発話文への発話者情報付与の基本設計 ： 『現代日本語書き言葉均衡コーパス』収録の小説を対象に

著者	宮嵜 由美, 柏野 和佳子, 山崎 誠
雑誌名	言語資源活用ワークショップ発表論文集
巻	1
ページ	38-48
発行年	2017
URL	http://doi.org/10.15084/00001456

発話文への発話者情報付与の基本設計
-『現代日本語書き言葉均衡コーパス』収録の小説を対象に-

宮嵜由美（国立国語研究所音声言語研究領域）†
柏野和佳子（国立国語研究所音声言語研究領域）
山崎誠（国立国語研究所言語変化研究領域）

Fundamental Planning of Annotation of Speaker's Information to Utterances :Focused on Novels in "Balanced Corpus of Contemporary Written Japanese"

Yumi Miyazaki (National Institute for Japanese Language and Linguistics)
Wakako Kashino (National Institute for Japanese Language and Linguistics)
Makoto Yamazaki (National Institute for Japanese Language and Linguistics)

要旨

現在，国立国語研究所音声言語研究領域では，『日本語日常会話コーパス』（以下，CEJC）の開発が行われている。多様な話し言葉の会話行動の収録を目指す上記プロジェクトの理念と同様，本プロジェクトの目指す，書き言葉における会話場面の「発話」への発話者情報付与も重要な“日本語の会話”の一端を担うものである。

すでに公開されている『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）の約 6 割を占める書籍のサンプルには，会話場面における大量の発話文が存在する。発話文は地の文とは言語的に異なる特徴を持つことが多いため，分析に当たっては別に扱うことが妥当であるが，現在の検索環境では難しい。

そこで，本稿では，BCCWJ 収録の小説を対象に，小説特有ともいえる発話部分特定の問題点（かぎ括弧で括られない例や非現実場面での発話など）を提示する。機械抽出のみでは同定の難しい発話箇所と発話者情報付与について，その基本設計の「発話認定箇所」基準を中心に提案する。

１．はじめに

現在，前述の CEJC や国語研究所『日本語歴史コーパス』には発話者情報が付与されているものの，BCCWJ 収録の会話文には発話者情報が付与されていない。この現代の書き言葉を収録する BCCWJ の会話文にも発話者情報が付与されれば，より深い分析や他のコーパスとの比較にも寄与できるものと考える。

そこでまず，本稿では BCCWJ 収録の小説・物語への発話者属性情報を付与するにあたり，どのように発話箇所を認定していくかを問題とする。なぜなら，実際に作業をしてみると，小説という書き言葉媒体では作者個別の文体的特徴が多くみられ，会話場面における“声

†

に出したと想定される発話”の認定にもかなりの困難が生じるためである。

例えば，発話箇所を示すことの多いカギ括弧を頼りに機械的に抽出する方法をとった場合，「釣りぼり」など看板を示す文字列も抽出され，分析対象外となる箇所も少なくない。逆に，カギ括弧で括られない場合にも，声に出したと想定される発話が多数存在し，小説の会話場面における発話の姿が十分に反映されないのが現状である。

そこで本稿では，発話箇所認定の原則としてまず，「A.発話が一重カギ括弧（以下，カギ括弧）に囲まれているかどうか」，「B.声に出したと想定される発話であるかどうか」を頼りに，以下５つの基準の提案を行う。

➢発話箇所認定の基本基準

１）カギ括弧に括られた声に出したと想定される部分

２）カギ括弧に括られた１）に準ずる部分

３）カギ括弧に括られた当該の文字列の強調などを示す部分 …<非発話>

４）カギ括弧に括られない声に出したと想定される部分

５）「場面設定」を考慮した１）に準ずる部分

上記基準に従い，具体的にどのような会話場面と，そこにどのような形式で表現される発話のバリエーションが生じているのか，発話者情報や発話状況の属性付与の概要とともに報告する。

２．作業対象

2.1 BCCWJ における「発話箇所」の収録状況

表 1 BCCWJ における発話文の割合

レジスター	サンプル数	<speech>タグを含むサンプル数	<quote>タグを含むサンプル数	発話箇所を含むサンプル数	発話箇所を含むサンプル数の割合（%）
図書館書籍（LB）	10,551	5,105	8,978	9,987	94.65
ベストセラー（OB）	1,390	917	1,080	1,321	95.04
Yahoo!知恵袋（OC）	91,445	0	0	0	0.00
法律（OL）	346	0	308	308	89.02
国会会議録（OM）	159	159	122	159	100.00
広報紙（OP）	354	244	354	354	100.00
教科書（OT）	412	0	0	0	0.00
韻文（OV）	252	0	68	68	26.98
白書（OW）	1,500	0	1,352	1,352	90.13
Yahoo!ブログ（OY）	52,680	0	0	0	0.00
出版書籍（PB）	10,117	3,479	8,646	9,250	91.43
出版雑誌（PM）	1,996	844	1,787	1,844	92.38
出版新聞（PN）	1,473	199	1,455	1,457	98.91
合計	172,675	10,947	24,150	26,100	15.12

本プロジェクトで対象とする BCCWJ には，表１，レジスター欄に示す日本語の「書き言葉」のデータが収録されている。

さらにデータには，「カギ括弧」で括られた箇所に<speech>あるいは<quote>によってタグ付けが施されている。まず，この２つのタグを暫定的な発話箇所[1]とみなし，集計したものが表１である。

この<speech>もしくは<quote>タグにより，多くの発話部分を機械的に抽出することが可能である。本プロジェクトではその出現箇所の多い，図書館書籍，出版書籍，ベストセラーを対象に，さらに NDC 番号によって分類される 913 番台「文学：日本文学：小説，物語」を作業対象の出発点とした。この，NDC913 番台の<speech>もしくは<quote>によって括られた暫定的な発話箇所はおおよそ 23 万箇所に及ぶ。

３．「発話認定箇所」と「発話者情報」

3.1 発話認定箇所と具体的データ例

前述の通り，本プロジェクトで認定する基本的な発話箇所とは，原則として「A.カギ括弧で括られた」「B．声に出したと想定される発話（以下，声に出した発話)」を指す。

ただし，対象とする小説や物語によっては，場面の流れや作家個別の文体など，例１に示す二重下線部（以下，下線部）のような，声に出した発話が必ずしもカギ括弧で括られていない場合が多数ある。

例１（ｻﾝﾌﾟﾙ ID：LBp9_00190）
<speech>[2]
<paragraph>
<superSentence><sentence>Ⅰ[3]「神林家は、わしと東吾と二人だけの兄弟である。
</sentence>
<sentence>Ⅱ東吾の同意なくば、この話は成り立たぬのだ」</sentence>
</superSentence><br type="automatic_original" />
</paragraph>
</speech>
</quotation>
<paragraph>
<sentence> Ⅲどうじゃ、承知してくれるか、と重ねて通之進がいい、東吾は畳に手を突いて、深く頭を下げた。</sentence>

【出典】平岩弓枝（2001）「春の高瀬舟」文藝春秋

[1] <quote>タグは１発話内における<speech>の内側に括られる場合があり，必ずしも<speech>タグから独立した発話文とはならない。さらに<speech>タグ部分が，必ずしも発話箇所であるとは限らない。その詳細と具体例は「４.非発話認定箇所」に示す通り。

[2] 抽出個所の多くの前後には，例１に示したような<speech><paragraph>(<superSentence>)<sentence>のタグが付与される。本稿ではスペースの都合上<sentence>タグ以降を例として提示する。

[3] 発話と認定した箇所の冒頭に付与したローマ数字は 3.2.1 に示す「図１:属性付与の作業例」と対応するものであり，暫定的に筆者が付与したものである。

この下線部が，カギ括弧で括られていないものの，声に出した発話と認定できる根拠は，同文中に「と重ねて通之進がいい」と声に出した発話を意味する動詞が付与されている点にある。このような出現例への発話者情報付与例は 3.2.1 や 5 で詳しく述べる。

カギ括弧に拠らない声に出した発話の認定には，発話部分の認定が作業者による恣意的なものであってはならない点を十分に考慮する必要がある。しかし,「人間」が何を頼りに，どのような箇所を「発話」と認定するのかという認知過程のデータの蓄積も兼ね，機械抽出だけでは同定の難しい発話箇所の認定について以下，具体例とともに検討していく。

3.2　カギ括弧に括られた声に出したと想定される発話

3.2.1 発話認定箇所とそこに付与される属性

まず，A.カギ括弧で括られ，B.声に出した発話と判断される発話認定箇所について，必ず，話者が特定できる〈話者名〉を付与する。その他，発話と認定した箇所にどのような発話者情報が付与されているか，その内容の概略を表２に，データ入力の具体例を図１に示す。

表２　発話者情報の概略

		発話者属性	内容(概略)
①	話者ID	話者名	小説内での「発話者」の呼び名
		性別	男／女／その他／不明
		年代	若年層（~19 歳）／成年層（20 歳～59 歳）／老年層（60 歳以上）ただし，６歳以下は幼年とし，若年層を選択の上，備考欄に記載
		年代の確信レベル	書籍内に記載がない場合に「？」を付与
		非人間	ファンタジー小説，ホラー小説などに登場する人間以外の話者に「○」を付与
		会話モード	方言／外国人との会話／日本語以外での会話 通話／テレパシー／声に出した引用／独話／沈黙　など
		会話認定情報	カギ括弧がないが，声に出した発話である場合／非発話（看板，メモ，語の強調等）／心内発話　など
		備考	上記属性の補足情報
②		職業	書籍内で記載のある場合に付与
		相手	誰に対する発話かを小説内の話者名を用い付与

現在の作業段階として，まず，表２①部分の話者 ID 情報付与作業が行われており，②については，筆者が作業対象の一部のデータ（現在 100 サンプル程度）に情報付与を行っている。

例１の会話例に，表２①部分の属性を付与した作業例を図１に示す。原著では同一話者による改行が挿入されないひとつのカギ括弧内の発話であっても，作業ファイルでは，図１Ⅰ，Ⅱのように<sentence>タグを境に新たに情報付与行が設けられ，その行ごとに話者

ID を付与していく。例１の場合，具体的には〈話者名〉神林通之進，〈性別〉男，〈年代〉成年層，の話者 ID 情報がⅠ，Ⅱ，Ⅲにそれぞれ付与される。

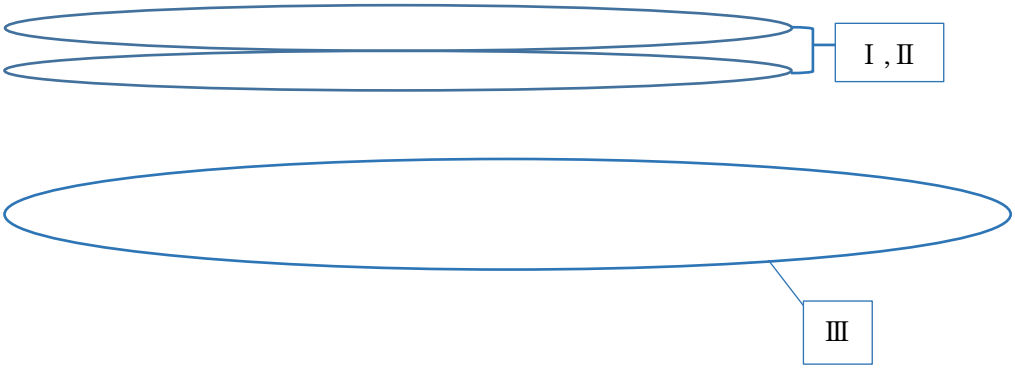

図１　属性付与作業例

ただし，Ⅲのような，カギ括弧では括られていないものの，声に出した発話として認定されるものには，〈原文〉，〈会話認定情報〉，〈備考〉欄にその旨を入力する。詳細は「５.カギ括弧が付与されていない声に出した発話」に提示する。

3.3 カギ括弧に括られたその他の発話

その他，上述 3.1 の原則に準ずるものとして，A.カギ括弧が付与され，B.声に出してはいるものの，発話者は聞き手を意識しないと想定される「独話」や，会話場面においてA.カギ括弧が付与されているが，発声の伴わない「沈黙」があげられる。また，A.カギ括弧が付与されている点を考慮し作業対象とした，「心内発話」もここで取り上げる。

3.3.1 独話

例２は，アスタシュールという男が場内のアナウンスを聞き，ひとり呻く場面である。

例２(ｻﾝﾌﾟﾙ ID：LBh9_00135)

```
<sentence>　「む…」</sentence>
<br type="automatic_original" />
</paragraph>
</speech>
</quotation>
<paragraph>
<sentence>　すでに真昼の陽光が射す丘の上である。</sentence>
<br type="automatic_original" />
</paragraph>
```

<paragraph>
<sentence>　アスタシュールが軽い呻きをあげて立ち止まった。</sentence>
【出典】伏見健二（1993）「叛逆の獣将」中央公論社

この独話を受け，それを耳にした別の人物が「どうしたんだよ」と，カギ括弧で括られた声に出した発話が続くことから，当該下線部分もカギ括弧に括られた声に出した発話と認定した。この場合，話者 ID 情報のほか，聞き手を意識しない発話として〈会話モード〉に「独話」と記入する。

3.3.2 沈黙

「沈黙」は発言の裏返しの行為ではない。意見に対する反感や内容の吟味などその機能はさまざまである。本プロジェクトでは，沈黙は小説における会話場面内で，ある一定の意味をもつ発話の一部として抽出し，発話者情報を付与している。例３の下線部がその具体例である。

まず，「沈黙」を意味するカギ括弧で括られた三点リイダの場合であるが，①発話中のいわゆる“言いよどみ”を表す「沈黙パターン(a)」と，②カッコ内が沈黙のみで表される「沈黙パターン(b)」との２つに分類できるものと考える。

例３(ｻﾝﾌﾟﾙ ID：LBq9_00101)
<superSentence><sentence>　「…どないしたん、由香ちゃん。</sentence>
<sentence>泣いたら疲れたか？」</sentence>
</superSentence><br type="automatic_original" />
</paragraph>
</speech>
</quotation>
<quotation>
<speech>
<paragraph>
<sentence>　「…」</sentence> >

沈黙パターン(a)

沈黙パターン(b)

【出典】佐藤ケイ（2002）「Last kiss」メディアワークス;角川書店

「沈黙パターン(a)」の場合，発話開始の際のいわゆる言いよどみとして，開始括弧からクオーテションマーク＋終了括弧までが発話箇所と認定され，通常の発話と同様に話者 ID 情報が付与される。

「沈黙パターン(b)」の場合，(a)と同様，話者 ID 情報が付与され，さらに〈会話モード〉に「沈黙」，〈会話認定情報〉に「保留[4]」が付与される。

3.3.3 心内発話

カギ括弧が付与されているものの，声に出していない発話として，心内発話がある。例

4 「保留」とは，カギ括弧で括られてはいるものの，声に出して発話されていないものに付与される。バリエーションについては別稿にあらためたい。

４がその具体例である。これも，声に出すことで聞き手の反応を想定するものではないが，カギ括弧が付与されることを考慮し，3.1.A，B で示した原則に準ずる発話とし発話者情報付与の対象とした。

例４(ｻﾝﾌﾟﾙ ID PB29_00066)
<sentence>　<squote_A>「あれはきっと悪い夢を見たのだわ」</squote_A>と彼女は自分に言い聞かせた。</sentence>
<sentence>夢の記憶を両親に確かめるのが、なんとなく憚られた。</sentence>
【出典】森村誠一（1989）「黒魔術の女」光文社

この場合，“自分に言い聞かせた”との心内発話を意味する名詞や動詞を抽出対象の根拠とし，話者 ID 情報のほか，＜会話認定情報＞に「心内発話」と付与される。

４．非発話認定箇所

カギ括弧に括られた文字列ではあるが，当該の文字列の強調などを示すものであり，発話と認定されないものとして，例５に示す抽出例がある。

例５ （ｻﾝﾌﾟﾙ ID: LBp9_00237）
<sentence><pquote_1>「客のめし」</pquote_1>の味も、食糧の豊かな時代にあっては<pquote_2>「豚がわり」</pquote_2>に動員された屈辱を救いきれない。</sentence>
【出典】森村誠一（2001）「鍵のかかる棺　下」徳間書店

これらカギ括弧で括られた下線部は，文中における当該の語の強調を示すものであり，発話とはみなさない。このような発話として認定されないカギ括弧のデータには，話者 ID 情報を付与せず，〈会話認定情報〉を「保留」とし，発話と認定しない根拠を〈備考〉に記す。その他，声に出して読まれていない，看板，手紙やメモなどもこれにあたる。

５．カギ括弧が付与されていない声に出した発話

カギ括弧は付与されていないが，声に出した発話と判断される場合として，例１同様，例６の下線部がある。

例６ （ｻﾝﾌﾟﾙ ID: LBp9_00203）
<sentence>[それにしても…]と、蘭の方は深い安堵の吐息とともに言った。</sentence>
【出典】岩崎正吾（2001）「遥かな武田騎馬隊」角川春樹事務所

この場合，同文中にある“言った”との声に出した発話を意味する動詞を，声に出した発話と認定する根拠とする。この場合話者 ID 情報は，カギ括弧のある発話と同様に付与され，〈会話認定情報〉に「タグなし」を付与する。さらに，作業データの〈原文〉の当該発話部分を，新たにブラケット[ ]で括る。〈備考〉には，発話認定に至る根拠を示す。例えばこの場合「言った。とある」と記す。ブラケットの付与範囲は，各作業者が，地の文との境界

と判断する箇所までとする。

次にあげる例７もカギ括弧は付与されていないが，声に出した発話の例である。この例は，発話と地の文との境界を，従来の記号とその機能を頼りに機械的に抽出するには困難な例でもある。

この場合，まず，同文中にある“返事をした”との声に出した発話を意味する動詞から発話情報付与対象と判断される。また，この場合どこを地の文との境界とするかであるが，下線部ハイフンが三点リイダに相当する機能を持って付与されているものとし，その直後までをブラケットで括る。

例７（ｻﾝﾌﾟﾙID：PB59_00081）

：<sentence>　[ふうん―]と、中禅寺は感心したような馬鹿にしたような返事をした。</sentence>

<sentence>それから徐に横に視線を送って、電柱に凭れかかっていた風采の上がらない男に向けてこう云った。</sentence>

【出典】京極夏彦（2005）「百器徒然袋-雨」講談社

今日の書き言葉，打ち言葉[5]では，例えば「すみません。。」，「ありがとうございます、、」など，句読点等が，言いよどみや感情表現として使用される現象が多くみられ，本プロジェクトでも人的判断により地の文との区別を行っている。必要があれば，〈備考〉にその旨を記載する。この発話文と地の文の境界をどこに見出すか，その根拠の蓄積は，発話箇所認定に関わる新たな提案ができるものと考える。

６．場面設定による発話表示　－非現実場面での発話－

小説・物語特有の場面設定として，非現実場面があげられる。具体的には，夢の中での会話場面や，ＳＦ小説などでのロボットの会話場面，ファンタジー小説のテレパシーを使った会話場面など，まさに多種多様な場面設定がある。その多種多様性が，小説や物語という書き言葉媒体の醍醐味ともいえよう。

これらの場面設定における発話も，「声に出した発話」に準ずる「発話」と認定し，どのような場面設定での発話であったかを判別可能な状態とする情報を付与している。

6.1「夢の中」での会話例

例８は，「夢の中」という場面内での会話場面の例である。小説という書き言葉媒体の場合，場面が転換された場合や，作家固有の文体によって，声に出したと想定される発話が必ずしもカギ括弧で括られているとは言えない例のひとつとでもある。

例８作品の原著では，「夢の中」という場面での主人公の声に出した発話にはカギ括弧が，主人公以外の声に出した発話には二重カギ括弧が付与されるといった規則性がみられる。この二重カギ括弧部分は，例８の最終行にある「声」という声に出した発話を意味する名詞から，「夢の中」という場面設定における声に出した発話と認定される。

[5] 携帯メールや無料通信アプリケーション「LINE」でのやり取りでは，既に句読点は感情を表す表現として機能の拡張がみられる。

例８（ｻﾝﾌﾟﾙ ID: LBh9_00122）
<sentence>　『拓ちゃん、ごめんなさいね、駄目なのよ。</sentence>
<sentence>あたし達のせいなの』</sentence>
</superSentence><br type="automatic_original" />
</paragraph>
</speech>
</quotation>
<paragraph>
<sentence>　あ。</sentence>
<sentence>この声は、夢ちゃんのママだ。</sentence>
【出典】新井素子（1993）「緑幻想」講談社

この場合，まず話者 ID 情報を付与し，さらに〈会話モード〉に「夢の中」が，〈会話認定情報〉に「保留」が付与される。

6.2「ファンタジー小説」－テレパシーによる会話例－

小説や物語の場面設定によっては，直接的な発声は伴わないものの，カギ括弧で会話が繰り広げられる場合がある。例９はファンタジー小説において，「意識の中」「声が響いてきた」「語りかける」など，発声は伴わないが，声に出したものと同等とされる発話を意味する名詞や動詞群から，テレパシーによる会話であることを示す例である。

例９（ｻﾝﾌﾟﾙ ID: LBd9_00046）
<sentence>　イーノゥが白い耳をピンと立てる。</sentence>
<sentence>すると、キャロルたち全員の意識の中に、マクスウェルの声が響いてきた。
</sentence>
<superSentence><quote><sentence>「聞こえるかい？</sentence>
<sentence>　私は君たちと共にいる。</sentence>
<sentence>イーノゥの目を通してすべてを見ることができるし、こうして君たちに語りかけることもできるのだ。</sentence>
<sentence>何かあった時の判断は私がしよう。</sentence>
<sentence>しかし、イーノゥの頭脳に私の意識が無理矢理入り込んでいるので、そうたくさんは話すことができない。</sentence>
<sentence>そのへんはフラッシュ、君にまかせようと思う」</sentence>
【出典】木根尚登（1989）「キャロル」CBS・ソニー出版

この場合，テレパシーの発信者に<話者名>等の話者 ID を付与する。さらに，〈会話モード〉に「テレパシー」と付与し，直接の発声を伴わない旨を〈会話認定情報〉に「保留」を記入することで，3.1 の A，B で示した原則的な「発話箇所認定」の両条件を満たすわけではないが，それに準ずるものとして検索可能となる。

７．おわりに

以上，本稿ではじめに示した発話箇所認定の基本基準１）～５）と，具体的な発話例を照らし合わせると，図２のように示すことができる。発話はＡとＢの両方の条件を備えているものが原則であるが，小説・物語という書き言葉媒体と作家の個性ともいうべき文体のあり方を考慮した上で，ＡもしくはＢの条件を備えているものを提示した。

図２　本稿で抽出した発話形態

2017 年 1 月現在，BCCWJ に収録されている小説・物語の約 2000 ファイルへの発話属性付与作業が行われている。そこには本稿で示した具体例以外にも多種多様な場面設定や話者設定があり，多種多様な発話の表現形態が存在している。作家によっては，場面転換（夢の中と現実との区別など）や発話モード（通話やテレパシーなど）の転換の演出として，カギ括弧以外の記号が使用されている例も多くみられる。

読み手である人間が，何を基準にどこまでを発話対象とするか。発話者情報属性付与作業を通し，その過程を俯瞰し整理した上で，階層的な構造を持つ日本語の書き言葉における会話行動の姿を提示していくことができると考えられる。

今後も，汎用性のある書き言葉媒体の会話場面における発話者情報コーパス構築を目指すとともに，データ整備作業を通し，人間が発話と認定する要因とその階層性についても考察していきたい。

謝　辞

本研究は国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」（代表：小磯花絵），JSPS 科研費 15H03212（代表：山崎誠），16K02714（代表：宮嵜由美）の助成を受けたものです。

また，本プロジェクトの発話者属性付与作業については，国立国語研究所技術補佐員，立花幸子さん，田嶋明日香さん，平本智弥さんにご協力いただきました。ここに感謝致し

ます。

文　献

小磯 花絵・土屋 智行・渡部 涼子・横森 大輔・相澤 正夫・伝 康晴（2016）「均衡会話コーパス設計のための一日の会話行動に関する基礎調査」国立国語研究所論集 (10), pp.85-106

小西 光・中村 壮範・田中 弥生・間淵 洋子・浅原 正幸・立花 幸子・加藤 祥・今田 水穂・山口 昌也・前川 喜久雄・小木曽 智信・山崎 誠・丸山 岳彦(2015)「『現代日本語書き言葉均衡コーパス』の文境界修正」国立国語研究所論集 (9) pp. 81-100

砂川 有里子（1988.a）「引用文の構造と機能 : 引用文の 3 つの類型について」文藝言語研究. 言語篇 13,pp. 73-91

砂川 有里子（1988.b）「引用文の構造と機能(その 2) : 引用句と名詞句をめぐって」文藝言語研究. 言語篇 14, pp.75-91

村井 源（2016）「主体語彙辞書を用いた物語テキスト中の主体推定システムに向けて」人間科学とコンピュータシンポジウム発表論集 pp.209-214

山崎 誠（2007）「『現代書き言葉均衡コーパス』の設計」特定領域研究「日本語コーパス」平成 18 年度研究成果報告書『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法―現代日本語書き言葉の文字数調査― pp.20-27

関連 URL

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