

国立国語研究所学術情報リポジトリ

日中対訳コーパスの構築と公開に向けて

メタデータ	言語: Japanese 出版者: 公開日: 2023-11-24 キーワード (Ja): キーワード (En): 作成者: 宮本, 華瑠 メールアドレス: 所属:
URL	https://doi.org/10.15084/0002000113

日中対訳コーパスの構築と公開に向けて

宮本 華瑠（大阪大学）

Toward the Construction and Publication of a Japanese-Chinese Bilingual Corpus

Haru Miyamoto (Osaka University)

要旨

昨今、公開された日中対訳コーパスには、北京日本学研究センターの『中日対訳コーパス』、情報通信研究機構の『NICT 多言語対訳コーパス』、JST・NICT 共同で構築された『アジア学術論文抜粋コーパス(ASPEC)』、『GSK 通訳データベース (JNPC コーパス) 日中・日西サブコーパス』などがあげられる。しかし、『中日対訳コーパス』は現在、個人・機関問わず対訳コーパスの入手はできなくなっている。『NICT 多言語対訳コーパス』は機械翻訳の研究またはシステム開発の一環として構築されたものでデータは非公開となっている。そして、『ASPEC』コーパスと『JNPC コーパス』は、専門用語など特殊用語の偏りが多く、広く一般的に用いられている言語使用を代表できるとは言い難い。以上のことから、日中対照研究を行う際に利用できるコーパスは極めて限定的で、言語資源が乏しい状況に陥っていることが読み取れる。発表者は個人利用を目的に2009年から対訳文の収集を続けており、その成果物を個人利用だけでなく、オープンにすべきであると考えている。収集済みデータには、雑誌『Taiwan Panorama』約45万字、『聞く中国語』2018年1月～2023年7月（67冊）のデータ約242万字、『人民網』ニュース（日中対訳文）2014年7月～2023年8月現在のデータ約272万字が含まれる。本稿では主に、日中対訳コーパスの紹介（2節）、実用に向けた活用方法の紹介（3節）、著作権問題に関する示唆（4節）、データ公開に向けての告知事項（5節）、今後の課題（6節）について述べる。

1. はじめに

本稿は日中対訳データの公開と実用に向けて、発表者が個人で収集を行った日中対訳コーパスデータを紹介するものである。発表の目的は、日中対訳言語資源の存在を広めると共に、現時点での乗り越えるべき問題点を開示し情報共有を促すためである。データは今後も収集を続け、バージョンはデータが整い次第定期的に更新を行う予定である。

言語資源の著作権に関しては法律専門家への業務委託を視野に入れており、現段階の課題は「軽微な使用」の壁を如何に乗り越えるべきかである。データの試作ではなく、収集済みデータ全文を配布するには、今後著作権法の更なる展開が鍵となってくる。

2. 日中対訳コーパスの紹介

対訳データは二つの捉え方が可能である。即ち、起点言語だけをデータとして切り取った場合は「单一言語資源」、起点言語と目標言語がセットになっている対訳文の場合は言語間の対照分析ができるパラレルコーパスデータとして捉えることができる。そして、目標言語を対象にすれば不自然性を隠れ持った言語資源として考えることができる。本稿でいう日中対訳コーパスとは、日本語を中国語に訳した場合と、中国語を日本語に訳した場合の両方を含む。本稿では便宜上、日本語が起点言語の場合は「日→中」、中国語が起点言語の場合は「中→日」と記す。

2.1 既存の日中対訳コーパス

2.1.1 北京日本学研究センターの『中日対訳コーパス』

「中日対訳コーパス」は北京日本学研究センターで 2003 年に公開したコーパスである。格納されたデータには、文学作品が約 1130 万字、作品は日中それぞれ 22 篇と 23 篇となり、文学以外は約 575 万字、日本原作が 14 篇、中国原作 14 篇、共同 2 篇となる。その全作品名を次に示す。

【日本原作：36 作品】

明日来る人、坊っちゃん、越前竹人形、布団、雁の寺、破戒、鼻、金閣寺、こころ、高野聖、黒い雨、野火、ノルウェイの森、羅生門、青春の蹉跌、飼育、死者の奢り、砂の女、徐陽、痴人の愛、友情、雪国、日本戦後名詩百家集、百言百話、ひとりっ子の上手な育て方、激動の百年史、日本経済の飛躍的な発展、心の危機管理術、近代作家入門、マッテオ・リッチ伝、日本列島改造論、日本国憲法、サラダ記念日、タテ社会の人間関係、適応の条件、五体不満足

【中国原作：39 作品】

人大報告 96、人大報告 97、人大報告 98、人大報告 99、我的父亲邓小平、我的父亲邓小平 2、邓小平文选第一卷、邓小平文选第二卷、邓小平文选第三卷、中日飞鸿、毛泽东选集第一卷、毛泽东文选集第二卷、毛泽东选集第三卷、毛泽东选集第四卷、毛泽东传、中日外交两个基本文件、插队的故事、盖棺、丹凤眼、辘轳把胡同 9 号、关于女人、活动变人形、红高粱、金光大道、家、轮椅上的梦、呐喊、彷徨、青春之歌、倾城之恋、棋王、人到中年、人啊人、上海的早晨（上）、霜叶红似二月花、天云山传奇、小鲍庄、骆驼祥子、钟鼓楼

作品名から読み取れるように、「坊っちゃん」の発行は 1906 年、芥川の「羅生門」「鼻」は 1915 年と 1916 年、川端の「雪国」は 1948 年であり、中国原作も同様 50 年～100 年前の作品が多く含まれている。そして、次の図 1 は中国原作の日本語訳で使われた語彙と「出典」との対応を明らかにしたものである。

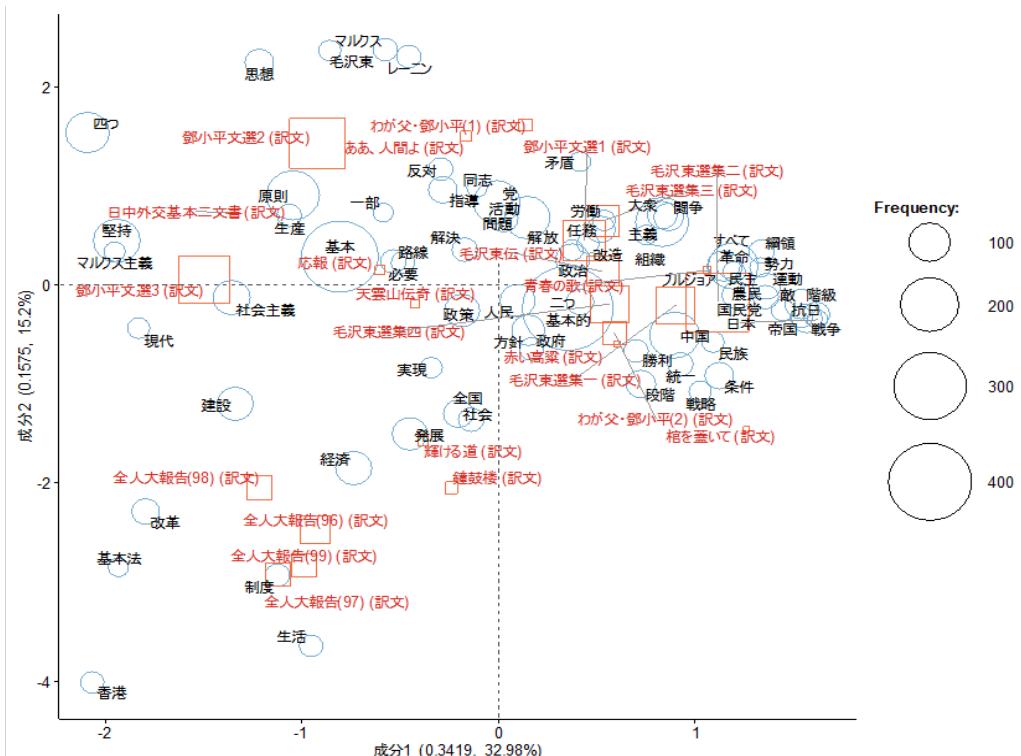


図 1 中国原作の日本語訳 語彙対応関係

上記図 1 を読み解くポイントは「原点からの距離が特徴の強さを意味する点にある。点線の交わっている場所が、横軸 0 で縦軸 0 という場所「原点」であり、原点に近い位置にある語はあまり特徴がない語になる。特徴がないというのは、外部変数の値に関係なしに、まんべんなく出現している語である(樋口他 2022;60)」とされる。また、丸いバブルの大きさは共起語の使われた回数を意味し、四角いバブルの大きさは出典の文量を意味する。

以上のことから『毛沢東選集』『毛沢東伝』『鄧小平文選』『わが父・鄧小平』『日中外交基本二文書』で用いられる特徴語として「大衆、闘争、綱領、革命、勢力、民主、敵、国民党、抗日、帝国、統一、マルクス、レーニン、毛沢東、四つ、マルクス主義、社会主義、ブルジョア、勝利、制度、戦争、中国、民族、段階、条件、戦略、思想、堅持、主義、組織、すべて、運動、農民、階級、日本」などが確認でき、「改革、香港、生活、基本法」は『全人大報告』で用いられている特徴語であることが明らかとなった。いずれにしても、中国語原文には全国人民代表大会関連が 4 作品、鄧小平論集が 5 作品、日中政治関連が 2 作品、毛沢東論集が 5 作品となり、文体的偏りが眼立っている。

『中日対訳コーパス』は現在、新規の利用申し込みができない状況である。

2.1.2 情報通信研究機構『NICT 多言語対訳コーパス』

日中対訳コーパスには他に NICT のデータがあげられる。機械翻訳のシステム開発の一環として構築されたもので、京大コーパス(1995 年の毎日新聞の記事から抜粋された約 4 万文)を日本語原文とし、その中国語訳を作成して構成された対訳コーパスである(張他 2005; 510)。格納されているデータは全て日本語を中国語に訳した対訳文(日→中)となっている。NICT の研究成果として公開された日中対訳言語資源には「日英中基本文データ」と特許に関する専門用語のデータなどは確認できるが、「NICT 日中対訳コーパス」は公式サイトでは現在未公開となっている。

表 1. NICT 日中対訳コーパスの詳細

	日本語	中国語
文	38,383	
単語	947,066	877,859
語彙	36,657	33,425
一回出現の語彙	15,036	13,238
平均文長(文字)	24.7	22.9

—張他(2008 ; 261)

2.1.3 JST・NICT 共同で構築された『アジア学術論文抜粋コーパス (ASPEC)』

「アジア学術論文抜粋コーパス (ASPEC)」(Asian Scientific Paper Excerpt Corpus) に格納されているデータは全て日本語を中国語に訳した対訳文(日→中)である。並列文からなる日英論文抄録コーパス (ASPEC-JE) が 3MB、並列文からなる日中論文抄録コーパス (ASPEC-JC) が 680KB 格納されている。ASPEC は、2006 年から 2010 年まで日本で行われた日中機械翻訳プロジェクトの成果の一つとして申し込みの申請を行うことで使用可能になる。配布用のデータには、train.txt : 672,315 対、dev.txt : 2,090 対、devtest.txt : 2,148 対、test.txt : 2,107 が含まれている。

2.1.4 GSK 通訳データベース『日中・日西サブコーパス (JNPC)』

「日中・日西サブコーパス (JNPC)」は特定非営利活動法人 言語資源協会(GSK)会員限定で公開しており、提供元は通訳コーパス作成共同研究者グループ（代表・立教大学 松下佳世）になっている。格納されたデータは、公益社団法人・日本記者クラブで行われた通訳付きの記者会見における、登壇者の原発話（中国語・スペイン語・日本語）と通訳者の訳出を、映像、音声、文字情報を組み合わせた形でデータベース化したものである。原発話と訳出の書き起こしには、自動音声認識技術が用いられており、それぞれタイムスタンプが付与されているため、スプレッドシート等にエクスポートして分析することが可能である。

記者会見は中国語が平均約1時間半、スペイン語が約1時間10分、冒頭の発言ならびに質疑応答からなる。会見数は中国語10件（同時通訳8件、逐次通訳2件）、スペイン語11件（同時通訳6件、逐次通訳5件）である。

2.2 自作対訳データ「日中対訳 EGA コーパス」の紹介

2.2.1 雑誌『Taiwan Panorama』

雑誌『Taiwan Panorama』は台湾で発行された日本語訳文付き（中→日）総合誌で、中国語漢字表記は全て繁体字になっている。1976年1月創刊時には『光華書報 SINORAMA』として20年以上発行されてきたが、2000年1月号から日中対訳付初版が発売され、2006年2月号からは名称を『Taiwan Panorama』に変更している。雑誌は現在Web版でも購読できるようになっている。

『Taiwan Panorama』の対訳文データの収集を開始した時期は2009年からであり、収集済みデータは2008年6月号から2009年5月号までの12ヵ月分に留まった状態である。データ量は、中国語原文総字数205,251字¹、平均文長は93.5字；日本語訳文総字数は250,034字、平均文長は113.9字となっており、トータル2195対の対訳文が含まれている。

	A 年号	B 開始 ページ	C タイトル	D 中国語	E 日本語
1	2009年5月号	82	愛すべき中華料理美食家が勧める「銀翼」と「欣葉」	感、費心地為小吃「變身」，例如，將碗裡油膩的豬肉換成雞肉，再塑成扁薄短小的賣相。美其名為「扁標」；芋頭，番薯原本韌性程度不同，經巧手操合為「芋薯甜糕」，口感極美。其實，小吃入饗的風尚並非由阿嬤開頭，老牌台菜「欣葉」，打從22年前台北市雙城街的11張桌子開始，對台式宴席的想法就充份的創造。	かがされひ。欣葉されたヒルトンホテルのシェフは、宴席に向けてリビング料理を改良し、もじ米料理の碗模様も脂っこい豚肉を鍋肉に変えて小ぶりに仕立て、扁糕と名づけるなど工夫した。
2183	2009年5月号	82	愛すべき中華料理美食家が勧める「銀翼」と「欣葉」	「蓬萊閣」早已消逝。記憶中台北「滙家菜」(如鳳尾大蝦、糖醋魚、肝煎、四色火鍋)雖然在溫州紙北投找回生命，又流入鄉城婚禮從小喜愛燉菜，也會與人合夥開食館的李秀英，當年就矢志將「台灣」端上這常檯面。她延攬了原本在北投餐廳做大廚的師傅陳清南與另一位資深主廚翁茂貴，共同設計出一道筵席料理。	台湾の小皿料理が宴席に並ぶようになったのは、陳前總統が最初ではない。32年前に台北市双城街に11卓で開店した台湾料理の老舗欣葉は、台湾スタイルの裏席という新しいコンセプトの店である。
2184	2009年5月号	82	愛すべき中華料理美食家が勧める「銀翼」と「欣葉」	她的創意可追溯至1970年代。當時，百年前迪化街的巨富出入如江山樓、「蓬萊閣」早已消逝。記憶中台北「滙家菜」(如鳳尾大蝦、糖醋魚、肝煎、四色火鍋)雖然在溫州紙北投找回生命，又流入鄉城婚禮從小喜愛燉菜，也會與人合夥開食館的李秀英，當年就矢志將「台灣」端上這常檯面。她延攬了原本在北投餐廳做大廚的師傅陳清南與另一位資深主廚翁茂貴，共同設計出一道筵席料理。	このセイバは1970年代に遡る。100年前に迪化街の富商が出入りした江山樓や鳳凰閣はすでなく、記憶の中の台湾宴席料理は、田舎の結婚式などにこのセイバは1970年代に遡る。100年前に迪化街の富商が出入りした江山樓や鳳凰閣はすでなく、記憶の中の台湾宴席料理は、田舎の結婚式などに
2185	2009年5月号	82	愛すべき中華料理美食家が勧める「銀翼」と「欣葉」	可怕由於和當時主流的外省宴席完全不合，欣葉度過了一天只有2位客人。「含蓄」的草創期，之後轉型獨立以菜脯蛋、蛋黃肉、煎豬肝等為招牌的輕食「烹風格」，邁進結合起飛的1980年代，欣葉的清粥小菜帶動飲食風潮。平價的台菜竟然也闖出天地。又陸續開創風格殊異的涮涮鍋、日式料理及咖哩屋，事業不斷擴展。	しかし、主中の中国風好みに合わせ、欣葉は一日のお客が二人という修羅たる創業期を逃さなければならなかった。その後は才千大根入り玉子焼きやレバ炒めなど、小皿料理を看板に路線変更した。経済発展が始まつた1980年代、欣葉のあ弱と小皿料理が人気を呼び、低価格の台湾料理の宴席もそれ
2186	2009年5月号	82	愛すべき中華料理美食家が勧める「銀翼」と「欣葉」	2000年起，欣葉把台菜餐飲帶入新加坡、北京與日本，每道菜餚從選料、刀工、火候、油膩控制到調味等環節，都研究出現代標準作業流程。繼承企業的總經理李鴻鈞更常請請廚師赴國試吃與觀察，研究精進新菜肴成果：道地的台式芋頭、以鹽漬滷成的101食主軒，即是現階段的最新菜肴成果：道地的台式芋頭、以鹽漬滷成的101食主軒，即是現階段的	に連れて客がくづくになった。さらにはしゃらしゃら、日本風料理店、カレー店2006年、欣葉は台湾料理レストランをシンガポールや北京、日本に開店。厳選した材料で標準化した調理プロセスを取り入れていた。家業を継いだ李鴻鈞社長は、シェフを通じて外国に視察に出かけ、より先端されたレストランを目指す。台北101大楼55樓、充满时尚感的101食主軒は、最新の視察の成果を発揮している。台湾料理に食前酒から前菜、メインコース、デザートと続くフレンチのコースを取り入れ、日本風の盛り付けで興奮を添える。
2187	2009年5月号	82	愛すべき中華料理美食家が勧める「銀翼」と「欣葉」	以前は台湾料理を看板に並べていたが、今はしゃらしゃら、日本風料理店、カレー店2006年、欣葉は台湾料理レストランをシンガポールや北京、日本に開店。厳選した材料で標準化した調理プロセスを取り入れていた。家業を継いだ李鴻鈞社長は、シェフを通じて外国に視察に出かけ、より先端されたレストランを目指す。台北101大楼55樓、充满时尚感的101食主軒は、最新の視察の成果を発揮している。台湾料理に食前酒から前菜、メインコース、デザートと	たとえば新華四喜と名づけられた前菜を見ると、カラスミやイクラ、明太子寿司などが白い器に盛り付けられ、フレンチの一皿のよろこび美しい。看板料理のレバーソースでは、取り扱われた食材に調理法は昔ながらの調味料に濁れ込んで蒸す手順を急いで炒めることに見え、醤油と砂糖をからめて水分を飛ばしている。レバの口当りを残しながら、あっさりと仕上げて、赤ワインによく合う。
2188	2009年5月号	82	愛すべき中華料理美食家が勧める「銀翼」と「欣葉」	以前は台湾料理を看板に並べていたが、今はしゃらしゃら、日本風料理店、カレー店2006年、欣葉は台湾料理レストランをシンガポールや北京、日本に開店。厳選した材料で標準化した調理プロセスを取り入れていた。家業を継いだ李鴻鈞社長は、シェフを通じて外国に視察に出かけ、より先端されたレストランを目指す。台北101大楼55樓、充满时尚感的101食主軒は、最新の視察の成果を発揮している。台湾料理に食前酒から前菜、メインコース、デザートと	たとえば新華四喜と名づけられた前菜を見ると、カラスミやイクラ、明太子寿司などが白い器に盛り付けられ、フレンチの一皿のよろこび美しい。看板料理のレバーソースでは、取り扱われた食材に調理法は昔ながらの調味料に濁れ込んで蒸す手順を急いで炒めることに見え、醤油と砂糖をからめて水分を飛ばしている。レバの口当りを残しながら、あっさりと仕上げて、赤ワインによく合う。
2189	2009年5月号	82	愛すべき中華料理美食家が勧める「銀翼」と「欣葉」	以前は台湾料理を看板に並べていたが、今はしゃらしゃら、日本風料理店、カレー店2006年、欣葉は台湾料理レストランをシンガポールや北京、日本に開店。厳選した材料で標準化した調理プロセスを取り入れていた。家業を継いだ李鴻鈞社長は、シェフを通じて外国に視察に出かけ、より先端されたレストランを目指す。台北101大楼55樓、充满时尚感的101食主軒は、最新の視察の成果を発揮している。台湾料理に食前酒から前菜、メインコース、デザートと	たとえば新華四喜と名づけられた前菜を見ると、カラスミやイクラ、明太子寿司などが白い器に盛り付けられ、フレンチの一皿のよろこび美しい。看板料理のレバーソースでは、取り扱われた食材に調理法は昔ながらの調味料に濁れ込んで蒸す手順を急いで炒めることに見え、醤油と砂糖をからめて水分を飛ばしている。レバの口当りを残しながら、あっさりと仕上げて、赤ワインによく合う。
2190	2009年5月号	82	愛すべき中華料理美食家が勧める「銀翼」と「欣葉」	以前は台湾料理を看板に並べていたが、今はしゃらしゃら、日本風料理店、カレー店2006年、欣葉は台湾料理レストランをシンガポールや北京、日本に開店。厳選した材料で標準化した調理プロセスを取り入れていた。家業を継いだ李鴻鈞社長は、シェフを通じて外国に視察に出かけ、より先端されたレストランを目指す。台北101大楼55樓、充满时尚感的101食主軒は、最新の視察の成果を発揮している。台湾料理に食前酒から前菜、メインコース、デザートと	たとえば新華四喜と名づけられた前菜を見ると、カラスミやイクラ、明太子寿司などが白い器に盛り付けられ、フレンチの一皿のよろこび美しい。看板料理のレバーソースでは、取り扱われた食材に調理法は昔ながらの調味料に濁れ込んで蒸す手順を急いで炒めることに見え、醤油と砂糖をからめて水分を飛ばしている。レバの口当りを残しながら、あっさりと仕上げて、赤ワインによく合う。

図2 「Taiwan Panorama」対訳文データ

2.2.2 雑誌『聞く中国語』

「聞く中国語」は日本で発行されている語学月刊誌として、ニュース、エッセイ、対談、人物紹介、日常会話集、歌詞、ドラマセリフなど幅広い題材が扱われている。使用されている中国語漢字表記は全て簡体字になっており、（対訳ではなく、日本語だけの記事もあるが）中国語には基本的に日本語訳が付いている。各紙面には写真や絵などが満遍なく使われ、紙

¹ 半角、全角問わず1文字としてカウントしている。

面配置のバリエーションも豊富である。1冊あたり約130ページで構成されており、学習者のために漢字に発音記号のピンインを付している。データ収集を行う際には対訳文の所在を確認しながら手入力で作業を行っている。現時点の入力済みデータは2018年1月号から2023年7月号までの67冊分である。データ詳細は次の表2で示す。

表2. 雑誌「聞く中国語」入力済みデータ統計表

年	総字数	中国語字数	中国語平均文長	日本語字数	日本語平均文長
2018年	515604	212900	43.3	302704	61.6
2019年	419560	171449	41.0	248111	59.4
2020年	437739	177764	37.4	259975	54.7
2021年	395844	162202	35.3	233642	50.9
2022年	399846	162223	35.8	237623	52.5
2023年1月～7月	254923	105586	47.1	149337	66.6
統合	2423516	992124	40.0	1431392	57.6

2.2.3 中国日報社のニュースサイト『人民網』

『人民網』は、中国共産党中央委員会の機関紙『人民日報』で広く知られている中華人民共和国のメディア機関「人民日報社」が1997年1月1日に開設したニュースを主体とするネット情報交流プラットフォームである。『人民網』日本語版には中国語と日本語訳文付きニュースが掲載されている。

『人民網』の対訳文データは現在2014年7月から2023年8月現在まで98ヶ月分の収集が完了している（表3）。

表3. 『人民網』2014年7月～2023年8月現在 データ統計表

年度	総字数	対訳文	中国語字数	中国語平均文長	日本語字数	日本語平均文長
2014年(6ヶ月)	226,175	961	92,440	96.2	133,735	139.2
2015年	450,536	2,115	187,526	88.7	263,010	124.4
2016年	392,339	1,843	161,215	87.5	231,124	125.4
2017年	446,479	2,062	180,478	87.5	266,001	129.0
2018年	454,538	2,070	182,082	88.0	272,456	131.6
2019年	255,066	1,190	102,736	86.3	152,330	128.0
2020年	184,718	882	74,243	84.2	110,475	125.3
2021年	152,826	732	60,468	82.6	92,358	126.2
2022年	134,872	679	53,570	78.9	81,302	119.7
2023年(8ヶ月)	27,308	141	10,680	75.7	16,628	117.9
統合	2,724,857	12675	1,105,438	85.6	1,619,419	126.7

2.2.4 日中対訳 EGA コーパスの文体的特徴

ここでは「聞く中国語」及び「人民網」のデータを用い、語彙の文体的偏りについて確認を行う。

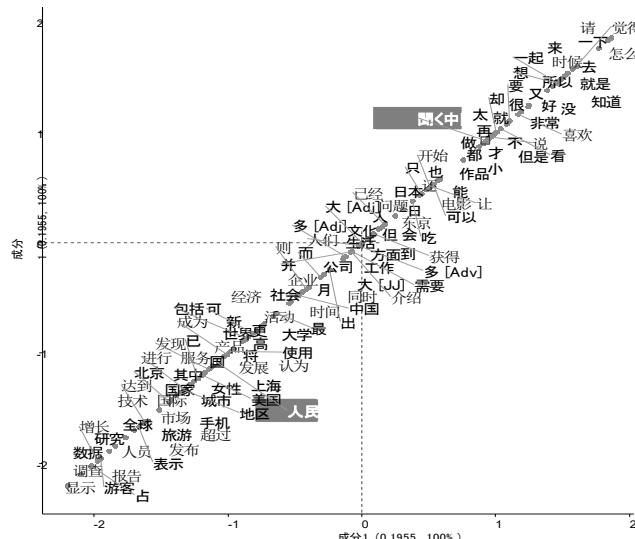


図 3. 言語資源の文体的偏りの対応分析

左の図 3 から読み取れるように、日中対訳 EGA コーパス（以降 EGA コーパス）は文体的偏りが見られず、自然言語を満遍なく収集できたデータサンプルであると考えることができる。即ち、原点からの距離が特徴の強さを意味し、原点に近い位置にある語はあまり特徴がない語になる。特徴がないというのは、外部変数の値に関係なく、まんべんなく出現している語であることを現す。同様、原点から離れた語群は「聞く中国語」と「人民網」各々が持つ特徴語の集まりであることが読み取れる。

3. データの応用事例

本稿では主に KH Coder を用いた応用事例を紹介する。その理由の一つに KH Coder はもっとも有効なテキストマイニングソフトであると考えているからである。

KH Coder には三つの機能がある。「1つ目は、テキストから自動的に語を取り出し、統計的な分析を行う機能。2つ目は、分析者が注目したいコンセプトを取り出し、統計的な分析を行う機能。3つ目は、語やコンセプトの統計分析をもとに、もとのテキストを検索・閲覧するための機能。(樋口他 2022;15)」である。KH Coder を用いた研究事例は 5,000 件を上回っている。今回、ソフトウェア KH Coder（無料版）を用いた応用事例では、雑誌「聞く中国語」2018 年 1 月～2023 年 7 月号（約 242 万字）のデータを使用している。

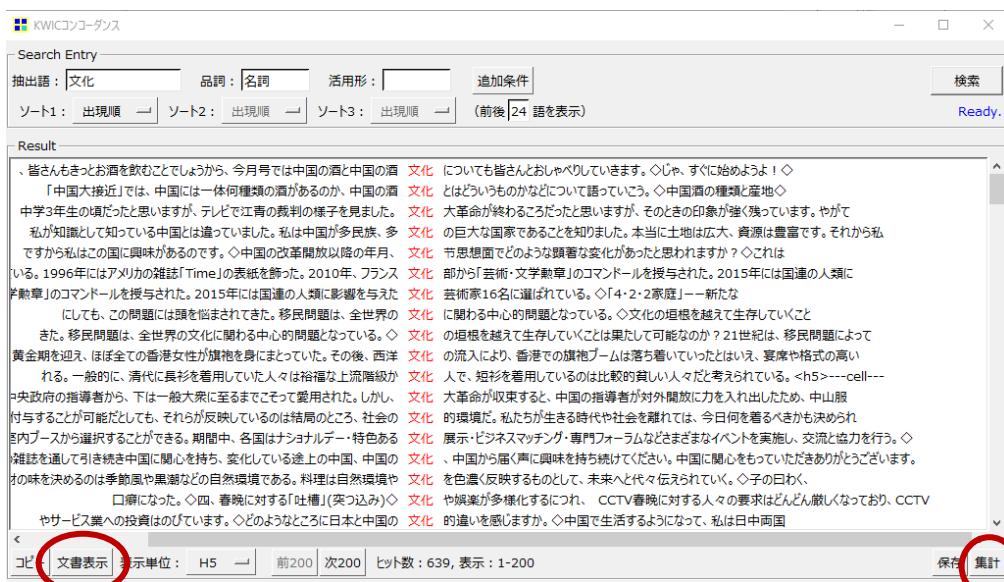


図 4. KH Coder KWIC コンコーダンスの一例

図4では抽出語を「文化」にした場合の例である。現在分析対象にしているのは日本語訳文になっているが、パネル左下の「文書表示」をクリックすることで中国語原文も同時に確認できる（図5）。ファイル内では「<<前」や「次>>」をクリックすることで、ヒットされた用例をスムーズに閲覧することが可能となる。更に図4右下の「集計」をクリックすると、図6のように抽出語「文化」を中心に、その前後の文脈から5グラムの範疇でもっとも共起しやすい語を確認することができる。

ここで注意すべき点は、中国語を日本語に訳した訳出文を構成する語彙は多かれ少なかれその起点言語の影響を受けていることである。

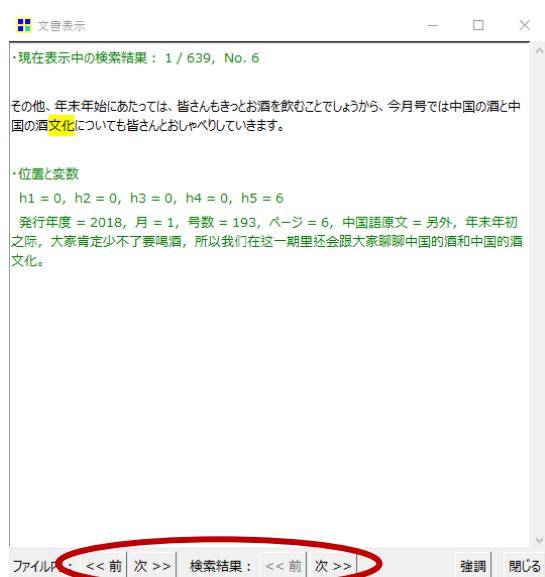
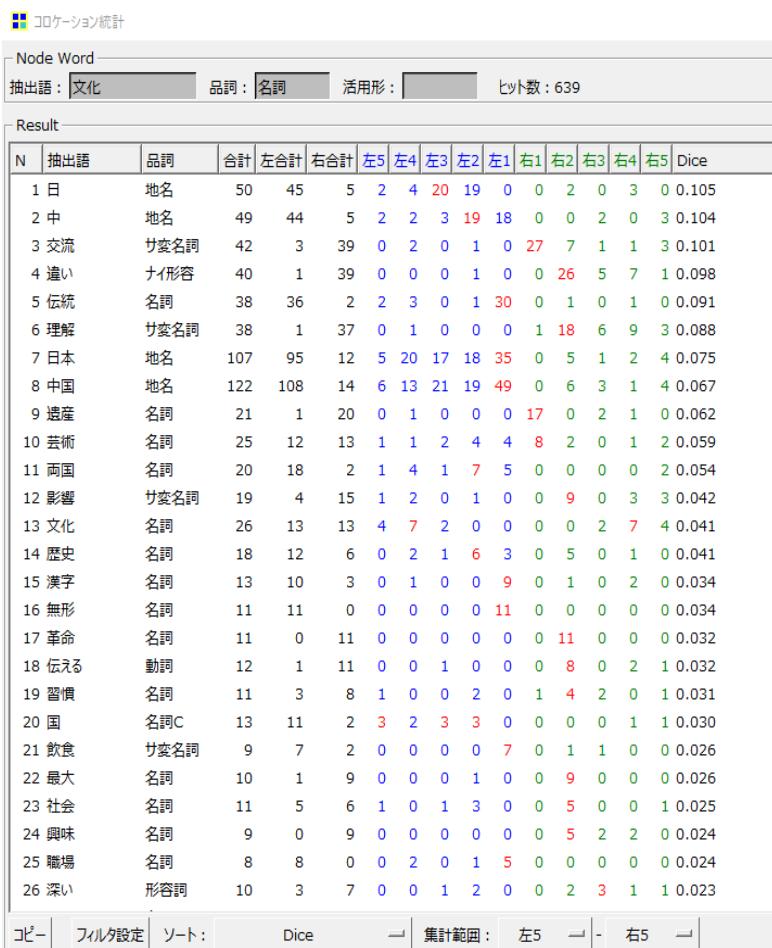


図5. 文書表示機能

図6で得られた結果はあくまでも中国語から（中→日）訳出された日本語として考え、自然言語として使用されている日本語とは区別すべきである。発表者はこの点を切口に、抽出

語と共に起しやすい語群の
日中における違いを観察
することで日中同形語
「文化」の日中間のズレ
は何か、について考察を行
っている。図6から読み取
れるように、中国では「日文化、
日○文化、
中文化、中○文化、
文化交流、文化～違い、
伝統文化、文化～理解、
日本文化、日本～文化、
中国文化、中国～文化、
文化遺産、
芸術、
両国文化、
両国～文化、
歴史文化、
歴史～文化、
漢字文化、
無形文化、
革命文化」
などがよく使われて
いることが明らかとなっ
た。



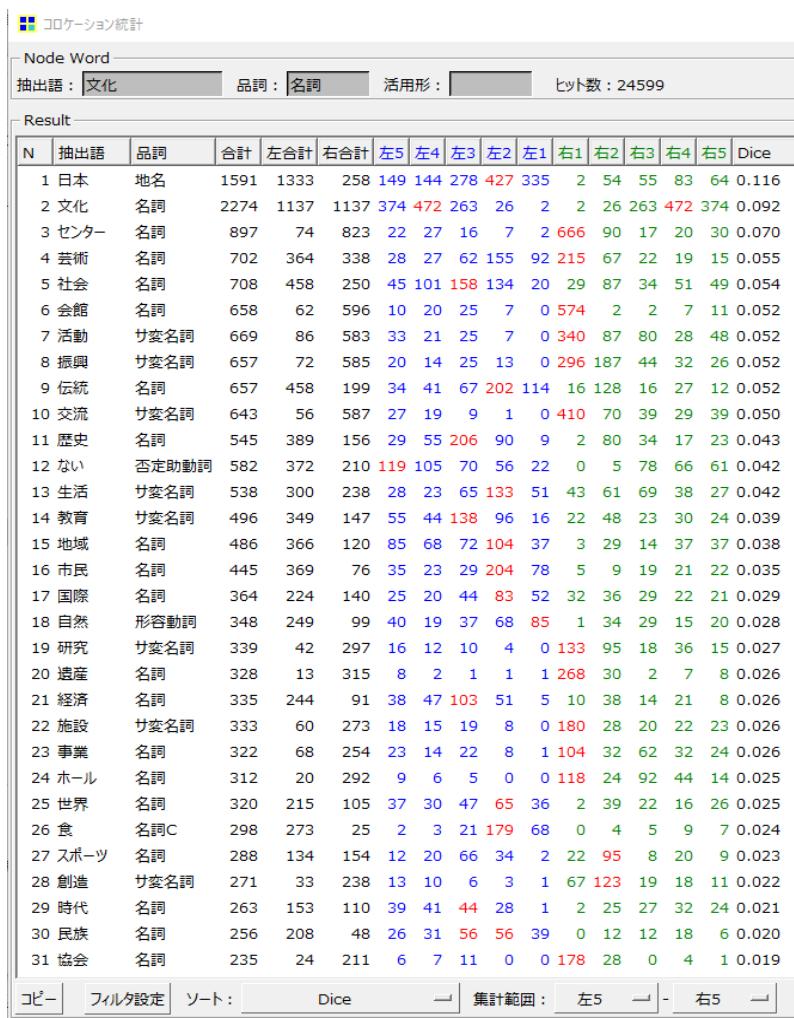
24599例見られ、日本語高頻度コロケーションパターンとして「日本文化、日本～文化、文

図6. コロケーション統計の例

実際、「現代日本語書き言葉均衡コーパス（BCCWJ）」を確認したところ（図7）、「文化」は

化～文化、文化センター、芸術、社会、文化会館、文化活動、文化振興、伝統文化、伝統～文化、文化交流、歴史、生活、教育、地域、市民、国際、自然、文化研究、文化～研究、文化遺産、経済～文化、文化施設、文化事業、文化ホール、世界、食、スポーツ、時代、民族、文化協会、文化主義、中国、発展、文化産業、政治、都市、アメリカ、文化〇革命、新しい、文化人類、影響」などがあげられる。

以上のことから日中同形語「文化」は一見日中間で同形同義語として対応しているようで実際は異なった使われ方が多く存在していることが明らかである。また、日中同形語「文化」の共起ネットワークを用いた比較を行うことで、「文化」という語が含まれた文脈を構成する語群(内容語)が日中間でどのように重なり合うかの考察が可能である(図8)。本稿では日中同形語「文化」を応用事例として取り上げることに留めるが、日本語の「文化」は中国語と違って「会館」



センター、振興、保護、市民、活動、文化財」と結びつく傾向があることが読み取れる。

図7. 日本語「文化」の高頻度コロケーション (BCCWJ)

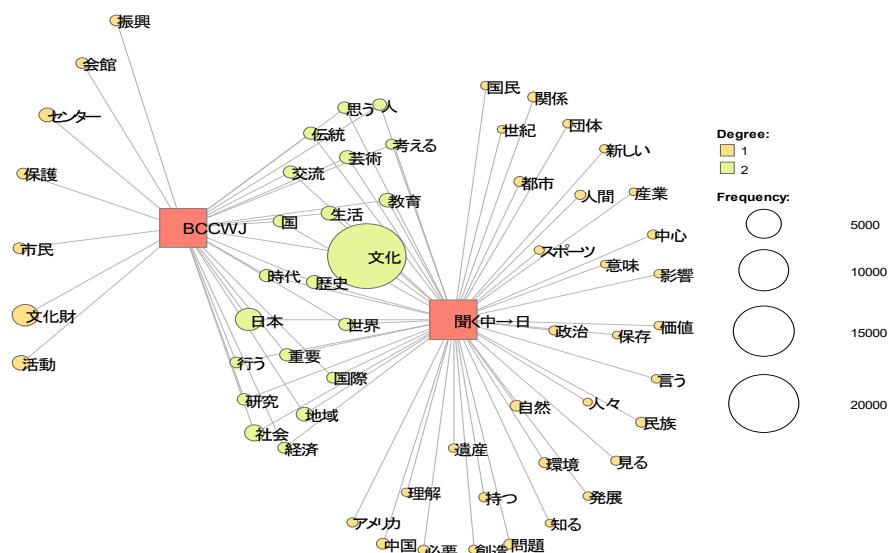


図 8. 高頻度コロケーションの比較（上位 80）

4. コーパスデータの構築及び配布に関わる著作権法

コーパスデータは個人使用の範囲では論文などで引用する際に出典を提示することで原則、著作者の承諾は不要となっているが、その成果物をオープンにしようとすると様々な問題に直面する。特に用例提示が必要とされる言語研究では、著作物の第三者への全文共有は著作権問題がかかわってくる。

機械学習用に加工された学習用データセットに対して現著作者の権利を主張できるかに関しては、加工後の学習用データセットがどのような形になっているかにより結論が代わり得る。基本的に、自然言語処理分野では、データ自体は使い捨てとなるため、学習用データセットの公衆送信は著作権法 30 条の 4 が適用され、現在は原則著作権者の承諾が不要となっており、機械学習等のためであれば公衆送信も認められている。ところが、翻案に関する平成 13 年の判例によれば学習用データセットが生データの本質的な特徴の同一性を維持し、生データの本質的な特徴を直接体感できるような場合には原著作者の権利を主張できることとされている。

4.1 著作権法「第三十条四」とは

著作権法第 30 条 4 は次のように定義されている。

著作物は、次に掲げる場合その他の当該著作物に表現された思想又は感情を自ら享受し又は他人に享受させることを目的としない場合には、その必要と認められる限度において、いずれの方法によるかを問わず、利用することができる。ただし、当該著作物の種類及び用途並びに当該利用の態様に照らし著作権者の利益を不当に害することとなる場合は、この限りでない。

- 一 著作物の録音、録画その他の利用に係る技術の開発又は実用化のための試験の用に供する場合
- 二 情報解析（多数の著作物その他の大量の情報から、当該情報を構成する言語、音、映像その他の要素に係る情報を抽出し、比較、分類その他の解析を行うことをいう。第四十七条の五第一項第二号において同じ。）

—『著作権関係法令・条約集（令和元年版）』pp25-26

4.2 「著作権者の利益を不当に害する」とは

まず疑問になる点として「著作権者の利益を不当に害する」とはなにかである。要するに「著作権者の著作物の利用市場と衝突するか、あるいは将来における著作物の潜在的販路を阻害するか」という観点から、最終的には司法の場で個別具体的に判断されることになる。とされている（上野 2021）。即ち、データの扱いには「著作物の利用市場との衝突や潜在的販路を阻害する」要素を避けるための工夫が必要となるということである。

4.3 「試験の用に供する」とは

次の疑問は「試験の用に供する」という表現であるが、平成 30 年改正前は、「電子計算機による情報解析……を行うことを目的とする場合には」と規定されていたため、情報解析を行う者が自ら著作物等の利用を行う場合が想定されていたが、同改正がこれを「情報解析…の用に供する場合」に変更したことによって、情報解析を行う者が自ら著作物等の利用を行う場合のみならず、情報解析を行う他人のために著作物等を複製したり、譲渡・公衆送信したりすることも権利制限の対象になり得るのである。例えば、情報解析を行う他人のために著作物等を収集して学習用データセットを作成することや、情報解析を行う複数事業者でこれを共有することも許容され得るのである（同上）。上記でいう「複数事業者」の範囲が明確でないこともあります、「著作物を享受する利用法を制限」にも関わることで、著作物の

共有、提供には処理内容（目的）の明確化と、契約に相当するなにかが必要となる。例えば、研究計画、情報処理の仕様書、処理の証跡のような研究上かならず必要なものがあげられる。これは万が一訴訟になった場合の証拠でもあり、安全管理上望ましいことだとされている。対訳文データの共有を著作権上問題がないものとするためには、共有先で研究計画や処理内容が準備されていることが望ましい。

4.4 「情報解析」の範疇

「情報解析」の範疇に関して上野（2021）は次のように述べている。

起草者によれば、情報解析には①ウェブページや書籍等の中に含まれる特定の単語、文字列の用いられ方を分析し、多数のウェブページ、書籍等の間の異同の調査などの統計的な処理を行うウェブ情報解析や言語解析等②音声や映像、画像等に関し、それらを構成する音の波形、影像や文字列等が、どのような事物を意味するかについて、その波形の構成比、輝度・色彩、文字の構成比・出現頻度等を分析し、あらかじめ用意しておいた事物ごとの標準データパターン（特徴）のデータベースと照らし合わせて、その資料がどの事物の標準データパターンに近いのか判別（識別）を行う音声、影像、画像解析等」がこれに当たるとされる[加戸 13]。また、文化庁の解説においては、「深層学習（ディープラーニング）の方法による人工知能の開発のための学習用データとして著作物をデータベースに記録するような場合も対象となるものと考えられる」と述べられている[文化庁 18]。したがって、例えば、画像認識や自動翻訳のAI開発のためにネット上の画像や文章を大量に収集することや、ディープフェイク技術開発のために特定の芸能人の音声や肖像写真を大量に収集することなども「情報解析」に当たると考えられよう。コンピュータを用いない情報解析も許容され得るのである。例えば、大量の新聞記事を情報解析するために人手でコピーをすることや、大量のテレビ番組を情報解析するために人手でハードディスクに録画する場合であっても、「情報解析」に当たり得るのである。

4.5 日本初の事例

日本で著作法第30条4を初めて適用し公開されたコーパスには「昭和・平成書き言葉コーパス（SHC）」があげられる。SHCは高度な検索システムを用いる点に関しては本稿がオープンにしようとするデータとは開示に用いるツールが大きく異なる。小木曾他（2023）は、高樹町法律事務所の小林利明弁護士にも助言を受け、著作物の利用行為が「軽微である」べき点を考慮し、SHC中納言では原文の表示される文脈長を、前後20語～30語範囲に絞ることにしている。

この「軽微である」とは何かについて発表者は、上野（2021）の著者に尋ねたことがある。その際いただいたコメントとして「『軽微』かどうかを問題にされているのは興味深い話であり、著作権法30条の4は、いわば入力の段階はカバーしているが、他方で出力の段階はカバーしていないとされ、その出力段階については、同法47条の5が必要になり、そこでは「軽微」であることが求められるところだということである（逆に言えば、入力段階など、30条の4だけでカバーできる範囲であれば軽微である必要はないことになる）。もっとも、この30条の4と47条の5の関係というのは非常に難問で、実は、知財学者も今さらながらこの問題に頭を悩ませているところである」と貴重な観点をご教示いただいた経緯がある（引用部分に関しては上野氏の許可を得たものである）。

本稿で紹介した自作日中対訳コーパスは全体の分量を「軽微な使用」に適した文長に調整し、更に段落の順番を全てシャッフルしたものをオープン用データとして用いる方法を視野に入れている。その引き換えに文脈情報が大きく乱れることになるが、一先ず日中対訳コーパスの試作としてEGA-Ver.1の配布を試みる。

5. 日中対訳コーパスの試作データ「EGA-Ver.1」の配布

EGA コーパスの試作データとして、2023年12月10日よりEGA-Ver.1の配布（無償）を開始する。データ使用には著作権法の制約があり、データの入手の際には「使用者申込」が必要となる。データの使用を希望される場合は次の「使用申込フォーム」のリンク（URL：<https://forms.gle/SFTi52Tow24shRqZA>）もしくは右側のQRコードより必要事項を記入し（誓約事項への同意を含め）送信することで申し込みは完了となる。お申込み内容に基づき当方から順次データを送付する。



同時に、2024年4月1日以降、日中対訳コーパスの全文を用いた共同研究を募集する。

[\(https://e-ga.jp/page0006.html\)](https://e-ga.jp/page0006.html)

6. まとめと今後の課題

EGA コーパスは現在データの拡大を続けており、今後は中国語を日本語に訳した（中→日）データ及び、日本語を中国語に訳した（日→中）データの収集にも力を入れる予定である。収集済み日→中データには村上春樹の「1Q84」があり、入力済みデータ量は日本語原文約42万字、中国語訳文約28万字となる。

現段階では、データ全体を開示することは難しい状況であるが、近い将来大規模日中対訳コーパスの構築と公開の実現に向けて作業を続けていく所存である。

謝 辞

コーパスデータの収集及び活用手法にあたり、指導教官として終始多大なご指導を賜った、大阪大学人文学研究科基盤日本語学講座の教授石井正彦先生に深謝申し上げる。本稿は2023年3月17日に行われた自然言語処理学会ワークショップで多くの先生方からいただいた適切なご助言を賜った成果でもある。ここでコメントをいただいた先生方々に感謝の意を表す。最後に、幾度に渡り著作権法に関わる知見をご教示頂いた上野達弘先生にこれ以上なくお礼を申し上げる。

文 献

上野達弘(2021) 「情報解析と著作権——『機械学習パラダイス』としての日本」

https://www.jstage.jst.go.jp/article/jjsai/36/6/36_745/_pdf

小木曾智信・近藤明日子・高橋雄太・間淵洋子（2023）「『昭和・平成書き言葉コーパス』

の構築と公開」日本語学会2023年度春季大会ワークショップ予稿集 pp143-156、

共益社団法人著作権情報センター（2019）『著作権関係法令・条約集（令和元年版）』CRIC
張玉潔,馬青,内元清貴,井佐原（2005）「NICT 多言語コーパスにおける日中対訳データの構

築」言語処理学会年次大会発表論文集 2005年3月.pp510-513

張玉潔,王主龍,内元清貴,馬青,井佐原（2008）「日中対訳コーパスにおける単語・句の翻訳

対応関係の付与」言語処理学会第14回 年次大会発表論文集 2008年3月.pp261-264

樋口耕一・中村康則・周景龍（2022）『KH Coder OFFICIAL BOOK II 動かして学ぶ！はじめてのテキストマイニング 一フリーソフトウェアを用いた自由記述の計量テキスト分析』ナカニシヤ出版

樋口耕一著（2020）『社会調査のための計量テキスト分析【第2版】 内容分析の継承と発展を目指して』

文化庁著作権課（2019）「デジタル化・ネットワーク化の進展に対応した柔軟な権利制限規定に関する基本的な考え方」（著作権法第30条の4、第47条の4及び第47条の5関連）https://www.bunka.go.jp/seisaku/chosakuen/hokaisei/h30_hokaisei/pdf/r1406693_17.pdf

北京日本学研究センター（2013）『日中対訳コーパス(CD-ROM)』

前川喜久雄（2009）「代表性を有する大規模日本語書き言葉コーパスの構築」『人工知能学会誌』2009年24巻5号 pp. 616-622, https://doi.org/10.11517/jjsai.24.5_616

前川喜久雄（2010）コーパス構築と著作権保護.研究開発における情報利用と著作権[特集].
人工知能学会誌 25巻5号pp628-632

関連 URL

コーパス検索アプリケーション『中納言』(2023年8月17日現在)

<https://chunagon.ninjal.ac.jp/>

文化庁 Web サイト (2023年8月17日現在)

https://www.bunka.go.jp/seisaku/bunkashinkai/chosakuen/hosei/h20_05/shiryo1_2.html

情報処理学会 Web サイト(2023年3月現在) <https://www.ipsj.or.jp/faq/chosakuen-faq.html>

Web サイト『人民網』(2023年8月17日現在)

<http://j.people.com.cn/95961/index.html>