「みんなで翻刻」による翻刻テキストの分析の試み

加納 靖之（東京大学 地震研究所／地震火山史料連携研究機構）
橋本 雄太（国立歴史民俗博物館 研究部）

市民参加型の史料翻刻プロジェクト「みんなで翻刻」で生成されたテキストに対して，既存の計量テキスト分析用のツールを利用して，頻出語の計数や共起関係の分析を実施した．また，歴史地名データを利用して，テキスト中の地名の同定についても検討した．「地震」という語には，方角や地名に関する語だけでなく，被害に関する語が伴なうことが多いことがわかった．一定の分析結果が得られたものの，分析に利用する辞書の整備や地名の同定方法を洗練されたものにすることが今後の課題である．

Preliminary analyses of the full text produced by “Minna de Honkoku”

Yasuyuki Kano (Earthquake Research Institute, The University of Tokyo)
Yuta Hashimoto (Research Department, National Museum of Japanese History)

We made preliminary analyses of the full text produced by “Minna de Honkoku” which is a project for crowdsourced transcription of Japanese documents on historical earthquakes. We used existing tools for text mining and dictionary to extract frequent words and co-occurrence network. We also tried to extract ace names using an integrated historical gazetteer. The word “earthquake” co-occur with the words that describes direction and place as well as damage. We need good dictionary and gazetteer to obtain better results of text mining.

1．まえがき

京都大学古地震研究会では，2017 年 1 月に「みんなで翻刻【地震史料】」[1] [2][3][4]を公開した（図 1）．「みんなで翻刻」は，Web 上で歴史史料を翻刻するためのアプリケーションであり，これを利用した市民参加型の翻刻プロジェクトである．ここで，「みんなで」は，Web でつながる人々（研究者だけでなく一般の方をふくむ）をさしており，「翻刻」は，くずし字等で書かれている史料（古文書等）を，一字ずつ活字（テキスト）に起こしていく作業のことである．

「みんなで翻刻【地震史料】」では，正式公開から約 1 年 9 か月にあたる 10 月 31 日現在で，掲載されている資料 473 点のうち 469 点の翻刻がひととおり完了している．コマ数では，7087 コマ中 7031 コマで 99%が完了していることになる．総入力文字数は約 531 万文字である．参加登録者は約 4500 人となっている．

掲載されている資料の大部分は東京大学地震研究所図書室が所蔵する資料のうち「古文書」に分類されデジタル画像化されているものである．ほかに，遠野市立博物館の所蔵資料や，京都府立京都学・歴彩館の「京の記憶アーカイブ」[5]や国文学研究資料館の「新日本古典籍総合データベース」[6]でインターネット公開されている資料もふくんでいる．大部分はくずし字の版本であるが，活字やくずし字ではない手書きの資料もあり，これらもすべて翻刻が進んでいる．

古地震（歴史地震）の研究においては，伝来している史料を翻刻し，地震学的な情報（地震発生の日時や場所，規模など）を抽出するための基礎データとする．過去の人々が残した膨大な文字記録のうち，活字（テキスト）になってデータとして活用しやすい状態になっている史料は，割合としてはそれほど大きくはない．

本稿では，「みんなで翻刻」によって生成することができた大量のテキストデータを地震学に活用することを想定して，テキストデータの分析を試行し，本格的な分析に向けての課題を検討する．特に，テキストデータからの語の抽出および頻出語や共起関係の分析と，地名の抽出や同定に注目した分析をおこなった．

図 1　みんなで翻刻のトップページ
Figure 1　Landing page of “Minna de Honkoku.”

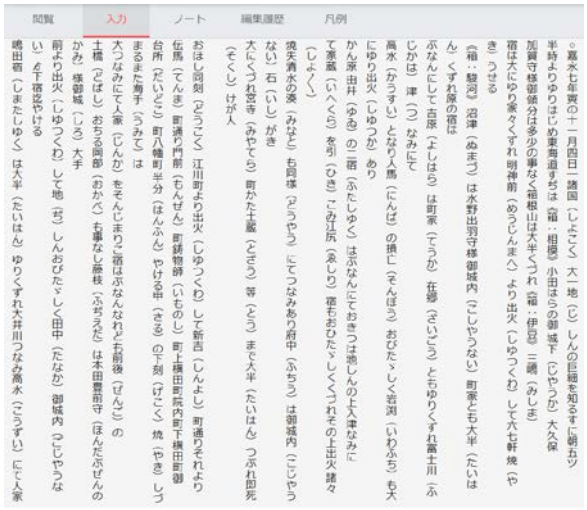

図 2　入力画面の例．「諸国海辺地震津波書」（東京大学地震研究所蔵）の 4 コマ目に対応する．
Figure2　Example of editor of “Minna de Honkoku.”

2．計量テキスト分析

「みんなで翻刻」で生成されたテキストデータに対して，既存のツールによる計量テキスト分析を行なった．対象としたテキストデータは，2018 年 5 月 19 日の時点で「みんなで翻刻」に入力されていた全文テキストである．

現時点では「みんなで翻刻」のシステムには全文を出力する機能はないため，データベースから直接ダンプしたテキストデータを用いた．すべての資料のテキストが一体となったテキストデータである．このテキストデータには，ルビや割書などの情報が独自のマークアップ記法で書きこまれている（図 2）．これらのマークアップは，今回の分析には不要としてすべて削除する前処理を施した．具体的には，正規表現でマークアップにマッチしたものを削除するという単純な手法である．

テキストデータの分析には KH Coder [7] [8]を利用した．KH Coderha 計量テキスト分析（テキストマイニング）のために開発されたフリーソフトウェアである．テキスト分析には辞書が必要となる．KH Coder は標準で IPA 辞書を用いるが，本稿では，利用する辞書を UniDic [9]の「近世口語（洒落本）UniDic」 [10]に変更して分析を実施した．標準搭載の IPA 辞書および UniDic のサイトで公開されているすべての辞書をそれぞれ用いて予備的な分析を実施し，より妥当な語の分解ができたのが「近世口語（洒落本）UniDic」辞書であり，これを採用することにした．

KH Coder では，テキストデータ中に見出し行を入れることで、1 つのテキストファイルをいくつかの部分に区切ることができる．そこで，資料名を見出し行として資料ごとに区切ることとした．これにより資料ごとの分析や資料間の比較も可能になる．このようにして準備したテキストデータは，空白文字や改行も合わせて，約 239 万文字であった．このテキストデータを KH Coder に読み込み，「前処理」（「テキストデータから語を取り出して以後の分析の準備をする処理」）を実施した．

KH Coder には複数の分析機能が備わっているが，まず，頻出語の計数を行った．のべ 155 万語が検出され，ユニークな語数は約 44,000 であった．それぞれの語は，複数の表現をまとめた頻度になっている．例えば「地震」であれば，「地震」，「地しん」，「ぢしん」，「なゐ」の頻度を合計した数となっている．図 3 は，KH Coder で頻出語のうち上位 150 語までを表示させた画面の一部である．表 1 に上位 100 位までの語をしめす．頻出語の上位には「地震」「崩」「水」「人」「山」「火」「町」「寺」「宿」「川」「破損」などが挙がった．これらは，地震とその被害に関する語であり，既刊の地震史料集（たとえば，『増訂大日本地震史料』，『新収日本地震史料』など）による翻刻からの印象とほぼ同じである．計算機による機械的な処理により，定量的に得られた結果をもとに，内容に関する議論がおこなえる可能性がある．

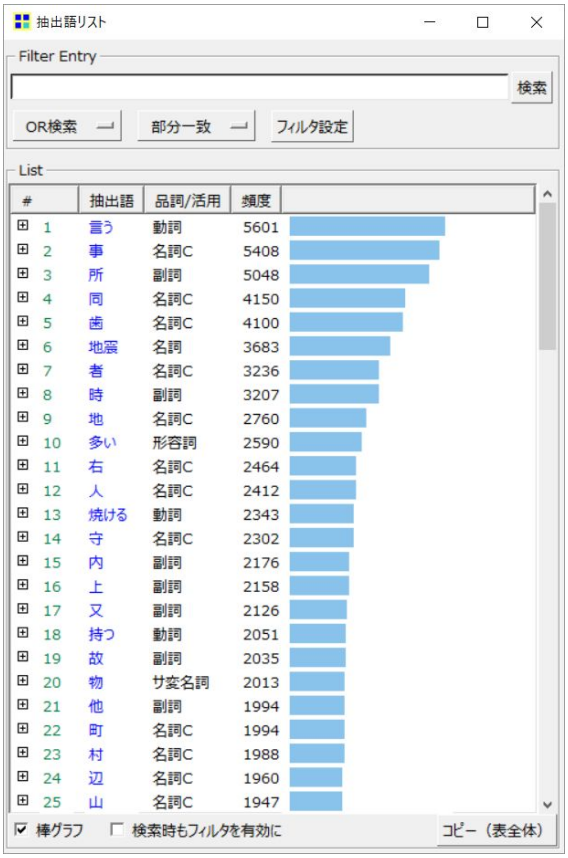

図 3　KH Coder による頻出語上位 25 位までの表示
Figure 3　Top 25 words from full-text of “Minna de Honkoku” extracted using KH Coder.

表 1. 頻出語上位 100 位
Table 1　Top 100 words from full-text of “Minna de Honkoku” extracted using KH Coder.

順	抽出語	頻度	順	抽出語	頻度	順	抽出語	頻度	順	抽出語	頻度
1	言う	5601	26	下	1866	51	左	1176	76	国	885
2	事	5408	27	水	1764	52	南	1176	77	マツヘイ	881
3	所	5048	28	少し	1750	53	今	1174	78	御座	881
4	同	4150	29	辺り	1690	54	由	1125	79	分	878
5	歯	4100	30	屋敷	1649	55	北	1114	80	印	861
6	地震	3683	31	中	1632	56	皆	1113	81	身	854
7	者	3236	32	残る	1577	57	破損	1111	82	数	854
8	時	3207	33	間	1551	58	西	1088	83	然	844
9	地	2760	34	テイ	1544	59	東	1076	84	木	780
10	多い	2590	35	後	1536	60	潰す	1073	85	図	775
11	右	2464	36	潰れる	1523	61	余	1067	86	猶	772
12	人	2412	37	気	1464	62	知る	1038	87	大	758
13	焼ける	2343	38	川	1458	63	然る	1026	88	同所	758
14	守	2302	39	為	1452	64	レ	1023	89	前	754
15	内	2176	40	郡	1427	65	頃	1023	90	海	748
16	上	2158	41	橋	1405	66	出火	1018	91	半	746
17	又	2126	42	方	1393	67	理	959	92	聞く	739
18	持つ	2051	43	至る	1377	68	成す	950	93	目	726
19	故	2035	44	焼く	1368	69	高い	936	94	凡そ	724
20	物	2013	45	出でる	1334	70	夜	928	95	立つ	723
21	他	1994	46	湯	1288	71	因る	914	96	天	711
22	町	1994	47	火	1280	72	船	906	97	津波	706
23	村	1988	48	宿	1190	73	残	897	98	心	693
24	辺	1960	49	家	1186	74	取る	890	99	儀	690
25	山	1947	50	日	1183	75	入る	888	100	金	686

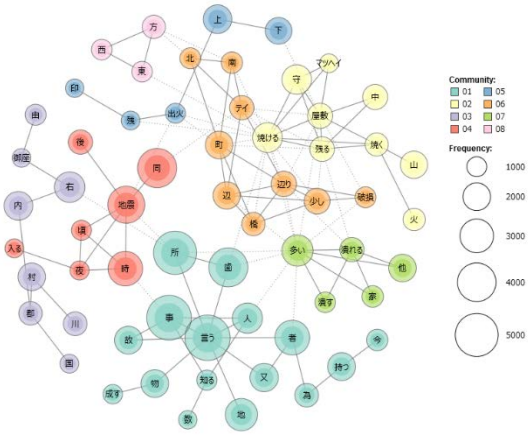

図 4　KH Coder による共起ネットワーク表示
Figure 4　Co-occurrence network produced using KH Coder.

共起関係についても分析した．図 4 は解析例を KH Coder で表示させたものである．上位 120 の共起関係を表示させている．また，代表的な語の共起関係を表 2 にまとめた．「地震」という語には，方角や地名に関する語だけでなく，被害に関する語が伴なうことが多いことがわかった．それぞれの資料で対象となっている地震によって，被害のあらわれ方が違うことから，資料ごとにより詳細に分析することによって，テキスト分析から地震の様相を抽出できる可能性がある．

このようなテキスト分析には，適切な辞書が必要であることは言うまでもない．本稿で利用した「近世口語（洒落本）UniDic」は，文字どおり近世口語文を分析するための辞書であり，地震について書かれた資料の分析に最適の辞書ではない．今回分析した結果をみると，適切な語の分解ができていない部分があった．また，地名や人名についても，適切に判別できていないものが相当数存在する．

たとえば，34 位に挙がっている「テイ」は，「丁」をふくみ，その一部は「町」と同じ意味で用いられている．また，5 位に挙がっている「歯」は「は」や「ハ」で，その多くは助詞の「は」であり，語としては分解できているが，本来助詞であるものが名詞として認識されている．頻出語の 64 位に挙がっている「レ」は感嘆詞として，77 位に挙がっている「理」は名詞として認識されているが，これらの大部分はそれぞれ送りかなの「レ」と「リ」である．77 位に挙がっている「マツヘイ」は姓の「松平」であり，固有名詞としては認識されているが，人名の扱いが不適当であるといえる．これらを正しく認識できるように辞書を修正する必要がある

表 2 主な共起関係
Table 2 Example of Co-occurrence

主な語	共起関係にある語
地震	同，時，頃，夜，後，所，出火
時	地震，頃，夜，事
所	歯（ハ）言う，右，多い
出火	地震，残，町，上，焼ける
焼ける	焼く，屋敷，守，残る，出火，残，町，テイ（丁），橋，辺り，南，多い，歯（ハ）
潰れる	潰す，多い，家，他，残る，辺り

より正確な分析を実施するためには，資料の年代や地震記事であることに対応した辞書が必要となる．UniDic のような既存の辞書を利用しつつ，今回試行したような，実際の史料テキストでの分析の結果を再帰的に反映させることによって，よりよい辞書を作成できるだろう．また，地名や人名など固有表現を抽出する方法も課題である．なお，地名の抽出については，次節で予備的な分析を実施している．

「みんなで翻刻」のようなクラウドソース型のプロジェクトでは，翻刻テキストが入力されるごとに，そのテキストを機械的に処理することによって，キーワードや頻出語を抽出することができる．例えば，これを史料ごと，あるいはプロジェクト全体の画面にタグクラウド等として表示することができるだろう．また，翻刻に参加したテキストが，全体としてどのような意味をもつのかを大雑把につかむこともできるだろう．このような工夫で，さらにプロジェクトを盛り上げ，参加への動機付けを強くすることができると考えられる．

さらに，辞書の作成にあたっても，クラウドソース型の展開をすることが可能かもしれない．翻刻と同時に，あるいは，辞書作成の別プロジェクトとして，うまく分析できなかった語を分類あるいは再定義してもらうなどするシステムを開発することが考えられる．翻刻はできないけれども，テキストベースの辞書作成であれば参加できる，あるいは，辞書や語彙そのものに興味がある，という層も一定程度存在することが想定される．このような層を取り込むことができれば，「みんなで翻刻」そのものややオープンサイエンスの新たな展開をもたらす可能性がある．

３．歴史地名データを利用した分析

「みんなで翻刻」の翻刻テキストには，多くの地名がふくまれている．前節での計量テキスト分析の結果においても述べたように，地震の際には被害の発生した場所等が記録されるからである．同時に，地名の同定に課題があることも既に述べた．

これらの地名を機械的に処理する際には地名辞書が必要となる．適切な辞書があれば，テキストデータからの地名の抽出の精度を上げることができる．また，地理的な座標（緯度，経度など）の情報をふくむ地名辞書があれば，地名から座標に変換するジオコーディングを実施することができる．例えば，翻刻テキストに登場する被害地点を地図上に直接プロットし，被害分布図や震度分布図を作成する，あるいは，震央と被害地点との距離と被害の軽重の関係を議論するなど，地震学的な分析に利用しやすくなる．現代の地震学の解析手法を過去の地震に適用するためにも，地名から座標への変換は重要である．

人間文化研究機構の歴史地名データ[11]を利用した分析をおこなった．「みんなで翻刻」の翻刻テキスト全体から，歴史地名データにマッチしたものを抽出し，地図上にしめしたのが図 5 である．多くの地点が抽出されていることがわかる．しかしながら，実際の地震の被害記録の場所と一致しているわけではない．これにはいくつかの理由が考えられる．まず同名の地点の存在である．たとえば，1666 年，1751 年，1847 年などに地震の被害の記録がある越後高田を例として，「高田」を検索し，マッチした地点を図 6 にしめす．日本の各地に「高田」が存在することがわかる．この場合，前後に越後に関する記述が存在する場合であればその文脈により，また，越後高田で地震が発生したことを事前情報として持っていれば，正しい「高田」を選ぶことができる．

ほかにも，人名と地名で同じものがある，複数の地名を組み合わせてひとつの地名となっているものがあるなど，単純な検索では正しい地点を特定できない場合がある．

テキストからの機械的な歴史地名の抽出や特定は，現時点では不完全ではあるが，少なくとも，ある地名に対応する地点の候補を挙げることには利用できるだろう．これをもとに，人間による判断あるいはテキスト分析を組み合わせた文脈の判断により，より正解に近い地点に絞り込んでいくような手順が考えられる，こうして得られる「正解」を集約し，地震の資料に対応した地名辞書を作成することで，初めて分析するようなテキストであっても，精度よく地点を同定することができるようになる．

歴史地名辞書の作成の作業の一部または全部についても，「みんなで翻刻」のようなクラウドソースによって行なうこともできるだろう．過去の地名に興味を持つ層も一定程度存在すると考えられ，その協力を得てよりよい歴史地名辞書をつくることができれば，オープンサイエンスのプロジェクトとしての展開として有意義なものになるだろう．また，OpenStreetMap(OSM) [12]のような地理情報を作成するプロジェクトと連携することも可能かもしれない．

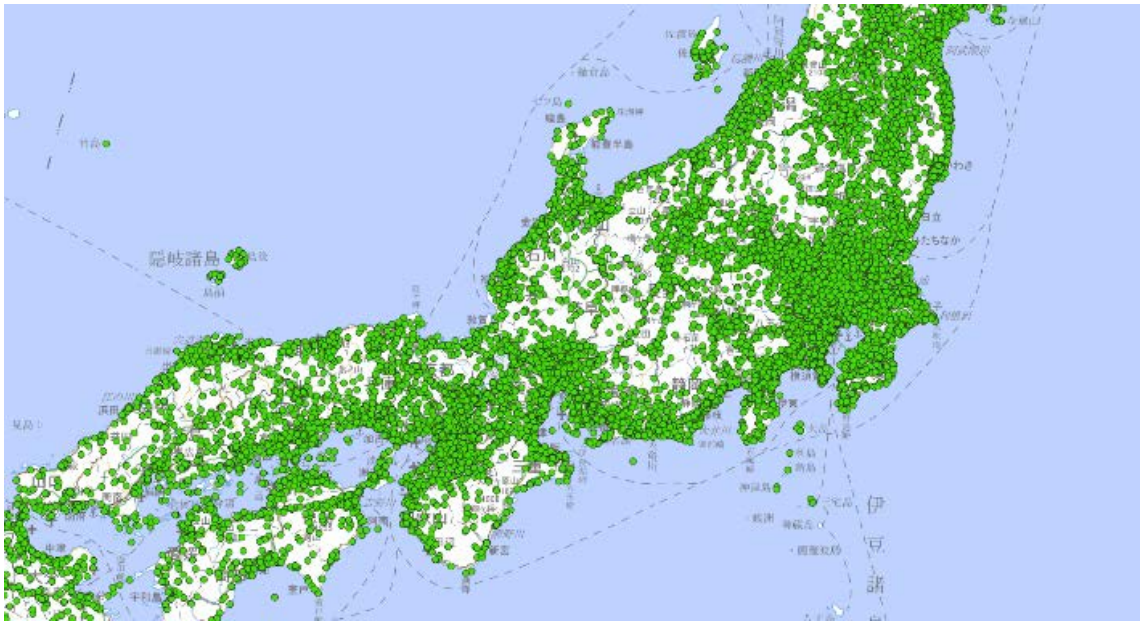

図 5　「みんなで翻刻」テキスト中の地名分布
Figure 5　Distribution of points appeared in full-text of “Minna de Honkoku”

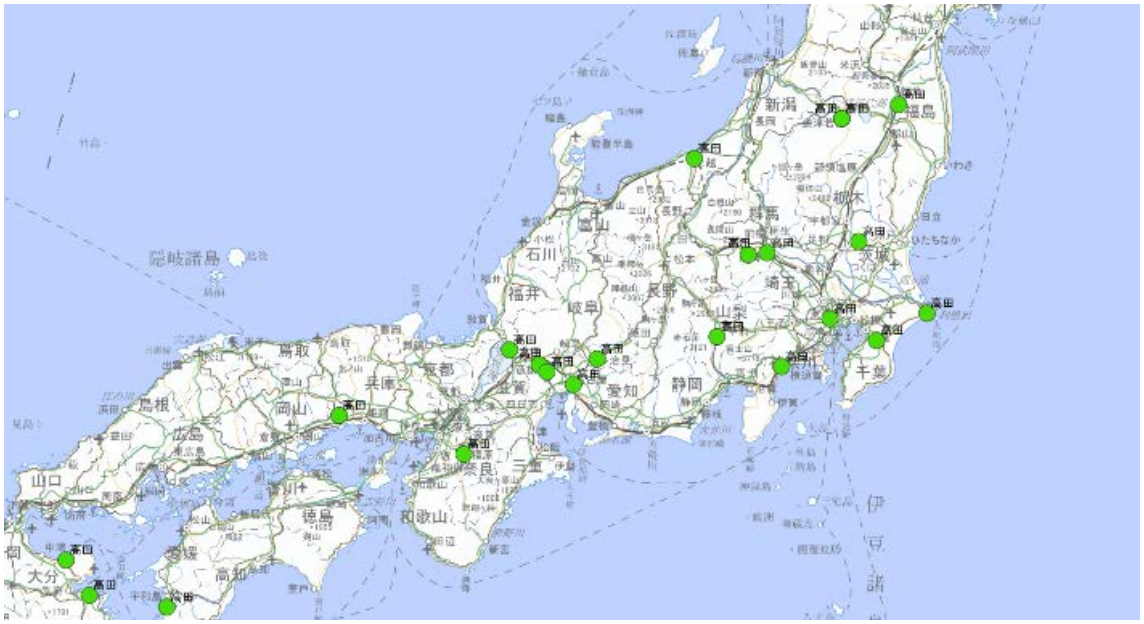

図 6　歴史地名データでの「高田」の分布
Figure 6　Distribution of “Takada” appeared in full-text of “Minna de Honkoku”

4．あとがき

「みんなで翻刻」によって生成された翻刻テキストの分析には，既存のツールと辞書，および歴史地名データを単純に適用するだけでは限界があることがわかった．ここで試行したようなテキスト解析の手法を「みんなで翻刻」のシステムへ実装することを通して，市民参加型のプロジェクトをさらに推進することができるだろう．そのためにも，機械的な処理の手法を洗練する必要がある．

謝辞

「みんなで翻刻」の参加者および資料を提供していただいた東京大学地震研究所図書室，遠野市立博物館に感謝いたします．図 5，図 6 の描画には QGIS [13]を用いました．

参考文献

[1] 京都大学古地震研究会：みんなで翻刻，入手先〈https://honkoku.org/〉（参照 2018-11-01）.
[2] 京都大学：Web アプリケーション「みんなで翻刻【地震史料】」の公開－市民参加で地震史料を後世に残し、新たな史料発掘へー，入手先〈http://www.kyoto-u.ac.jp/ja/research/research_results/2016/170110_1.html〉（参照 2018-11-01）.
[3] 加納靖之：みんなで翻刻―これまでとこれから，リポート笠間，63，入手先〈http://kasamashoin.jp/ 2017/12/63_201630.html〉（参照 2018-11-01）.
[4] Hashimoto, Y., Kano, Y., Nakanishi, I.: Minna de Honkoku: Learning-driven Crowdsourced Transcription of Pre-modern Japanese Earthquake Records, DH 2018, 入手先〈https://dh2018.adho.org/en/minna-de-honkoku-learning-driven-crowdsourced-transcription-of-%E2%80%A8pre-modern-japanese-earthquake-records/〉（参照 2018-09-02）.
[5] 京都府立京都学・歴彩館：京の記憶アーカイブ，入手先〈http://www.archives.kyoto.jp/〉（参照 2018-11-01）.
[6] 国文学研究資料館：新日本古典籍総合データベース，入手先〈https://kotenseki.nijl.ac.jp/〉（参照 2018-11-01）
[7] 樋口耕一：KH Coder: 計量テキスト分析・テキストマイニングのためのフリーソフトウェア，入手先〈http://khc.sourceforge.net/〉（参照 2018-11-01）.
[8] 樋口耕一：社会調査のための計量テキスト分析―内容分析の継承と発展を目指して，p.233 ナカニシヤ出版，京都（2014）.
[9] 国立国語研究所 コーパス開発センター：「UniDic」国語研短単位自動解析用辞書，入手先〈http://unidic.ninjal.ac.jp/〉（参照 2018-11-01）.
[10] 小木曽智信，市村太郎，鴻野知暁：近世口語資料の形態素解析の試み，第 4 回コーパス日本語学ワークショップ予稿集, pp.145-150（2013）.
[11] 人間文化研究機構：歴史地名データ，入手先〈https://www.nihu.jp/ja/publication/source_map〉（参照 2018-11-01）.
[12] OpenStreetMap Japan：歴史地名データ，入手先〈https://www.nihu.jp/ja/publication/source_map〉（参照 2018-11-01）.
[13] QGIS Development Team: QGIS Geographic Information System. Open Source Geospatial Foundation Project. 入手先〈http://qgis.org〉（参照 2018-11-01）.