	2-91-18	1
040770	鏐	2-91-32	1
041663	陡	2-91-68	1
041895	隟	2-91-77	1
042864	鞚	2-92-05	1
043917	颿	2-92-37	1
043964	颱	2-92-41	1
043973	颺	2-92-42	1
044022	飧	2-92-45	1
044202	餛	2-92-59	1
044637	駉	2-92-82	1
044665	駔	2-92-83	1
044677	駙	2-92-84	1
044928	騶	2-93-02	1
045124	骯	2-93-07	1

046431	鰺	2-93-73	1
046432	鰌	2-93-74	1
046626	鱶	2-93-94	1
046667	鳰	2-94-02	1
047022	鵐	2-94-23	1
047208	鷯	2-94-34	1
047388	鷸	2-94-44	1
047733	麹	2-94-55	1
048278	龕	2-94-65	1
048351	鑿	2-94-66	1
048519	齣	2-94-72	1
048543	齦	2-94-73	1
048590	齟	2-94-76	1
050083	吕	2-08-79	1
050925	亹	2-01-20	1
051896	帮	2-08-86	1
053107	偶	2-01-59	1
053464	嬝	2-05-74	1
053509	槃	2-86-79	1
053616	贏	2-87-91	1
053620	鄺	2-90-29	1
054860	牖	2-80-17	1
056042	暐	2-14-11	1
056043	曙	2-14-16	1

056062	熱	2-80-07	1
056069	厩	2-81-11	1
056112	芃	2-85-90	1
056116	芋	2-85-91	1
056179	茨	2-86-16	1
056229	荏	2-86-34	1
056281	蘋	2-86-59	1
056326	蕢	2-86-83	1
056375	薜	2-87-13	1
056458	鱚	2-93-84	1
057196	耆	2-82-32	1
057313	聃	2-85-11	1
057315	聱	2-85-13	1
057368	籯	2-83-81	1
057591	虓	2-87-25	1
057637	髀	2-93-15	1
057641	骷	2-93-09	1
057850	蕨	2-86-29	1
065806	欽	2-90-48	1
067184	餈	2-92-61	1
079011	鰥	2-93-71	1
079145	路	2-89-41	1
079566	澎	2-79-39	1

X0213 非漢字で表現できる文字は、次の 38 字（表 7 参照）である。『太陽コーパス』での使用度数順に示す。

表 7：X0213 非漢字で表現できるもの

文字番号	字形	面区点	度数
063004	â	1-09-56	51
063017	é	1-09-63	30
063020	ê	1-09-64	21
063070	ǔ	1-10-68	13
063003	ä	1-09-58	10
063034	î	1-09-68	10
063068	ü	1-09-81	10

062960	Ś	1-10-05	9
063018	è	1-09-62	8
063065	ţ	1-10-55	6
063069	û	1-09-80	5
063051	ö	1-09-76	4
063060	ś	1-10-16	4
069682	↰	1-03-28	4
063002	à	1-09-54	3

063014	ç	1-09-61	3
062833	æ	1-09-60	2
062845	œ	1-11-10	2
063048	ñ	1-09-71	2
063052	ô	1-09-74	2
063053	ǒ	1-08-87	2
063055	ō	1-09-94	2
063221	Ⅰ	1-13-21	2
062588	Ⴝ	1-06-57	1
062952	Ô	1-09-43	1
062968	Ǔ	1-09-50	1
063001	ά	1-09-55	1

063005	ă	1-10-41	1
063009	å	1-09-59	1
063023	ē	1-09-93	1
063030	ĥ	1-10-65	1
063031	í	1-09-67	1
063033	ï	1-09-69	1
063049	ó	1-09-73	1
063066	ú	1-09-79	1
063222	Ⅱ	1-13-22	1
063223	Ⅲ	1-13-23	1
063224	Ⅳ	1-13-24	1

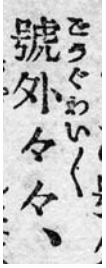
3.2　踊字タグ

JIS X 0208 には「二の字点」や「くの字点」がないため、『太陽コーパス』では「々」「ゝ」で代用し、踊字タグを使って表現している。次の例は、平仮名「る」の繰り返し符号に使われた「二の字点」を「ゝ」で代用した例である。

（例 3）　意を戦況に注がる<踊字　種類="二字点">ゝ</踊字>の致す處なりと雖も、〔t189501〕

また、繰り返し符号が複数連続する場合にも踊字タグが使われている。この場合は、JIS X 0208 で表現できないからタグを用いているわけではなく、繰り返し符号で表現された語が何であるのかを、タグの属性で示すのが目的である。

（例 4）　號外<踊字　値="號外">々々</踊字>〔t189511〕



踊字タグの再処理結果をまとめると表 8 のようになる。JIS X 0213 を用いると、『太陽コーパス』に出現するすべての繰り返し符号を表現することができる。なお、X0208 非漢字に分類されたものは、（例 4）に類するものである。

表 8：踊字タグの再処理結果

	のべ字数	異なり字数	用例
X0208 非漢字	1,444	4	々、ゝ、ゝ、ゞ
X0213 非漢字	16,575	4	二の字点、くの字点（上、上濁点、下）
X0213 外字	0	0	
計	18,019	8	

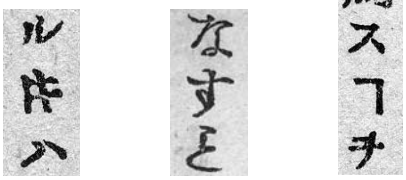
3.3　合字タグ

『太陽コーパス』では、仮名合字等に対して、合字が表わす語形を入力し、合字タグを付与している。合字タグの使用例を示す。

（例 5）　郵便切手ヲ代用スル<合字>トキ</合字>ハ五厘又ハ壹錢切手ニ限ル〔t189501〕

（例 6）　其用意をなす<合字>こと</合字>肝要なるべし。〔t189501〕

（例 7）　條約締盟國ハ日本國中何レノ地ニ於テモ通商貿易ヲ爲ス<合字>コト</合字>ヲ得領事裁判權ハ三年以内ニ之ヲ廢止シ〔t189501〕



合字タグの再処理結果をまとめると表 9 のようになる。『太陽コーパス』に出現する合字に対しては、JIS X 0213 はあまり効果がない。

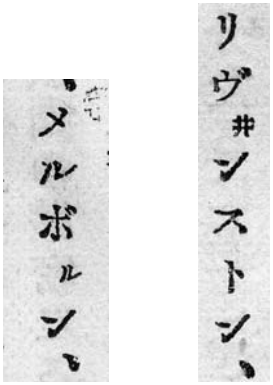
表 9：合字タグの再処理結果

	のべ字数	異なり字数	用例
X0213 非漢字	23	2	コト、より
X0213 外字	8,531	5	こと、トキ、トモ、かしこ、まゐらせ候
計	8,554	7	

3.4　小書タグ

『太陽コーパス』では、主として小書きの仮名を表わすために、小書タグを設けている。外来語表記のための片仮名が目立つ。

（例 8） メルボ<小書>ル</小書>ン〔t189510〕
（例 9） 亞弗利加の中心に深進せしリヴ<小書>ヰ</小書>ンストン〔t189510〕



小書タグの再処理結果をまとめると表 10 のようになる。小書タグのうち、JIS X 0213 を用いて符号化できるものは、のべ字数で 3 割程度にとどまる。なお、X0208 非漢字に分類した小書きのワは、コーパス開発時の単純なバグであろう。

表 10：小書タグの再処理結果

	のべ字数	異なり字数	用例
X0208 非漢字	1	1	ワ
X0213 非漢字	56	7	ハ、ヒ、フ、ヘ、ホ、ム、ル
X0213 外字	130	8	キ、コ、ヰ、ヱ、ヲ、八、忘、●
計	187	12	

なお、片仮名以外の文字について小書タグを使ったものには、次のようなものがある。（例 10）は伏せ字、（例 11）は割注に相当するものと思われる。

（例 10） 亞細亞洲にては●●、●●に在り、近南洋にては●<小書>●</小書>●<小書>●</小書>●●群島、●●大島に在り、〔t189510〕
（例 11） 大兒子、小兒子、大則以王、小則以覇、大小王<小書>忘</小書>覇<小書>八</小書>兒子〔t190114〕

4．おわりに

『太陽コーパス』の文字関連タグに対する再処理結果を加え、JIS X 0213 による符号化をまとめると表 11 のようになる。のべ字数では、JIS X 0208 によるカバー率が 99.79%であったのに対して、JIS X 0213 では 99.93%となる。カバー率の増

加はわずかに 0.15 ポイント程度である。しかし、異なり字数では、JIS X 0208 のカバー率 79.16%に対して、JIS X 0213 は 92.06%と、12.90 ポイントも増加している。

『現代日本語書き言葉均衡コーパス』の書籍（生産実態サブコーパスと流通実態サブコーパス）では、JIS X 0208→JIS X 0213 によるカバー率の増加は、のべ字数で平均 0.03 ポイント、異なり字数で平均 7.75 ポイントであるから、現代語を扱うときよりも、JIS X 0213 を用いる効果が大きい。『太陽コーパス』のような近代語文献を電子化する際には、JIS X 0208 文字セットよりも、JIS X 0213 文字セットの方が適していると言えよう。

表 11：『太陽コーパス』の JIS X 0213 による符号化

	のべ字数	異なり字数
第 1 水準漢字	5,650,619	2,721
第 2 水準漢字	888,886	2,864
X0208 非漢字	7,798,780	318
第 3 水準漢字	2,685	446
第 4 水準漢字	1,371	426
X0213 非漢字	16,874	51
（小計）	14,359,215	6,826
X0213 外字	9,890	592
計	14,369,105	7,418

最後に、JIS X 0213 文字セットで表現できない 592 字を、『太陽コーパス』での使用度数順に表 12 に示す。

表 12：JIS X 0213 外字

文字番号	字形	度数
合字こと		8,487
小書ヰ		105
T003	苍	62
047550	𪜈	55
019790	牒	32
067692	𣏌	27
058594	颺	24
合字まゐらせ候		22
020571	猪	21
012384	揪	20

文字番号	字形	度数
073740	璨	18
057321	蚋	17
058090	圻	14
T014	纊	14
合字トキ		14
027431	紙	11
044757	騙	11
003007	厘	10
007255	竄	10
011016	惣	10
038183	軏	10

020803	玀	9
小書ヲ		9
010846	鬼	8
015158	椴	8
020483	猓	8
014585	枴	7
063059	r	7
小書ヱ		7
004658	嚩	6
021193	璀	6
065520	縩	6
074686	滀	6
T025	閧	6
013004	攪	5
020687	獞	5
059905	嶚	5
066170	厄	5
074461	誓	5
小書●		5
合字トモ		5
001706	漸	4
002225	劊	4
011737	扃	4
025000	秖	4
028693	翎	4
030766	芾	4
034535	褻	4
035801	謏	4
037152	趍	4
037993	躝	4
042302	霉	4
044479	饞	4
048306	鼉	4
048652	鮂	4
050664	麦	4
058129	耦	4
077953	晉	4

080074	洌	4
000575	佚	3
001171	僾	3
004183	嘍	3
004259	嘵	3
004981	坭	3
005403	塿	3
007191	寘	3
009463	廕	3
012241	掇	3
012712	撙	3
013751	旰	3
015675	櫆	3
018819	灡	3
019378	爦	3
020335	狯	3
020672	獎	3
024552	礉	3
030411	舺	3
033980	衅	3
034974	覷	3
039942	醬	3
043609	賴	3
044963	鷙	3
056123	芠	3
059815	碉	3
071309	垔	3
071516	粵	3
071768	蔡	3
075506	暺	3
083904	衙	3
T002	啖	3
合字かしこ		3
001134	夾	2
001584	冢	2
001953	刱	2
002041	剚	2

003478	咂	2
003673	哜	2
003806	啉	2
004202	嗎	2
004242	嘮	2
004252	嚫	2
004455	嚅	2
005133	垐	2
005261	堦	2
005404	墀	2
007285	寠	2
008263	嵉	2
008753	啻	2
008760	啻	2
011115	慦	2
011525	戇	2
012124	捂	2
012505	搪	2
012595	摟	2
012728	撦	2
012897	擰	2
012991	攙	2
013668	旐	2
013984	晾	2
014070	暯	2
014163	暟	2
015492	毃	2
017171	汸	2
017492	洚	2
017784	淯	2
019202	煞	2
019430	燇	2
020274	狉	2
020336	狹	2
020377	狗	2
020722	獷	2
020961	琅	2

022784	皤	2
023619	瞕	2
023633	瞪	2
024669	袂	2
025778	娖	2
026318	筃	2
026455	簌	2
027483	綁	2
027831	繁	2
028824	翾	2
029426	胭	2
032114	蘼	2
032411	藁	2
033770	蠹	2
035233	訐	2
035466	註	2
036843	賬	2
037787	蹏	2
039791	酗	2
040185	釭	2
040350	銘	2
040514	錁	2
041574	阱	2
044678	駃	2
044710	駧	2
045468	鬍	2
045524	鬣	2
047440	鸉	2
049191	齾	2
049534	胳	2
050020	灣	2
050031	贊	2
056037	撐	2
056114	芾	2
056412	薆	2
057125	螣	2
057128	螣	2

059535	峬	2
059665	勝	2
059875	舿	2
062581	𠂉	2
067039	荎	2
067352	搫	2
077217	遲	2
081515	髺	2
085829	簨	2
T007	礗	2
T016	睭	2
T020	蛵	2
T105	爭	2
T117	鑑	2
T125	鯽	2
T126	磔	2
T135		2
T136		2
T137		2
T138		2
T146		2
000484	伻	1
001119	僨	1
001154	僱	1
001505	顯	1
001955	刲	1
002094	剛	1
002217	剝	1
002683	匽	1
002854	卲	1
003235	另	1
003322	呷	1
003368	吵	1
003382	呖	1
003477	咁	1
003628	哎	1
003756	唬	1

003811	唇	1
003940	喔	1
004041	嗃	1
004139	嗶	1
004190	嘏	1
004195	嘍	1
004348	噙	1
004979	坫	1
005080	垌	1
005208	堊	1
005323	塍	1
005326	塏	1
005425	墀	1
005467	墦	1
005560	壖	1
006124	姒	1
006235	姱	1
006361	娀	1
006392	婇	1
006405	婑	1
006579	媸	1
006787	嫚	1
007957	岏	1
007958	岝	1
008175	崒	1
008326	嵎	1
008367	嵲	1
008391	嵽	1
008460	嶓	1
008556	巖	1
008579	巑	1
008758	帨	1
009052	幙	1
009081	幠	1
009245	庱	1
009605	弇	1
010394	忯	1

010675	悜	1
010751	愩	1
010766	悽	1
010839	悊	1
010937	愔	1
011167	憀	1
011225	憏	1
011475	憫	1
012131	捆	1
012316	摃	1
012453	摛	1
012491	搠	1
012672	摔	1
012693	撐	1
012951	攉	1
012957	攦	1
013627	旁	1
013664	旃	1
013685	旖	1
013686	旐	1
013691	旚	1
013780	旰	1
013963	晅	1
014045	暌	1
014817	棒	1
015232	橅	1
015356	槤	1
015718	檯	1
015903	欀	1
016017	欴	1
016476	殑	1
017166	汾	1
017169	汊	1
017179	沿	1
017510	浼	1
017594	浮	1
017596	涯	1
017685	淣	1
017714	淩	1
017793	淘	1
017853	湢	1
018256	溧	1
018320	溦	1
018388	澦	1
018392	澨	1
019134	煢	1
019379	燘	1
019896	牘	1
019987	牳	1
020054	犄	1
020067	犏	1
020251	犻	1
020391	狪	1
020394	猗	1
020452	狶	1
020456	狒	1
020487	猏	1
020618	猓	1
020680	猾	1
020692	獠	1
020879	玭	1
020988	琌	1
020990	琅	1
021041	琲	1
021208	瑫	1
021239	瑣	1
021564	瑩	1
021749	畀	1
021828	畟	1
022451	癡	1
022685	皂	1
022787	皾	1
022815	皽	1
023120	盱	1

023178	明	1
023373	睠	1
023555	瞅	1
024088	砢	1
024178	硐	1
024200	硜	1
024249	硢	1
024330	硜	1
024369	碩	1
024568	礧	1
024613	礸	1
024734	禍	1
024737	禍	1
024761	禋	1
025094	稕	1
025368	穱	1
025438	窕	1
025966	笻	1
026109	笄	1
026132	篂	1
026158	篳	1
026309	篛	1
026742	篫	1
026985	糕	1
027439	絛	1
027499	綍	1
027538	綞	1
027701	緼	1
028350	罿	1
028440	狌	1
028919	耟	1
028978	耮	1
029544	脰	1
029582	腎	1
029593	腄	1
029627	腒	1
029683	腠	1

029895	膴	1
030055	臞	1
030202	舄	1
030279	刮	1
030472	艎	1
030818	荼	1
031589	萩	1
031616	蒚	1
032354	蔰	1
032871	蚖	1
032941	蛆	1
033255	蜳	1
033359	蜾	1
033367	蜴	1
033415	蝖	1
033612	蟠	1
033747	蠡	1
034034	衍	1
034339	裍	1
034355	裗	1
034470	褒	1
035019	觚	1
035283	詈	1
035292	訞	1
035384	詖	1
035596	誶	1
035636	諉	1
035656	諕	1
035947	譊	1
035991	贗	1
036050	譸	1
036515	豜	1
036713	貦	1
036747	賚	1
036955	贈	1
036960	賴	1
037004	赮	1

037084	趣	1
037124	趍	1
037303	趫	1
037508	跕	1
037560	跹	1
037568	跲	1
037716	蹀	1
037806	蹚	1
037847	蹶	1
037881	躆	1
037886	蹺	1
037962	躐	1
038414	轄	1
039296	邶	1
039851	酎	1
039874	酺	1
039998	醔	1
040104	醨	1
040162	釗	1
040166	釵	1
040226	釬	1
040253	鉢	1
040262	鈰	1
040338	鈾	1
040369	鋍	1
040429	鋅	1
040437	鋑	1
040474	鋓	1
040489	鋐	1
040798	鏕	1
040865	鏟	1
040921	鏦	1
040938	鐫	1
041008	鐳	1
041045	鑐	1
041078	鑞	1
041221	閈	1
041262	閣	1
041289	閾	1
041428	闉	1
041723	隑	1
041810	隔	1
043414	頒	1
043598	顑	1
043605	顡	1
043813	颭	1
043931	飂	1
043955	飋	1
044084	詐	1
044126	飪	1
044236	餧	1
044294	餵	1
044362	饈	1
044799	騕	1
044847	驄	1
044921	驍	1
044962	驟	1
045305	髗	1
045841	魈	1
046039	鮀	1
046059	鮇	1
046095	鮋	1
046119	鮭	1
046407	鰶	1
046856	鶂	1
047937	鼓	1
048352	鼕	1
049388	耆	1
050022	琍	1
050860	緆	1
050971	嫧	1
053173	摔	1
053192	蓍	1
053438	嫫	1

053607		1
053614		1
054924		1
056017		1
056035		1
056080		1
056118		1
056132		1
056150		1
056248		1
056357		1
056359		1
056398		1
056402		1
056407		1
056410		1
057215		1
057361		1
057461		1
057530		1
057851		1
058025		1
058029		1
058660		1
059216		1
059236		1
059247		1
059271		1
059618		1
059803		1
059921		1
060001		1
060081		1
061327		1
061593		1
062583		1
062587		1
062589		1
063047		1
063084		1
065057		1
065167		1
065183		1
065417		1
065538		1
065646		1
065972		1
066007		1
066167		1
066767		1
067720		1
067843		1
068171		1
068275		1
070734		1
072382		1
072508		1
075331		1
076687		1
077626		1
078439		1
078618		1
079135		1
082432		1
082959		1
085185		1
085262		1
086272		1
086719		1
096371		1
096381		1
096461		1
096551		1
T001		1

T004		1
T005		1
T006		1
T008		1
T009		1
T010		1
T011		1
T012		1
T013		1
T015		1
T017		1
T018		1
T019		1
T021		1
T022		1
T023		1
T024		1
T102		1
T103		1
T104		1
T106		1
T107		1
T108		1
T109		1
T110		1
T111		1
T112		1

T113		1
T114		1
T115		1
T116		1
T118		1
T119		1
T120		1
T121		1
T122		1
T123		1
T124		1
T127		1
T128		1
T129		1
T130		1
T131		1
T132		1
T133		1
T134		1
T139		1
T140		1
T145		1
小書八		1
小書キ		1
小書コ		1
小書忘		1

文　献

池田証寿・白井純・高田智和(2002)「宋版漢字字体の処理」『京都大学大型計算機センター第 69 回研究セミナー報告　東洋学へのコンピュータ利用』pp.49-62

下田正弘・師茂樹(1999)「大正新脩大蔵経データベース（SAT）における外字問題」『人文学と情報処理』25, pp.35-43

須永哲矢・堤智明・高田智和(2011)「明治前期雑誌の異体漢字と文字コード」『人文科学とコンピュータシンポジウム論文集　「デジタル・アーカイブ」再考―いま改めて問う記録・保存・活用の技術―』pp.381-388

高田智和(2002)「漢字処理と『大字典』」、『訓点語と訓点資料』109, pp.99-107
高田智和・小林正行・間淵洋子・大島一・西部みちる・山口昌也(2009)『JIS X 0213:2004 運用の検証（国立国語研究所内部報告書 LR-CCG-09-01)』国立国語研究所
高田智和(2011)「現代日本語コーパスにおける文字処理」『人間文化研究情報資源共有化研究会報告集』2, pp.31-40
田中牧郎(2005)「漢字の実態と処理の方法」『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―（国立国語研究所報告 122)』博文館新社, pp.271-292
當山日出夫(2009)「『内村鑑三全集』デジタル版の文字処理について」『東洋学へのコンピュータ利用第 20 回セミナー』京都大学人文科学研究所附属漢字情報センター, pp.5-18
富田倫生(2000)「青空文庫と外字」、『人文学と情報処理』26, pp.23-30
安永尚志(1998)『国文学研究とコンピュータ』勉誠社