『太陽コーパス』の漢字表記語に対する熟字ルビについて

間淵 洋子（国立国語研究所 コーパス開発センター）†

On the Ruby Elements in Taiyo Corpus: The Case of Group-ruby

MABUCHI, Yoko (Center for Corpus Development, NINJAL)

１．はじめに

国立国語研究所コーパス開発センターでは，現在「通時コーパス」プロジェクトの一環として，形態論情報付きの近代語コーパスを構築している。これまでに，2012 年『明六雑誌コーパス』が公開されているほか，2014 年度には『国民之友コーパス』の公開も予定されており，今後も資料を拡充していく計画である。その一つが雑誌『太陽』であり，2005 年に公開された『太陽コーパス』を増補改訂し，新たに形態論情報付きコーパスとして構築し直す準備を進めている。

形態論情報を付与するにあたっては，言語の表層情報，すなわち表記を元に，語を割り当ててゆくという方法を取ることになるが，その際問題になるのが，異なる複数の語を同一の表記で表す「同表記異音語」の存在である（伝他 2008）。本発表では同一の表記で表される異なる語を「同表記異語」と呼ぶことにする。

例えば，「上方」という表記は，京阪地域を表す「かみがた」という語（A）と上の方を表す「じょうほう」という語（B）の二つの異なる語のいずれをも表し得る形である。

(1)　【上方】落語には「ぜんざい公社」という演目がある。（BCCWJ:PN2b_00005；毎日新聞 2002/9/27 朝刊）
(2)　またこれまでの推移から、主力商品、売れ筋商品の目標数を【上方】修正する。（BCCWJ:PB46_00066；宮崎文明 2004『商売の「計画・実施・検証」；実例による販売計画と商品管理の原則』）

「上方」の場合は，上例のように接続する語あるいは文脈によって，いずれの語かを判断することは容易で，例(1)は語(A)に，例(2)は語(B)にそれぞれ対応することが分かる。しかし，このような同表記異語の中には，それぞれの語同士が同義または極めて近接した意味を持つために，文脈を手掛かりに語を対応付けるのが困難な場合もある。特に，いわゆる「熟字訓」と呼ばれる表記と語の対応は，そもそも和語とそれに極めて近接的な意味を持つ漢語を結びつけることによって成り立っているために，文脈に基づく語の判定が容易でないことが多く，語種や文体といった手掛かりですら無力な場合も少なくない。

そのような場合に決定的な手掛かりとして用いることができるのが，表記に付された振り仮名（ルビ）である。ルビによって，読みを確定することができれば，その表記がどの語を表すための表層形かを知ることができる。

(3)　光代、【明日（あす）】は夙く發たうぞ。（太陽 1895 年 2 号 川上眉山「書記官」）
(4)　今では局長が引受て、萬事表面上商會の世話をして居る仲であつて見れば、すでに

† mabuchi@ninjal.ac.jp

【明日（みょうにち）】か二三日中に願書が出来て、（太陽 1895 年 2 号 川上眉山「書記官」）

『太陽コーパス』は，本文行の文字列に加えてそれに対応付けた形で詳細なルビ情報を付与しており，これを観察することによって，表記と語の結び付きの実態を把握できる可能性がある。

本発表では，同表記異語における表記と語の結び付きの実態把握を目指し，その第一段階として，『太陽コーパス』におけるルビ情報のうち，熟字訓を含む，複数文字列に対応するルビ情報（以下「熟字ルビ」と呼ぶ）を調査し，その一部について出現の実態を報告する。その上で，先に挙げたコーパスへの形態論情報付与における問題点に対する対処法を示す。

２．調査概要

２．１　コーパス

調査には，2005 年に公開された『太陽コーパス』CD-ROM を用いた。

『太陽コーパス』は，言文一致を経て口語体による書き言葉が安定し普及する時期（明治時代後期～大正時代）の書き言葉を代表できるコーパスとして作られたものであり，月刊総合雑誌『太陽』（博文館）の明治 28（1895）年，明治 34（1901）年，明治 42（1909）年，大正 6（1917）年，大正 14（1925）年について，著作権処理ができなかった記事を除くほぼ全文を対象にしたものである。

雑誌『太陽』は，分量の多さ，ジャンルの広さ，執筆陣の多彩さ，読者層の厚さなどの点で，当時の文献資料として格別の価値を持っていることが指摘されており（田中 2005），それを裏付けるように，2005 年の『太陽コーパス』公開以来，口語化，濁点や仮名遣い等の表記法の確立・統一化，新漢語の定着など，当時の語用の実態や現代語に連なる変化の過程を明らかにした論考が多く発表された。

『太陽コーパス』の収録記事が発行された明治～大正期は，ルビが活発に使用されていた（田中 2005）ことに加え，現在のように表記の規範意識が高くなく，語と表記の関係が多様であったことが知られている(今野 2012)。本研究で取り上げる漢字表記語に対して与えられた様々な「熟字ルビ」も，そのような明治～大正期の表記の多様性を映す現象の一つであり，『太陽コーパス』はこの現象を観察するのに有用な調査対象資料である。

２．２　調査対象表現

本研究で対象とする「熟字ルビ」の例を抽出するために，XML タグを利用した。

『太陽コーパス』は，XML 形式により様々な情報が付加されているが，その中に，本文行の文字に与えられた振り仮名（ルビ）の文字列を格納する要素「r」がある。

雑誌『太陽』における振り仮名の様相については，以下の報告がある（田中 2005）。

- 1910 年頃までは，記事の属する欄によって (1)総ルビ（ほとんどすべての漢字に振り仮名があるもの），(2)パラルビ（一部の漢字に振り仮名があるもの），(3)無ルビ（特別な場合を除き振り仮名がないもの）かが決まっている。
- 1910 年代からは，ほとんどすべての記事で総ルビになっている。
- 振り仮名を付けるか否かは著者の判断だけでなく編集部の判断も大きかったのではないかと推測されるが，必ずしも読みにくさや誤読の可能性などによるとは言いがたい。

・　全巻を通して常に総ルビであるのは，小説・戯曲ジャンルの記事である。

これを踏まえて，『太陽コーパス』においては，誌面に見られるルビについて，

・　小説・戯曲・詩歌の類の記事[1]については，原文にある振り仮名を入力する。

・　上記以外の記事については，原文に振り仮名があっても入力しない。

という方針に基づき，「r」タグに格納している。この結果，『太陽コーパス』全 3409 記事のうち，1 割程度に該当する 364 記事に対してルビが付与されている。

なお，「r」タグは，原則として漢字 1 字 1 字にルビを対応させる形（モノルビ）で付与されるが，熟字訓など複数文字連続にしか対応させられないルビについては，語の範囲で「r」タグを付与している（グループルビ）。今回は，この複数文字に対応付けられたルビを「熟字ルビ」と呼び，これらを調査対象表現とした。このようなルールで付された熟字ルビは，単なる字と音の結び付きではなく，語と語の結びつきを示す表現であるはずであり，同一表記を介した異語の関係性や，表記と語の結びつきの強さなどを観察するのに好対象と言える。

２．３　調査方法

XML 文書から XPath を用いて調査対象表現を抽出した。XPath は，XML 文書内で特定のノードの位置を指定することを目的とした構文である。コンピュータにおけるファイル・システムパスと同様に，スラッシュ区切りで階層を明示した式（ロケーション・パス）を使ってノードを参照することができ，比較的容易に XML 文書内の特定の要素や属性を抽出することができる。今回は，スクリプト言語 perl を用い，XPath 式「//r」を元に 2 文字以上の「r」要素（本文行文字列）と「rt」属性値（ルビ文字列）の対を抽出した[2]。

これによって抽出されるのは，『太陽コーパス』において，2 文字以上の本文文字列を範囲として r タグが付されているものすべてであり，延べ 26,165 件の表記とルビの対（ルビは語形を示していると考えられるので，以下「表記-語対」と呼ぶ）が抽出された。これらは，記事の異なりで 12,800 程度に，更に表記揺れ等を統一し，異なり 4,900 程度の表記-語対に集約した。

この表記-語対には，「茶-間／ちゃの-ま」のように，1 文字ずつ分割して r タグを付すことが可能なものや，「中合／なからい」のように音変化を伴うために分割はしがたいけれども，単漢字とそれに対応する読みが照合できるようなものも含まれる。今回の分析においては，語と語の結びつきを観察するという調査目的から外れるため対象から外した。同様に，表記事情が特殊で扱いに注意が必要となる固有名や動植物名なども，今回の試験的分析には不適切と判断し，対象から外した。その結果，分析対象として抽出されたのは，約 4,700 個の表記-語対である。

３．調査結果と分析

３．１　語種とルビのタイプ

調査対象として得られた表記-語対について，本文行文字列の語種，ルビ文字列の語種，ルビの表現タイプを分類した。分類結果を表 1 に示す。なお，語種の判断については，山田俊雄他編『新潮国語辞典　第二版』（新潮社）を参照した。

---

[1] 「記事」タグ内「ジャンル」属性に格納された NDC 番号のみを規準として対象記事を定めている。

[2] 作業環境：Windows8，Cygwin(1.7.17)，perl(5.14.2)＋XPath モジュール

語種：漢語（漢文由来表現を含む），和語，和語・漢語（対応する音読み・訓読みが両方あるもの），外来語（外国語含む），混種語，宛字（対応する漢語がないもの）
ルビの表現タイプ：表音的ルビ，表意的ルビ

表１　本文行文字列およびルビ文字列の語種別表記対数

		ルビ語種				総計
		漢語	和語	混種語	外国語	
表意ルビ		170	4200	7	259	4636
本文語種	漢語	160	3800	6	205	4171
	和語	3	290	1	26	320
	和語・漢語	2	42		3	47
	混種語		10		1	11
	宛字	5	58		24	87
表音ルビ			39		14	53
本文語種	和語		1			1
	宛字		38		14	52
総計		170	4239	7	273	4689

熟字ルビの大半は表意的なルビである。中でも，漢語の本文文字列と和語のルビ文字列を対応付ける表記対が圧倒的に多く，いわゆる熟字訓もここに含まれる。ついで，漢語に対して外国語，漢語に対して別の漢語，和語に対して別の和語を対応付ける表記対が多く見られる。それぞれ，以下に例を示す。『太陽コーパス』の熟字ルビには，語種の異なる語の結びつきのみならず，同一語種の別語を結びつける例が多いことが注目される。

漢語×和語：今日／きょう，先刻／さっき，住居／すまい
漢語×外語：硝子／ガラス，煙管／キセル，卓子／テーブル
漢語×漢語：真実／ほんとう，平常／ふだん，容貌／きりょう
和語×和語：左手／ゆんで，右手／めて，最早／もう

３．２　表記対の出現記事数

次に，出現記事数を計測した。度数分布を表２に示す。

表２　出現記事数と表記の分布

出現記事数	表記-語対数	累積%
1	3397	72%
2-3	742	87%
4-5	211	92%
6-7	117	94%
8-9	54	96%
10-141	192	100%

7 割程度が 1 記事のみに出現する表記-語対である。以下に例を示す。

出現記事数１の例：高利貸／アイスクリーム，争論／いさかい，同行／いっしょ

これらは，一時的な表記と語（読み）の結びつきである可能性もあるが，ここで示したのは単純な出現数であるため，当該の表記・語自体の出現可能性の差による部分が大きく，一概に定着していない・通用的でないとは言いがたい。中には辞書に掲載される表記-語対も含まれる。

一方，複数の記事に現れる表記-語対の場合は，この時代においてはある程度定着した通用的な表記-語対と言える。表３に出現記事数の多い順に表記-語対を挙げる（参考として，ルビ付与対象記事 364 記事に対する出現率を合わせて示す）。

「今日／きょう」のようないわゆる「熟字訓」のほか，「先刻／さっき」「四辺／あたり」といった現代では一般的と思われない表記-語対が上位に見られる点は注目される。

表３　出現記事数　上位 30 語

表記-語対	本文行語種	ルビ語種	ルビタイプ	出現記事数	出現率
一人/ひとり	漢語	和語	表意	141	39%
如何/いかん	漢語	和語	表意	134	37%
今日/きょう	漢語	和語	表意	120	33%
二人/ふたり	漢語	和語	表意	120	33%
何處/どこ	和語	和語	表意	115	32%
此方/こちら	和語	和語	表意	106	29%
何時/いつ	和語	和語	表意	92	25%
此處/ここ	和語	和語	表意	89	24%
其處/そこ	和語	和語	表意	85	23%
一寸/ちょっと	漢語	和語	表意	83	23%
可愛/かわいい	宛字	和語	表音	73	20%
何故/なぜ	和語	和語	表意	68	19%
昨日/きのう	漢語	和語	表意	64	18%
田舎/いなか	漢語	和語	表意	62	17%
明日/あす	漢語	和語	表意	57	16%
今朝/けさ	漢語	和語	表意	57	16%
這入/はいる	和語	和語	表意	54	15%
不可/いけない	漢語	和語	表意	51	14%
可笑/おかしい	宛字	和語	表意	51	14%
心地/ここち	漢語	和語	表意	50	14%
屹底/きっと	宛字	和語	表音	49	13%
先刻/さっき	漢語	和語	表意	47	13%
身體/からだ	漢語	和語	表意	46	13%
煙草/タバコ	漢語	外国語	表意	46	13%
四邊/あたり	漢語	和語	表意	44	12%
今年/ことし	漢語	和語	表意	44	12%
眞面目/まじめ	漢語	和語	表意	44	12%
眞似/まね	宛字	和語	表意	43	12%
甲斐/かい	宛字	和語	表意	42	12%
住居/すまい	和語・漢語	和語	表意	41	11%

３．３　文体別に見た表記に対する読みの出現割合

次に，　出現記事数の多い表記-語対について，ルビ付与対象記事における本文文字列と対応付けられる語（読み）の出現実態を追加調査した。調査は，本文文字列の表記をターゲットとして，XML 文書内を文字列検索し，そこに付加されたルビ情報を元に読みを確定させるという方法で行った。分析に際しては，『太陽コーパス』の記事に付与された文体情報（口語・文語）を用いた。以下に調査結果の一部を示す。

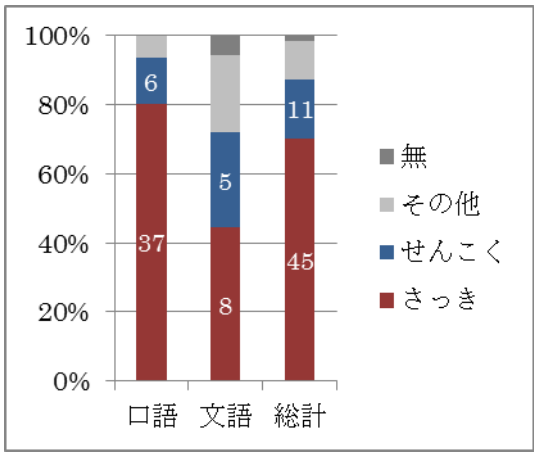

図１　本文表記「先刻」の読み

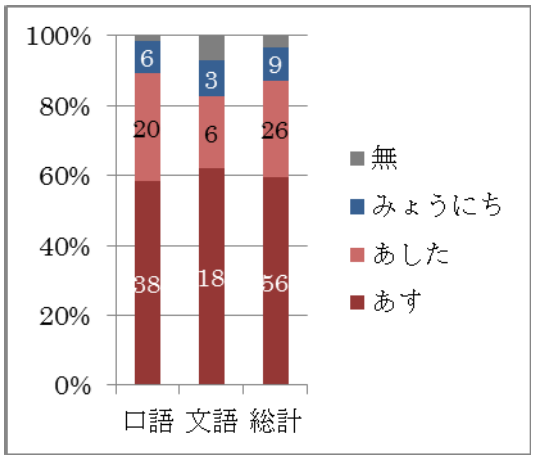

図２　本文表記「明日」の読み

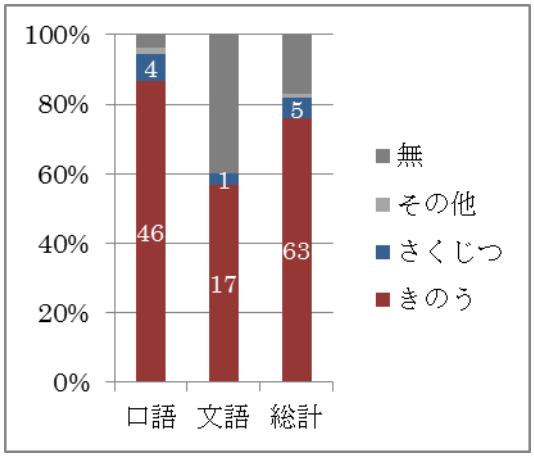

図３　本文表記「昨日」の読み

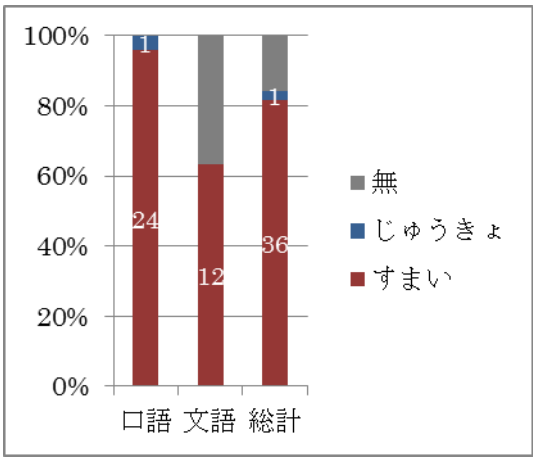

図４　本文表記「住居」の読み

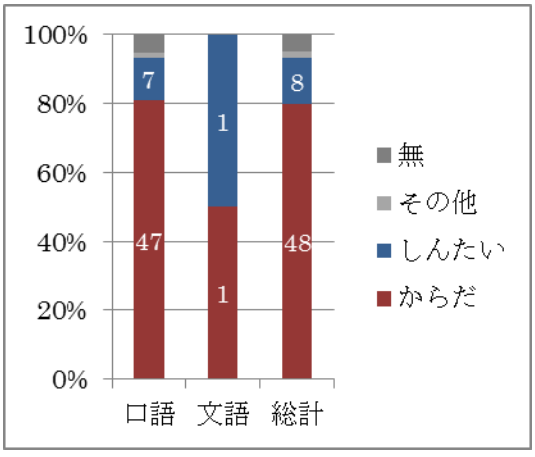

図５　本文表記「身体」の読み

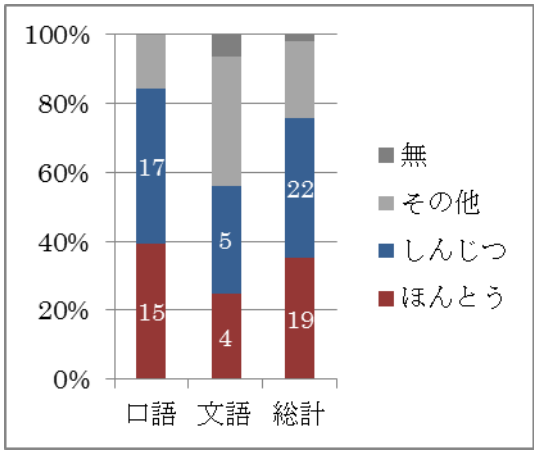

図６　本文表記「真実」の読み

図１の「先刻」を見ると，口語記事においては，圧倒的に和語「さっき」の語の表層形として用いられることの多いことが見て取れる。文語記事においても，「さっき」の割合が多いが，口語記事に比して字面通りに読む漢語「せんこく」の割合が多くなっている。

一方，図２の「明日」では，口語記事と文語記事でほぼ分布差異はなく，割合は，和語「あす」＞和語「あした」＞漢語「みょうにち」となっている。

図３，４の「昨日」「住居」は，いずれも口語記事で，ほぼ和語「きのう」「すまい」の表層形として用いられているが，文語記事では和語形に加えてルビのない形が多く見られる点に特徴がある。ルビのない形が漢語形なのか和語形なのかは不明である。

図５「身体」は，文語記事の数が少なく分析に用いることはできないが，口語記事において「からだ」と強力に結びついていることは疑いがない。

上５例については，文体差の見られる語もあるものの，元来の漢語形の読みを和語形の読みが大きく凌いでいるが，この中で現代語として一般的に「熟字訓」として認識されているのは，「明日／あす・あした」「昨日／きのう」であり，「先刻／さっき」「住居／すまい」「身体／からだ」といった表記と語（読み）の結びつきは強くない[3]。

図６「真実」は，同語種の別語との結びつきの例だが，口語記事で「ほんとう」「しんじつ」が同数ずつ，文語記事では「ほんとう」が「しんじつ」の倍以上を占め，「真実」と「ほんとう」の結びつきの強さが見て取れる。これも，現代においては希薄な結びつきの例である。

４．表記形にどの語を割り当てるべきか―形態論情報付与における指針

既存の形態論情報付き近代語コーパスである『明六雑誌コーパス』では，その形態論情報付与に関して，読みを与えるルールを以下のように定める（須永・近藤 2012）。

原則１　音読み・特に漢音を優先する。

原則４　原文にルビがあっても、その読みに従わない場合がある。

[1] 漢語に対し、外来語のルビがついている場合

[2] 原文のルビの読みが、熟字訓として定着しているとは言い難いもの

しかし，３節に示した調査結果より，少なくとも『太陽』の文芸記事においては，漢語に対して記事の文体や現代語の実態・内省による表記と語の結び付きを手掛かりに語を割当てることが不適切であることが分かった。

そこで，今後，『太陽コーパス』における形態論情報付与作業においては，今回の調査を元に，通用的と判断できる表記-語対については，以下のような方針で語の認定を行うことを検討したい。

(a) ルビに基づき語（読み）を認定する

(b) 表記に対する読みの分布に偏りがあるものについては，その情報に基づいて，ルビのない表記に熟字ルビに相当する読みを与える

５．まとめ

本発表では，『太陽コーパス』の XML タグを用いて，複数文字に対応するルビの抽出を試み，以下の調査報告を行なった。

[3] 例えば，BCCWJ コアでは「先刻-さっき」「住居-すまい」の表記-語対はなし，「身体-からだ」については書字形「身体」440 例中「からだ（体）」の語彙素認定は 71 例（16%）に過ぎない。

・ 本文表記およびルビの語種分類とそのクロス分布
・ 出現記事数による本文表記-ルビ対の定着度
・ 本文表記を元にした同表記異語の出現分布実態

また，調査結果を用いての，「時代的な表記と語の結びつき」を考慮した形態論情報付与を提案した。今回の手法あるいは時間的な制約によって残された問題点のいくつかを以下に示す。今後の課題としたい。

・ 文芸以外の記事についての実態把握

『太陽コーパス』の仕様上，文芸作品以外の記事にはルビが付与されておらず，それらの記事については実態の把握ができなかった。ルビの様態は多分にジャンルに依存するものであり，特に文学作品ではそれが顕著であることが予想される。更に，新聞等公共性の高いテキストにおけるルビの実態等とも合わせて考えるべきものである。

・ 著者別，年別の詳細な調査

テキストの表記は，著者の個人的な志向の強く現れる部分であるため，本来は，今回行った記事単位ではなく著者単位で集計を行うべきものである。また，今回指摘した表記と語（読み）の結びつきの定着度が現代語と『太陽』とで異なる点については，『太陽』の内部でも経年的な変化が見られる可能性もある。現代語との比較も合わせて今後取り組むべき課題と位置づける。

付記

本研究は，国立国語研究所共同研究プロジェクト「通時コーパスの設計」（プロジェクトリーダー：近藤泰弘）による成果の一部です。

コーパス

『太陽コーパス』: 国立国語研究所(2005a)『太陽コーパス―雑誌『太陽』日本語データベース―』（CD-ROM）博文館新社

『BCCWJ』: コーパス検索アプリケーション中納言（https://chunagon.ninjal.ac.jp/）

参考文献

今野真二(2012a)『百年前の日本語―書きことばが揺れた時代 (岩波新書)』岩波書店

今野真二(2012b)『ボール表紙本と明治の日本語』港の人

京極興一(1998)『近代日本後の研究―表記と表現―』東宛社

国立国語研究所(2005b)『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―』博文館新社

須永哲矢・近藤明日子(2012)「近代語コーパスのための形態論情報付与規程の整備」田中牧郎・岡島昭浩・小木曽智信・小野正弘・小島聡子・島田泰子・朱京偉・高田智和・張元哉・陳力衛・近藤明日子・須永哲矢『近代語コーパス設計のための文献言語研究 成果報告書 （国立国語研究所共同研究報告 12-03）』国立国語研究所、pp.93-117

武部良明(1981)『日本語表記法の課題』三省堂

田島優(1998)『近代漢字表記語の研究』和泉書院

田中牧郎(2005)「言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計」国立国語研究所(2005b)、pp.1-48

伝康晴・中村純平・小木曽智信・小椋秀樹(2008)「語種情報を用いた同表記異音語の解消」言語処理学会年次大会発表論文集 14、pp.69-72