

動的計画法を用いた狂言台本の語の自動対応付け

北崎 勇帆 (東京大学大学院 人文社会系研究科)

日本語史研究の一手法に、あらすじを同一としながら成立年代の異なる資料組の対応箇所を比較するという方法がある。本稿では書写年代の異なる狂言台本の組を例として、語と語の異同を自動で対応付ける方法を提案する。その際、文字・文単位ではなく語単位の対応付けを行うために動的計画法を二重に用いること、語同士の表記上の比較でなく、語彙素の比較を行うことによって、より適切な対応付けが行えることを示した。

Automatic Word Alignment for *Kyōgen* Scripts Using Dynamic Programming

Yuhō Kitazaki (Faculty of Humanities, The University of Tokyo)

When studying the history of the Japanese language, one possible approach is to compare various manuscripts of the same work from different eras. This paper proposes a method of aligning words between a pair of manuscripts using an iterative double dynamic programming algorithm. The material used in this study is a pair of *kyōgen* scripts. One of them is from the middle of the seventeenth century and the other from the end of the eighteenth century. The data suggests that it is more effective to compare lexemes than morphemes when aligning collations in texts.

1. はじめに

日本語の歴史的な変遷を検討する際に用いられる手法の一つに、同一のあらすじを持つ資料の対応箇所を比較するという方法がある。例えば、『天草版平家物語』(1593刊)とその原拠本とされる『覚一本平家物語』『百二十句本平家物語』の比較、朝鮮で編纂された日本語学習書『捷解新語』(1636頃成立)と、その後の改修本(1748刊)、重刊改修本(1781刊)の比較といった類である。史的变化以外にも、同時代の文体間の差に着目して『今昔物語集』『宇治拾遺物語』といった異なる説話集に収められる共通説話を対象とする、東西言語の差に着目して、江戸・大坂間で改作の見られる洒落本作品群『月花余情』組や『郭中奇譚』組を対象とするなどの方法が取られる。こうした手法には、用いることができる資料の組に限られているという制約はあるものの、「変化」「差異」といったものを捉えやすいというメリットがある。

また、近年は日本語学の研究対象となる言語資料を対象としたコーパスの開発・利用が進んでいる。現代語の書き言葉・話し言葉のみならず、古典語資料についても開発が活発に進められており、中古語資料については『土左日記』『源氏物語』等の文学作品を対象とする『日本語歴史コーパス 平安時代編』[1]が、中世後期については狂言台本を対象とする室町時代編[2]が、江戸時代については洒落本や人情本といった戯作を対象と

する江戸時代編[3]、明治の雑誌資料については『太陽コーパス』を始めとする近代雑誌コーパス[4]が公開されている。さらに、今後も「鎌倉時代編」を含め、資料が追加される予定であるという[5]。

こうしたタグ付きコーパスが利用される際には、それぞれの資料が独立して扱われるか、もしくは網羅的な検索・調査の対象の一つとして処理されることがほとんどであって、コーパス同士が何らかの形で関連付けて扱われたことはあまりなかったと言ってよい。

このような背景から稿者は以前、上述した洒落本作品の組を対象として、異同の対応する箇所を TEI P5 Guideline[6]に基づいてマークアップし、構築されたパラレルコーパスを基に対照表示を行うことを提案した[7]。当該発表は XML のマークアップによる対照コーパス設計の提案を主旨とするものであったため、対応付けの自動化のアルゴリズムについては未検討であり、語単位での細かな対応付けを行うことについても課題として残されたままであった。

そこで本稿では、類似するテキストを語単位で自動的に対応付ける手法の提案を行う。資料は、大蔵流狂言の祖本である大蔵虎明本『狂言之本』(1642書写)と、それから150年下る大蔵虎寛本『能狂言』(1792年書写)を対象とする。

以下、本稿の構成を示す。まず、第2章では、使用する資料を概説し、その冒頭部に見られる異同を示す。第3章で、自動対応付けの先行研究を

紹介し、第4章で稿者が行った実装の過程を示す。その実行結果と課題を第5章に提示し、第6章に展望を述べる。

2. 使用するデータ

本稿では対照に用いる資料の組として、大蔵流狂言の最古本である虎明本『狂言之本』（1642年大蔵虎明による書写、以下『虎明本』）、虎寛本『能狂言』（1792年大蔵虎寛による書写、以下『虎寛本』）の二つを用いる。

『虎明本』は台本の形式を持つ最古の狂言台本であり、全236番を収録する。形態論情報が付与されたタグ付きコーパスが『日本語歴史コーパス 室町時代編 I 狂言』として、2015年3月に国立国語研究所より公開された。『虎寛本』は『虎明本』から150年下の1792年に書写された、大蔵系の現行狂言とほぼ同じ台本である。所収曲は全165番、『虎寛本』所収の曲は概ね『虎明本』にも収録されている。書写年代は江戸後期であるものの、言語現象は近世前期のものを反映していると言われている[8]。『虎寛本』を翻刻したものとして、笹野堅『能狂言』（岩波書店）がある。この二本の狂言テキストを対照させることによって中世・近世間の言語変遷を観察する試みは多く、これまでも一定の成果が上がっている[9,10,11]。

今回は『虎明本』『虎寛本』より、大名狂言「粟田口」[a]を対象とし、語単位でのパラレルコーパスの構築を試みる。語単位での対応付けを行うためには当然ながら、元となる資料の電子テキストと、それを語ごとに区切ったデータが必要となる。

『虎明本』については、『日本語歴史コーパス 室町時代編 I 狂言』のデータを用いた。『虎寛本』については笹野堅『能狂言 上・中・下』を底本に電子テキストを作成し、形態素解析エンジンとして「MeCab」を、解析用辞書として「狂言 Unidic」(ver 0.9)を用いて形態素ごとに分解し、形態論情報を付与した後、人手による修正を行った。

図1に、『虎明本』『虎寛本』それぞれにおける「粟田口」の冒頭部の、対応する文を示す。粗筋・流れは同一であるが、文節・語単位での追加・変更・削除、表記単位の改変といった差異が見られる。

a 近頃大名らの中で「お道具比べ」（所有物を互いに比べること）が流行っており、今度は「粟田口」を比べようとのこと。大名（『虎明本』では果報者）は早速太郎冠者（召使）を呼び出し、「粟田口」が何なのか分からないまま、「その粟田口とやらを買い求めてくるように」と命じる。都に上った太郎冠者は粟田口を探し求め、「自分が粟田口だ」と名乗るすっぱ（詐欺師）に金を払って屋敷に連れ帰る。大名は粟田口を気に入り、何度も粟田口を呼んで遊ぶが、見ていない隙に逃げられてしまう。『虎明本』『虎寛本』ともに、大名狂言に分類される。

虎明本：罷出たる者は、此⁽¹⁾あたり⁽²⁾にかくれもなひ、⁽³⁾大きくわほうの者⁽⁴⁾で御ざる、
虎寛本：罷出たる者は、此⁽¹⁾當りに⁽²⁾隠れもない⁽³⁾大名⁽⁴⁾です。

虎明本：⁽⁵⁾去程に天下⁽⁶⁾おさまり⁽⁷⁾萬⁽⁸⁾めでたひ⁽⁹⁾おり⁽¹⁰⁾なれば、此間⁽¹²⁾方々の御⁽¹³⁾遊山⁽¹⁴⁾おびたしひ事⁽¹⁵⁾で⁽¹⁵⁾ござる、
虎寛本：天下⁽⁶⁾治り、⁽⁸⁾めでたい⁽⁹⁾御代⁽¹⁰⁾で御座れば、此間⁽¹¹⁾の⁽¹²⁾あなたこなたの御⁽¹⁵⁾道具くらべは、⁽¹⁴⁾おびたしひ事⁽¹⁵⁾で⁽¹⁵⁾御座る。

虎明本：⁽¹⁶⁾それに⁽¹⁷⁾つゐ⁽¹⁸⁾て⁽¹⁹⁾申付る事が御ざる、
虎寛本：⁽¹⁶⁾夫に⁽¹⁷⁾付、⁽²⁰⁾今度は持料の粟田口を、くらべさせられうとの御事で御ざるが、某が道具の内に粟田口が有るか、太郎冠者を呼出し、承うと存る。

虎明本：太郎⁽²²⁾くわじや⁽²³⁾有⁽²⁴⁾か⁽²⁴⁾やひ
虎寛本：⁽²¹⁾ヤイ / \、太郎⁽²²⁾くはじや⁽²³⁾有⁽²⁴⁾か⁽²⁴⁾やひ

図1 虎明本・虎寛本テキスト

以上のサンプルに見られる異同を分類すると、以下の通りになる[b]。

- ・表記レベルの異なり
 - (1) [あたり／當り]、(2) [かくれ／隠れ]、(6) [おさまり／治まり]、(8) [めでたひ／めでたい]、(14) [おびたしひ／おびたしひ]、(15) [ござる／御座る]、(16) [それ／夫]、(17) [つゐ・付]、(22) [くわじや／くはじや]、(23) [有／有る]、(24) [やひ／やひ]
- ・語レベルの異なり
 - 形態素を跨ぐ単位で異なるもの
 - (3) [大きくわほうの者・大名]、(4) [で御ざる・です]、(9) [おり・御代]、(10) [なれ・で御座れ]、(12) [方々・あなたこなた]、(13) [遊山・道具くらべ]
- ・片方に見られるが、片方には見られないもの
 『虎明本』にのみ見られるもの
 - (5) [去程に・×]、(7) [萬・×]、(18) [て・×]、(21) [ヤイ / \ ・×]
 『虎寛本』にのみ見られるもの
 - (11) [×・の]

b 括弧内の前者に虎明本の例、後者に虎寛本の例を挙げる。

- ・文以上のレベルでの異なり
- (19) 申付る事が御ざる、
- (20) 今度は持料の粟田口を、……

→『虎明本』では太郎冠者を呼び出してから「道具比べ」の事情を話す、『虎寛本』では呼び出す前に大名によって背景説明が行われるため、後者の文が余ってしまう。

3. 自動対応付け

自然言語処理の文脈において「自動対応付け」の議論は、要約テキストと要約元テキストとの対応付けを行う際に多く見られる。例えば、丸川ら(2002) [12]はローカルラインメント DP マッチングを用いて対応付けを行い、亀田ら(2013) [13]は最長共通部分列に基づく手法の優位性を提示し、立見ら(2014) [14]は文同士を紐付けるのではなく、本文の部分全体とあらすじ文との類似度を用いた DP マッチングの手法を提案している。本稿では類似する文同士が対応付けられていることを前提とし、語同士の同定を主眼に置くこととする。

語同士の対応付ける方法として、まず、それぞれの語と語の間の編集距離[a]を求めることが考えられる。しかしながら、例えば「この」と「此の」、「この」と「その」の編集距離はいずれも1であるが、意味的な側面から言えば前者は同一の語、後者は異なる語であるため、単に表記上での編集距離を求めるのは不適當であろう。日本語の文字列間比較を検討した師(2006) [15]もこの表記を基準とした編集距離の導出に疑義を呈しており[b]、形・音・意味の「文字素性」を用いた Chaon モデル[16]に基づく重み付けを提唱している。但し、単字と複数字の対応付けが困難であるという問題や、そもそもの資材となるデータベースの充実の問題もあり、実際にこの手法を採用するのは困難のようである。

本稿では、第一に、それぞれの語と語の編集距離を求め、第二に、その距離をコストとして最短経路問題へと帰着させる方法を取る。即ち、動的計画法を二重に用いることにより、語単位での対応付けを可能にする。さらに、師(2006)に見られた「表記情報のみに頼らない」という方針を参

a レーベンシュタイン距離は編集距離の一種であり、文字の挿入・置換・削除による、ある文字列から別の文字列への変換の手順の最小回数である。例えば、「まんごー」から「りんご」への距離は、「まんごー」→「まんご」(削除)→「りんご」(置換)により、2と求められる。この数値は、一手順前までの編集距離の最小値を用いた再帰的な漸化式を用いることにより、動的に算出することが出来る。

b 例えば、「有」と「無」、「无」と「無」のコストが同じであることや、逆に「芸(ゲイ)」と「芸(ウン)」のコストが0であることを問題としている。

考に、語形同士の比較ではなく、語彙素[c]同士の比較を行うことにより、この問題の解決を試みた。

4. 実装

上述の通り語彙素同士の比較を行うため、形態素ごとに区切ったリスト(結果表示用)と、それに対応する、語彙素ごとに区切ったリストを格納する。形態素ごとのデータと語彙素ごとのデータのサンプルを、図2・図3に示す。

『虎明本』形態素データ 1: "罷出","たる","者","は","","此","あたり","に","かくれ","も","なひ","","大","くわほう","の","者","で","御ざる","。"

『虎寛本』形態素データ 1: "罷出","たる","者","は","","此","當り","に","隠れ","も","ない","大名","です","。"

『虎明本』形態素データ 2: "去","程","に","天下","おさまり","萬","めでたい","おり","なれ","ば","","此","間","方々","の","御","遊山","おびたしひ","事","で","ござる","。"

『虎寛本』形態素データ 2: "天下","治り","","めでたい","御代","で","御座れ","ば","","此","間","の","あなたこなた","の","御","道具","くらべ","は","","おびたしい","事","で","御座る","。"

図2 形態素データ

『虎明本』語彙素データ 1: "罷り出でる","たり","者","は","","此の","辺り","に","隠れ","も","無い","","大","果報","の","者","なり","御座る","、"

『虎寛本』語彙素データ 1: "罷り出でる","たり","者","は","","此の","当たり","に","隠れ","も","無い","大名","です","。"

『虎明本』語彙素データ 2: "然る","程","に","天下","収まる","万","めでたい","折","なり","ば","","此の","間","方々","の","御","遊山","夥しい","事","なり","御座る","。"

『虎寛本』語彙素データ 2: "天下","収まる","","めでたい","御代","なり","御座る","ば","","此の","間","の","彼方此方","の","御","道具","比べ","は","","夥しい","事","なり","御座る","。"

図3 語彙素データ

c 言語学において、具体的な形としての語を「語形」(word form)、異なる形態から抽出できる抽象的な語を「語彙素」(lexeme)と呼ぶ。

これらのデータ同士の差分を求め、行列上にマッピングする。参考として、形態素の表記同士の編集距離を求めた場合を表1に、語彙素同士の場合を表2に示す。

網掛け部のうち、太字で示した箇所は、形態素であるか語彙素であるかによって編集距離が変わるものであるが、語彙素同士の距離を求めた場合、表記の差異に左右されずに対応付けが出来ることが分かる。

表1 形態素の編集距離をマッピングした行列

明\寛	...	此	當り	に	隠れ	も	ない
∴					∴		
此		0	2	1	2	1	2
あたり		3	2	3	3	3	3
に		1	2	0	2	1	2
かくれ	...	3	3	3	2	3	3
も		1	2	1	2	0	2
なひ		2	2	2	2	2	1

表2 語彙素の編集距離をマッピングした行列

明\寛	...	此	辺り	に	隠れ	も	無い
∴					∴		
此		0	2	2	2	2	2
辺り		2	0	2	2	2	2
に		2	2	0	2	1	2
隠れ	...	2	2	2	0	2	2
も		2	2	1	2	0	2
無い		2	2	2	2	2	0

以上の手順により、文中の全ての語同士の、語彙素に基づく編集距離をプロットした行列が生成される。さらにこれらの語彙素同士に対応関係を与えるため、行列のそれぞれのマスから右・下・右下1マスへ進むためのコストを移動元のポイントと移動先のポイントの和として設定し、ダイクストラ法[a]によって開始行・開始列から最終行・最終列までの最もコストの少ない最短経路を求める。その最短経路上を求める際に通ったマスの行・列の組を順番に出力し、実行結果として出力する。次節に結果を示す。

5. 実行結果と課題

実行結果を整形し、一例を表3・表4に示す[b]。

a ダイクストラ法は、単一始点の最短経路問題を解くためのアルゴリズムであり、動的計画法の一種である。導出には Python のグラフ処理ライブラリである NetworkX を用いた。

b 対応付けが期待され、且つ対応が示されている箇所

表3 実行結果1

ID	虎明本	ID	虎寛本
0	罷出	0	罷出
1	たる	1	たる
2	者	2	者
3	は	3	は
4	、	4	、
5	此	5	此
6	あたり	6	當り
7	に	7	に
8	かくれ	8	隠れ
9	も	9	も
10	なひ	10	ない
11	、	10	ない
12	大	11	大名
13	くわほう	12	です
14	の	13	。
15	者	13	。
16	で	13	。
17	御ざる	13	。
18	。	13	。

表4 実行結果2

ID	虎明本	ID	虎寛本
19	去	14	天下
20	程	14	天下
21	に	14	天下
22	天下	14	天下
23	おさまり	15	治り
24	萬	16	、
25	めでたひ	17	めでたい
26	おり	18	御代
27	なれ	19	で
27	なれ	20	御座れ
28	ば	21	ば
29	、	22	、
30	此	23	此
31	間	24	間
31	間	25	の
32	方々	26	あなたこなた
33	の	27	の
34	御	28	御
34	御	29	道具
34	御	30	くらべ

を太字で強調した。

34	御	31	は
35	遊山	32	、
36	おびたしひ	33	おびたしい
37	事	34	事
38	で	35	で
39	ござる	36	御座る
40	。	37	。

文節の初めと終わりが適切に対応付けられており、期待される語についても概ね対応付けが可能であるようだが、同じ行・列を複数回に渡って通った場合の、前後の余分な箇所処理が問題となる。また、「で 御ざる」と「です」、「なれ(ば)」と「で 御座れ(ば)」のように、一つの形態素に対して一つの形態素が対応していないものに目を向けると、そもそもその組が一語同士の対応ではないために、語彙素同士の対応付けという手法では限界が生じる。

さらに付け加えるならば、文法体系の変遷を検討するという観点から考えると、対照コーパスを用いて観察したいのは時代を超えて一致している箇所よりは、むしろ変化している箇所や改変が加えられている箇所であるため、「何が一致していて何が一致していないか」という二元的な対応付けでなく、「何が変化しているか」という一致・不一致の中間的な異同を同定する必要があるだろう。例えば、次の表5に囲ったような箇所を過不足なく同定することを今後の課題としたい。

表5 あいまいな対応付け

ID	虎明本	ID	虎寛本
98	色々	173	此
99	の	174	間
99	の	175	の
99	の	176	あなたこなた
99	の	177	の
100	お	178	御
100	お	179	道具
101	くらべ物	180	くらべ
102	、	181	は
103	数	182	、
104	を	182	、
105	つくさ	183	夥敷い
106	せ	184	事
107	らるる	185	事
108	事	185	事
109	で	186	で
110	ござる	187	御座る

6. 展望と利用可能性

本稿では狂言台本を対象にしたが、形態素ごとの分かち書きと、語彙素の情報が付与されたテキスト[a]の組であれば、基本的にはどのような資料組でも、DPを用いた語の対応付けが可能である。

日本語史研究の観点から言えば、冒頭に挙げたように、文体の異なる説話集に含まれる同一説話[17]、『天草版平家物語』とその原拠本『平家物語』、東西で改作のある洒落本作品などに適用可能である他、万葉集の原文と読み下し文[b][18]、異本間の異同を調査して校訂テキストを作成する等の利用法が考えられる。

しかしながら、古典テキストとその現代語訳、日本語テキストとその外国語訳のように、文法体系が著しく異なるものについては単純なDPマッチングでは対応関係を正確に捉えることができないため、更なる検討が必要である。そうした事例の一つとして、『源氏物語』の冒頭部の古典本文と現代語訳本文[c]を対象として同様の解析を行った結果を表6に示す。

表6 『源氏物語』の現古対照試行

ID	古典テキスト	ID	現代語訳
0	いづれ	0	帝
1	の	1	は
1	の	2	どなた
1	の	3	の
2	御	4	御代
3	時	5	で
4	に	6	あつ
5	か	7	た
5	か	8	か
6	、	9	、
7	女御	10	女御
8	、	11	や
9	更衣	12	更衣
9	更衣	13	が
10	あまた	14	大勢
11	さぶらひ	15	お
12	たまひ	16	仕え
13	ける	17	し
14	中	18	て

a 精度は落ちるものの、語彙素情報の有無については任意。

b 山田ら(2015)ではIBMモデルを用いて、万葉集の歌の原文(万葉仮名による)と訓読文の語の対応付けを行う手法が提案されている。発表時期の関係により、本稿では十分に参考にすることが出来なかった。

c 古典本文・現代語訳本文ともに、『新編日本古典文学全集』(小学館)を用いた。

14	中	19	おら
14	中	20	れ
14	中	21	た
14	中	22	中
15	に	23	に
16	、	24	、
17	いと	25	最高
18	やむごとなき	26	の
19	際	27	身分
20	に	28	と
21	は	29	は
22	あら	30	いえ
23	ぬ	31	ぬ
23	ぬ	32	お
23	ぬ	33	方
24	が	34	で
25	、	35	、
25	、	36	格別
25	、	37	に
25	、	38	帝
25	、	39	の
25	、	40	ご
25	、	41	寵愛
25	、	42	を
26	すぐれ	43	こうむつ
27	て	44	て
28	時	45	いらつしやる
28	時	46	お
29	めき	47	方
30	たまふ	48	が
31	あり	49	あつ
32	けり	50	た
33	。	51	。

謝辞

本研究に際し、『日本語コーパス 室町時代編 I 狂言』データ、「狂言 Unidic」(ver 0.9)ともに、国立国語研究所の共同プロジェクト「通時コーパスの設計」の成果を一部利用させて頂きました。

参考文献

- 1) 国立国語研究所編：日本語歴史コーパス平安時代編，入手先
http://pj.ninjal.ac.jp/corpus_center/chj/heian.html
 (参照 2015-11-01).
- 2) 国立国語研究所編：日本語歴史コーパス室町時代編 I 狂言，入手先

- 〈http://pj.ninjal.ac.jp/corpus_center/chj/muromachi.html〉(参照 2015-11-01).
- 3) 国立国語研究所編：日本語歴史コーパス江戸時代編，入手先
 〈http://pj.ninjal.ac.jp/corpus_center/chj/edo.html〉
 (参照 2015-11-01).
 - 4) 国立国語研究所：近代語のコーパス，入手先
 〈http://pj.ninjal.ac.jp/corpus_center/cmj/〉(参照 2015-09-01).
 - 5) 田中牧郎：『日本語歴史コーパス』の構築，日本語学，Vol.33，No.14，pp.56-67 (2014).
 - 6) TEI: P5 Guidelines：入手先
 〈<http://www.tei-c.org/Guidelines/P5/>〉，(参照 2015-11-01).
 - 7) 北崎勇帆：洒落本を対象とした東西対照コーパスの設計と建築，研究報告人文科学とコンピュータ(CH)，2015-CH-106，No.5，pp.1-6 (2015).
 - 8) 蜂谷清人：虎明本から虎寛本へ —語形・用法の変遷とその史的位置についての試論—，狂言台本の国語学的研究，笠間書院，pp.25-54 (1977).
 - 9) 柳田征司：虎明本狂言と虎寛本狂言との語彙の比較 —困惑の気持ちを表わす感情語彙に就いて—，安田女子大学紀要，Vol.1，pp.31-44 (1967).
 - 10) 神永正史：虎明本のテアル構文 —競合という観点から—，筑波日本語研究，Vol.14，pp.35-52 (2009).
 - 11) 深津周太：狂言テキストにおける感動詞「シシ申」の歴史的研究，HERSETEC テキスト布置の解釈学的研究と教育，Vol.2，No.2，pp.125-146 (2009).
 - 12) 丸川雄三，岩山真，奥村学，新森昭宏：ローカルアライメントを用いたテキスト間の柔軟な対応付け，研究報告自然言語処理，2002-NL-151，pp.23-28 (2002).
 - 13) 亀田堯宙，李元，内山清子，武田英明，相澤彰子：論文における要約記述に対応するパラグラフの同定手法，*JSAI2013*，(2013).
 - 14) 立見英士，笹野遼平，高村大地：小説における本文とあらすじ文の対応付け，言語処理学会第20回年次大会 発表論文集，pp.496-499 (2014).
 - 15) 師茂樹：文字オントロジに基づく文字オブジェクト列間の編集距離，CHISE Conference 2005 報告書 & CodeFest 京都 2005 資料集，pp.13-19 (2007)
 〈http://researchmap.jp/?action=cv_download_main&upload_id=17840〉(参照 2015-11-01).
 - 16) 守岡知彦，師茂樹：文字素性に基づく文字処理，研究報告人文科学とコンピュータ(CH)，2004-CH-62，pp.53-60 (2004).
 - 17) 田中牧郎，山元啓史：『今昔物語集』と『宇治拾遺物語』の同文説話における語の対応 —語の文体的価値の記述—，日本語の研究，Vol.10，No.1，pp.16-30 (2014).
 - 18) 山田祐実，大村舞，鴻野知暁，Kevin Duh，小木曾智信，松本裕治：万葉集を対象とした原文と読み下し文のアライメント，第8回コーパス日本語学ワークショップ予稿集，pp.243-252 (2015).