コーパスコンコーダンサ『ChaKi.NET』の「文書-部分構造行列」出力機能

浅原 正幸 (国立国語研究所) *
森田 敏生 (総和技研)

Document-Substructure Matrix Output Function on 'ChaKi.NET'

Masayuki Asahara (NINJAL)
Toshio Morita (Sowa Research Co., Ltd.)

要旨

コーパスを用いて統計処理を行う上で、「文書-単語行列」を作成をすることが多い。コーパスコンコーダンサ『ChaKi.NET』は従来より形態論情報に基づくクエリ Tag Search の Wordlist 機能を用いることにより、「文書-単語行列」を作成することが可能であった。今回この機能を拡張することにより、n-gram データや係り受け構造上の部分木などの「文書-部分構造行列」出力機能を実装した。さらに、既存の出力形式である Excel, CSV に加えて、R の dataframe 形式を出力できるようにした。ポスター発表では、当該機能のデモを行う。

1. はじめに

複数文書コーパスを用いて主成分分析や対応分析などの統計処理を行う際に「文書-単語行列」を作成をすることが多い（浅原ほか (2014)）。コーパスコンコーダンサ『ChaKi.NET』(Matsumoto et al. (2006))[(1)]は、Wordlist 機能を用いることにより文書-単語行列を容易に生成することができる[(2)]。特徴量空間として単一の単語表層形や語彙素のみならず、形態素系列（浅原ほか (2015)）や係り受け部分木（浅原・加藤 (2015)）などの部分構造データを用いることにより、より深い分析が行うことができる。しかしながら、部分構造データの枚挙においては、順列・組み合わせの枚挙といった煩雑な作業が伴う。プログラミングに不得手な研究者にとって、この作業が一つの障壁となっている。

今回『ChaKi.NET』の Wordlist 機能を拡張して、n-gram などの連続部分系列や連続部分木などを特徴量空間とする「文書-部分構造行列」を出力する機能を追加した。[(3)] 既存の出力形式である Excel 形式や CSV 形式に加えて、R の dataframe 形式を出力できるようにした。本稿では、これらの新機能を解説するとともに、非連続部分構造を枚挙する際の注意点について示す。

* masayu-a@ninjal.ac.jp

(1) http://osdn.jp/projects/chaki/

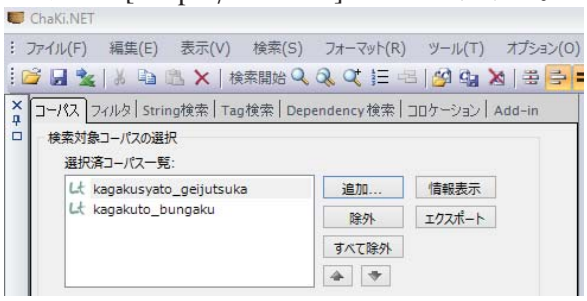
(2) http://qiita.com/masayu-a/items/66285bcb8d40c6bbb494

(3) ChaKi.NET 3.00β Revision 500

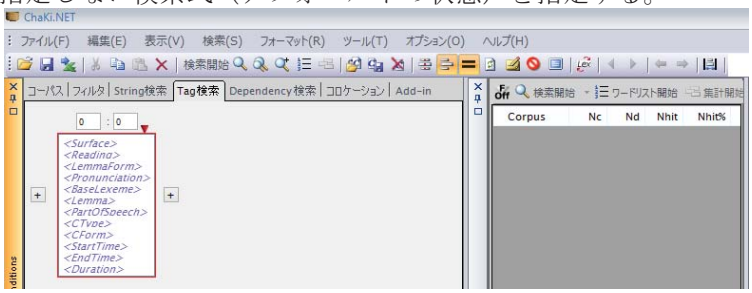
2. 『ChaKi.NET』の Wordlist 機能

最初に『ChaKi.NET』の Wordlist 機能を用いた「文書-単語行列」作成機能について解説する。あらかじめ分析対象のテキストを形態素解析器 MeCab などで解析して、ChaKi.NET 用の sqlite db ファイルを作成してあることを前提とする。後に述べる係り受け部分木に基づく分析を行う場合には、最初から係り受け解析器 CaboCha などで解析してあることが望ましい。(4)

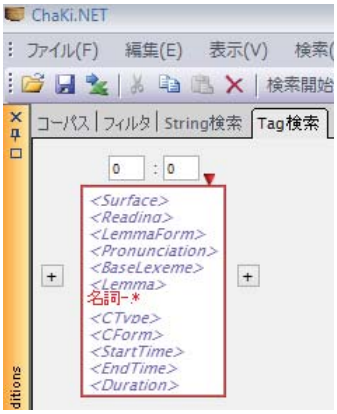
まず最初にコーパスを ChaKi.NET にコーパスを読み込ませる。sqlite db 化した複数ファイルを Search Conditions パネルの [Corpus/コーパス] タブに展開する。



Search Conditions パネルに [Tag Search/Tag 検索] タブを選択し、以下の図のように 1 形態素に対して何も指定しない検索式（デフォールトの状態）を指定する。



特徴量空間として、名詞しか定義しない場合には以下の図のように [PartOfSpeech] に名詞-* を選択する。検索窓が赤字の場合、当該検索クエリが正規表現であることを表す。



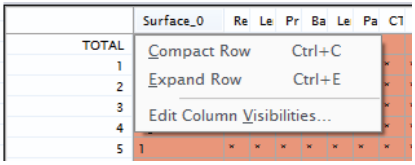
この状態で [Wordlist/ワードリスト開始] ボタンを押すと下図のように「文書-単語行列」が展開される。表中 1 列目から 9 列目が形態論情報を表す。10 列目、11 列目に選択したコーパ

(4) 複数のテキストファイルをバッチで係り受け解析を行い、sqlite db ファイルをに格納する方法については http://qiita.com/masayu-a/items/5e61dcf0ed7068c01f62 を参照すること。

スの頻度が示される。12 列目の [All] の列に全コーパスの頻度が示される。

	Surface_0	Re	Le	Pr	Ba	Le	Pa	CT	CF	kagakus	kagakut	All	Ratio(%)
TOTAL										1365	4978	6343	100
1	–	×	×	×	×	×	×	×	×	2	2	4	0.0630...
2	｜※[#「	×	×	×	×	×	×	×	×	1	0	1	0.0157...
3	2	×	×	×	×	×	×	×	×	1	0	1	0.0157...
4	4	×	×	×	×	×	×	×	×	1	0	1	0.0157...
5	5	×	×	×	×	×	×	×	×	1	0	1	0.0157...
6	68	×	×	×	×	×	×	×	×	1	0	1	0.0157...
7	あまり	×	×	×	×	×	×	×	×	1	0	1	0.0157...
8	アルキメーデス	×	×	×	×	×	×	×	×	1	0	1	0.0157...
9	いかん	×	×	×	×	×	×	×	×	1	3	4	0.0630...
10	インスピレーシ...	×	×	×	×	×	×	×	×	1	0	1	0.0157...
11	ヴォルテーア	×	×	×	×	×	×	×	×	1	0	1	0.0157...
12	うるか	×	×	×	×	×	×	×	×	1	2	3	0.0472...
13	エネルギー	×	×	×	×	×	×	×	×	1	3	4	0.0630...
14	かく	×	×	×	×	×	×	×	×	1	0	1	0.0157...
15	がら	×	×	×	×	×	×	×	×	1	6	7	0.1103...
16	キュービズム	×	×	×	×	×	×	×	×	1	0	1	0.0157...
17	こと	×	×	×	×	×	×	×	×	3	129	132	2.0810...
18	これ	×	×	×	×	×	×	×	×	20	41	61	0.9616...
19	これら	×	×	×	×	×	×	×	×	2	4	6	0.0945...
20	コンジェニアル	×	×	×	×	×	×	×	×	1	0	1	0.0157...

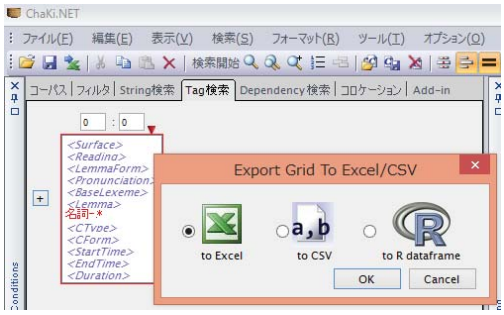
デフォールトの設定では形態素表層形のみが展開されている。各列のヘッダ部を右クリックすることにより、以下の図のように畳み込む [Compact Row Ctrl+C] か、展開する [Expand Row Ctrl+E] かが選択できる。



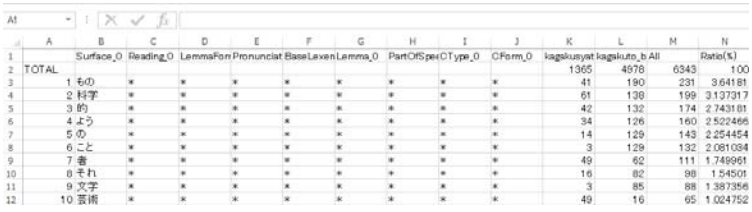
各列のヘッダ部を左ダブルクリックすることにより、当該列で昇順 → 降順にソートされる。以下の図は [All] 列 (全コーパス中の頻度) で降順ソートしたものである。

	Surface_0	Re	Le	Pr	Ba	Le	Pa	CT	CF	kagakus	kagakut	All	Ratio(%)
TOTAL										1365	4978	6343	100
1	もの	×	×	×	×	×	×	×	×	41	190	231	3.6418...
2	科学	×	×	×	×	×	×	×	×	61	138	199	3.1373...
3	的	×	×	×	×	×	×	×	×	42	132	174	2.7431...
4	よう	×	×	×	×	×	×	×	×	34	126	160	2.5224...
5	の	×	×	×	×	×	×	×	×	14	129	143	2.2544...
6	こと	×	×	×	×	×	×	×	×	3	129	132	2.0810...
7	者	×	×	×	×	×	×	×	×	49	62	111	1.7499...
8	それ	×	×	×	×	×	×	×	×	16	82	98	1.5450...
9	文学	×	×	×	×	×	×	×	×	3	85	88	1.3873...
10	芸術	×	×	×	×	×	×	×	×	49	16	65	1.0247...
11	場合	×	×	×	×	×	×	×	×	10	53	63	0.9932...
12	事	×	×	×	×	×	×	×	×	49	13	62	0.9774...
13	これ	×	×	×	×	×	×	×	×	20	41	61	0.9616...

この状態で [File/ファイル (E)] → [Send To Excel/CSV] を選択し、[to Excel] を選択するとと展開された「文書-単語行列」を保存することができる。尚、Microsoft Excel がインストールされていない機材の場合はこの機能が利用できない。



保存された Excel ファイルは以下のようになる。



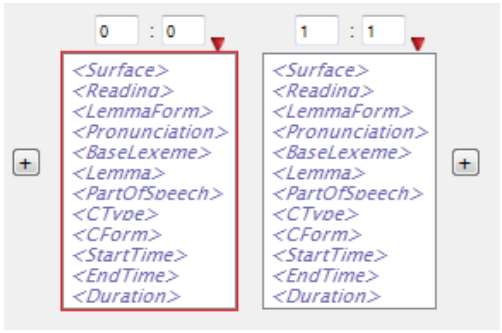
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Surface_0	Reading_0	LemmaForm	Pronunciat	BaseLexem	Lemma_0	PartOfSpee	CType_0	CForm_0	kagakusyat	kagakuto_b	All	Ratio(%)
2	TOTAL										1365	4978	6343	100
3	1	もの	*	*	*	*	*	*	*	*	41	190	231	3.64181
4	2	科学	*	*	*	*	*	*	*	*	61	138	199	3.137317
5	3	的	*	*	*	*	*	*	*	*	42	132	174	2.743181
6	4	よう	*	*	*	*	*	*	*	*	34	126	160	2.522466
7	5	の	*	*	*	*	*	*	*	*	14	129	143	2.254454
8	6	こと	*	*	*	*	*	*	*	*	3	129	132	2.081084
9	7	者	*	*	*	*	*	*	*	*	49	62	111	1.749961
10	8	それ	*	*	*	*	*	*	*	*	16	82	98	1.54501
11	9	文学	*	*	*	*	*	*	*	*	3	85	88	1.387356
12	10	芸術	*	*	*	*	*	*	*	*	49	16	65	1.024752

前の画面で [to CSV] を選択すると csv 形式のファイルが、[to R dataframe] を選択すると R 言語の dataframe 形式のファイルが出力される。

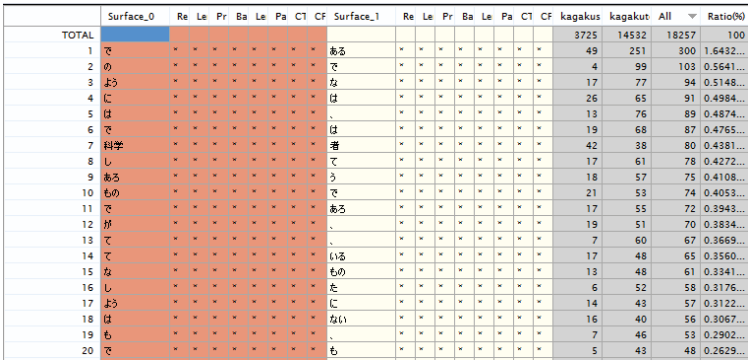
3.　文書-連続部分系列行列

以下では、文書-部分系列行列の展開方法について説明する。

Search Conditions パネルに [Tag Search/Tag 検索] タブを選択し、以下の図のように 2 形態素に対して何も指定しない検索式を指定することにより bigram 特徴量空間を考慮した文書-部分系列行列が展開できる。

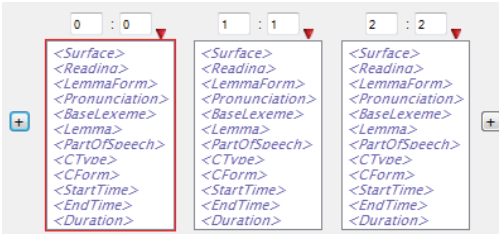


1 列目から 9 列目までが前件の形態論情報で、10 列目から 18 列目が後件の形態論情報である。19 列目以降に頻度情報が格納される。

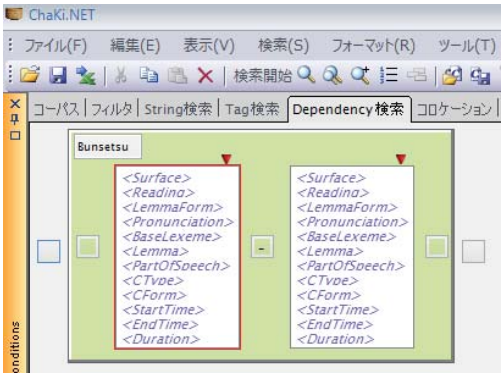


	Surface_0	Re	Le	Pr	Ba	Le	Pa	CT	CF	Surface_1	Re	Le	Pr	Ba	Le	Pa	CT	CF	kagakus	kagakut	All	Ratio(%)
TOTAL																			3725	14532	18257	100
1	で	*	*	*	*	*	*	*	*	ある	*	*	*	*	*	*	*	*	49	251	300	1.6432...
2	の	*	*	*	*	*	*	*	*	で	*	*	*	*	*	*	*	*	4	99	103	0.5641...
3	よう	*	*	*	*	*	*	*	*	な	*	*	*	*	*	*	*	*	17	77	94	0.5148...
4	に	*	*	*	*	*	*	*	*	は	*	*	*	*	*	*	*	*	26	65	91	0.4984...
5	は	*	*	*	*	*	*	*	*	、	*	*	*	*	*	*	*	*	13	76	89	0.4874...
6	で	*	*	*	*	*	*	*	*	は	*	*	*	*	*	*	*	*	19	68	87	0.4765...
7	科学	*	*	*	*	*	*	*	*	者	*	*	*	*	*	*	*	*	42	38	80	0.4381...
8	し	*	*	*	*	*	*	*	*	て	*	*	*	*	*	*	*	*	17	61	78	0.4272...
9	ある	*	*	*	*	*	*	*	*	う	*	*	*	*	*	*	*	*	18	57	75	0.4108...
10	もの	*	*	*	*	*	*	*	*	で	*	*	*	*	*	*	*	*	21	53	74	0.4053...
11	で	*	*	*	*	*	*	*	*	ある	*	*	*	*	*	*	*	*	17	55	72	0.3943...
12	が	*	*	*	*	*	*	*	*	、	*	*	*	*	*	*	*	*	19	51	70	0.3834...
13	て	*	*	*	*	*	*	*	*	、	*	*	*	*	*	*	*	*	7	60	67	0.3669...
14	て	*	*	*	*	*	*	*	*	いる	*	*	*	*	*	*	*	*	17	48	65	0.3560...
15	な	*	*	*	*	*	*	*	*	もの	*	*	*	*	*	*	*	*	13	48	61	0.3341...
16	し	*	*	*	*	*	*	*	*	た	*	*	*	*	*	*	*	*	6	52	58	0.3176...
17	よう	*	*	*	*	*	*	*	*	に	*	*	*	*	*	*	*	*	14	43	57	0.3122...
18	は	*	*	*	*	*	*	*	*	ない	*	*	*	*	*	*	*	*	16	40	56	0.3067...
19	も	*	*	*	*	*	*	*	*	、	*	*	*	*	*	*	*	*	7	46	53	0.2902...
20	で	*	*	*	*	*	*	*	*	も	*	*	*	*	*	*	*	*	5	43	48	0.2629...

trigram 以上の特徴量空間を規定するためには以下のように形態素の box を増やせばよい。



係り受け解析結果を格納することにより、文節境界の情報がデータベースに格納される。[Dependency Search/Dependency 検索] 機能を用いることにより、文節を越えない部分系列のみを展開することができる。以下の図は、文節内 bigram のみを特徴量とした文書-部分系列行列を展開するための式である。内側の形態素の boxes 間に - を入れることにより、2 形態素が隣接していることを表している。



4.　文書-非連続部分系列行列作成時の重複枚挙の問題

4.1　連続部分系列と非連続部分系列

前節では連続部分系列 (n-gram) を特徴量空間にした場合の「文書-部分系列行列」を展開する方法を述べた。本節では非連続部分系列 (p-mer) を特徴量空間にした場合の「文書-部分系列行列」の展開する方法と注意点について述べる。

非連続部分系列 (p-mer) とは、連続していないとびとびの部分列のことである。特に言及しない場合、非連続部分系列 (p-mer) は連続部分系列 (n-gram) を含むものとする。n-gram とは系列に対する長さ n の連続部分列 (substring) のことをいい，p-mer とは系列に対する長さ p の部分列 (subsequence) のことをいう。

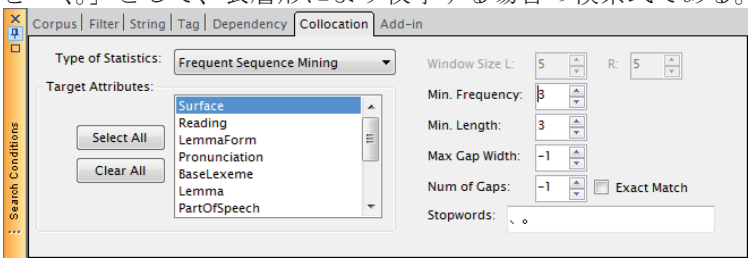
例えば “ABCDE” という系列に対して，3-gram は “ABC”, “BCD”, “CDE” の 3 種類あり，3-mer は “ABC”, “AB/D”, “AB/E”, “A/CD”, “A/C/E”, “A/DE”, “BCD”, “BC/E”, “B/DE”, “CDE” の 10 種類あり，それぞれ頻度は 1 である。p-mer の “/” は，そこにギャップがあることを意味している。

文全体にわたって非連続部分系列を枚挙する方法として、系列パターンマイニングアルゴリズム (Pei et al. (2001)) が知られている。ChaKi.NET には検索した文に対して、頻出系列パターンを枚挙する機能が実装されている。

4.2　既存の非連続部分系列枚挙機能

1 文書に対する非連続部分系列枚挙機能は以前から ChaKi.NET に実装されている。

[Search Condition] パネルから [Collocation/コロケーション] タブを選択し、[Type of Statistics] に "Frequent Sequence Mining"を選択することによって、頻出系列パターンの枚挙が行われる。以下の例では、最小頻度 3、最小系列長 3、最大ギャップ長 ∞、最大ギャップ数 ∞、ストップワードを 「、。」として、表層形により枚挙する場合の検索式である。

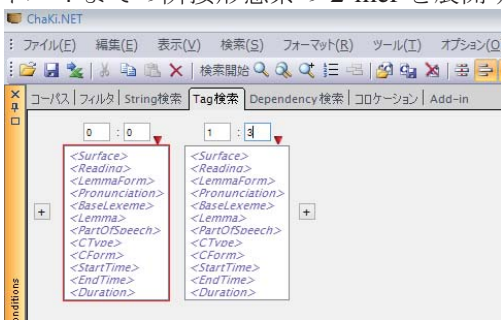


この手法では、1 文書毎に同じ作業を行う必要がある。

4.3　Wordlist 機能を用いた非連続部分系列枚挙

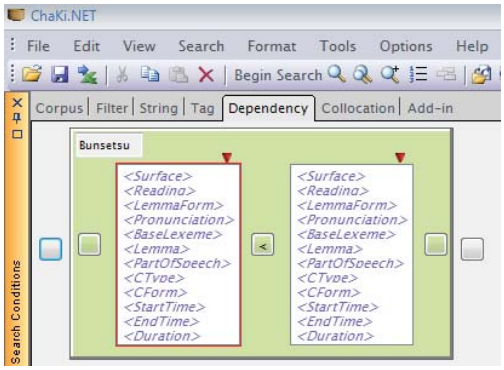
以下 Wordlist 機能を用いて、非連続部分系列を枚挙する方法について述べる。[Tag Search/Tag 検索] では、形態素の box の上についている index により、形態素の隣接性を規定することができる。

以下の例は Windows サイズ 4 までの隣接形態素の 2-mer を展開する検索式である。



Window サイズ n を広げると、各形態素位置に対して ${}_nC_p$ の組合せが展開されるので注意すること。

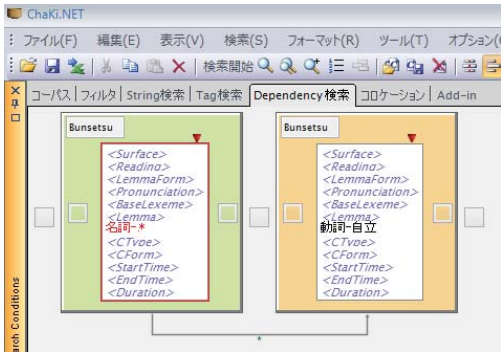
Window サイズを制限する他の方法として、文節境界により p-mer の枚挙を制限する方法がある。[Dependency Search/Dependency 検索] で以下の検索式を指定すると、文節内 2-mer を枚挙する。2 形態素 boxes 間の <は形態素の順序を規定する。この記号がない場合は、逆順についても枚挙してしまうので注意すること。



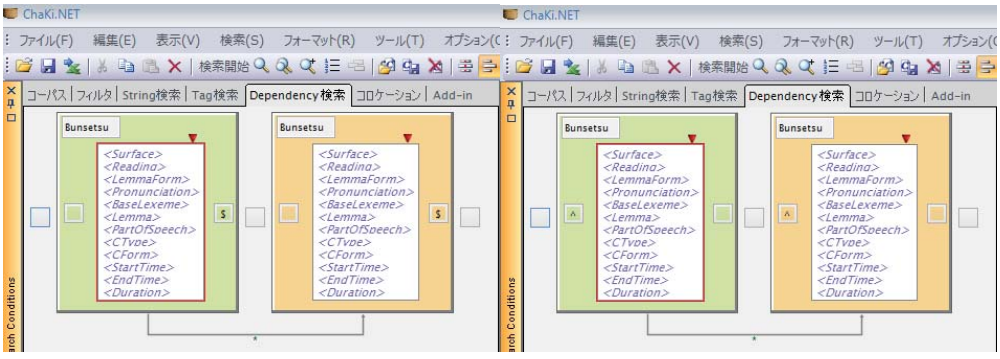
5. 文書-部分木行列

係り受け部分木を特徴量空間にする場合、[Dependency Search/Dependency 検索] を用いて Wordlist 機能を用いればよい。

以下の例では「動詞-自立」に係る「名詞」を枚挙する。しかし、文節内の形態素の位置を規定していないため、1 文節内に複数の名詞が存在する場合には、それぞれ別のものとして枚挙される。



残念ながら、文節内の形態素位置については先頭位置か末尾位置しか指定することができない。以下の例は各文節内形態素の出現位置を先頭位置にしたもの（左図）と末尾位置にしたもの（右図）である。



6. おわりに

本発表では、コーパスコンコーダンサ ChaKi.NET の「文書-部分構造行列」出力機能について紹介した。ChaKi.NET は他にも様々な機能がある (浅原・森田 (2013, 2014, 2015)) ので組み

合わせて利用されたい。

謝辞

本研究の一部は科研費基盤 (B)「言語コーパスに対する読文時間付与とその利用」(25284083)、科研費萌芽「近代語コーパスに対する統語情報アノテーション基準策定」(15K12888)、国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国語研「超大規模コーパス構築プロジェクト」によるものです。

参考文献

Matsumoto, Yuji, Masayuki Asahara, Kiyota Hashimoto, Yukio Tono, Akira Otani, and Toshio Morita (2006). “An annotated corpus management tool: Chaki.” *Proc. of LREC-2006*, pp. 1418–1421.

Pei, Jian, Jiawei Han, Behzad Mortazavi-Asi, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu (2001). “PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth.” *Proceedings of the 17th International Conference on Data Engineering*, pp. 215–224.

浅原正幸・加藤祥 (2015).「文体指標を特徴づける係り受け部分木の抽出」 第 8 回コーパス日本語学ワークショップ.

浅原正幸・加藤祥・立花幸子・柏野和佳子 (2014).「文体指標と語彙の対応分析」 第 6 回コーパス日本語学ワークショップ, pp. 11–20.

浅原正幸・加藤祥・立花幸子・柏野和佳子 (2015).「文体指標と語彙系列の対応分析」 第 7 回コーパス日本語学ワークショップ, pp. 7–16.

浅原正幸・森田敏生 (2013).「コーパスコンコーダンサ『ChaKi.NET』の連続値データ型」 第 4 回コーパス日本語学ワークショップ, pp. 223–232.

浅原正幸・森田敏生 (2014).「コーパスコンコーダンサ『ChaKi.NET』の連続値データ型 (2)―読み時間の表示―」 第 5 回コーパス日本語学ワークショップ, pp. 39–48.

浅原正幸・森田敏生 (2015).「コーパスコンコーダンサ『ChaKi.NET』のプロジェクト機能」第 7 回コーパス日本語学ワークショップ, pp. 103–112.