「文脈」の視覚化の試み

齊藤正高

要約：形態素解析とネットワーク視覚化の技法を組合せ，日本語テキストデータから語と語の繋がりを視覚化し，文書外の知識によらないで，一定の客観的構造を出力した。

キーワード：形態素解析・文脈・視覚化・日本語・有向グラフ

1. はじめに

本稿は，国際貿易や夫婦関係などの分析に用いられているネットワーク視覚化の手法と，テキスト処理の技法を組合せ，日本語の文献[1]における語の遷移関係[2]に注目し，その文献の構造（「文脈」）を視覚化する試みである。

実例としては，夏目漱石『夢十夜』[3]を用いた。

2.「文脈」について

辞書を参照すると，「文脈」とは①「文中の語の意味の続きぐあい」，②「文章の中での文と文との続きぐあい」，③「（比喩的に）筋道・背景」と定義[4]されている。

本稿では最も単純な①の意味として「文脈」を考え，文を単位とした関係（②）や比喩的な意味（③）はとらない。①の定義における「語の意味」については，「意味と形態の１対１対応」[5]から，意味を形態で近似する。したがって，「語の意味」は「語の形態」として処理し，「意味の続きぐあい」は「形態の遷移」として処理する。「遷移」とは，一つの文のなかである語のあとに次の語がつづくことである。

「語の形態」はテキストデータにChasen[6]などのソフトウェアによって，形態素解析をほどこすことで抽出でき，「形態の遷移」はGraphviz[7]などのソフトウェアによって視覚化することができる。

たとえば，Chasenは以下のように形態素解析の結果を出力する。

表層語	基本形	よみ	品詞
こんな	こんな	コンナ	連体詞
夢	夢	ユメ	名詞
を	を	ヲ	助詞
見	見る	ミ	動詞
た	た	タ	助動詞

また，GraphVizは有向グラフを記述するdot言語[8]をもっており，上の形態素分析の結果をdot言語で表現すると，次のようになる。

```
digraph a{
node[
shape=plaintext,
fontname="arialuni.ttf",
fontsize=10
];
こんな->夢;
夢->を;
を->見る;
見る->た;
}
```

1行目の“digraph”とは有向グラフであることを表し，“a”はグラフ名である。“digraph”の部分を“graph”とかけば，無向グラフを定義することもできる（ただしこの場合7行目以下の関係記述子は“--”とする必要がある）。2～6行目はノードの形状及びフォント名，フォントサイズの定義であり，日本語や中国語を用いる場合は必要である。

このdot言語を記述したファイルをutf-8の形式で保存し，GraphVizによって画像データを出力すると以下の画像を得ることができる。

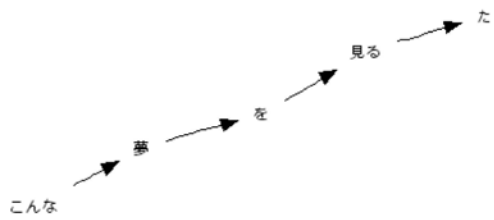

図１：文脈グラフの例

このような図を以下，「文脈グラフ」といい，語の部分を「ノード」，語と語を結んでいる矢印を「エッジ」という。

3.問題点

日本語文献から文脈グラフを抽出する場合，ノードの抽出法を中心に主に3つの問題がある。

3.1 機能語と内容語

機能語とは文法上必要な語のことで，直接文の意味を表すことはないが，非常に頻度が高い語である。日本語では助詞・助動詞などがこれにあたる。これらの機能語をグラフのノードとすると，そこにエッジが集中し，文脈グラフが複雑になりすぎ，視覚化した意味がなくなってしまう。したがって特に理由がない場合，助詞や助動詞をノードにしないことが賢明であろう。

一方，内容語は文の内容に関わるもので，名詞・動詞・形容詞などがこれにあたる。頻度は文献の内容によってさまざまである。これらはノードに採用しなくてはならない。

3.2 内容語の活用

日本語の動詞や形容詞には活用があるので，文献に用いられている表層語のレベルでは，文字処理上，同定が複雑になる。これをインデキシングし，索引語[9]にするため

には，基本形（辞書形）になおす必要がある。基本形は形態素解析ソフトの機能を利用すれば得られるが，結果を読むとき，ノードになっているのは，原文に内在する表層語ではなく，インデキシングされ抽象化された索引語である点には注意をしなければならない。

3.3 分析語彙の限定

文献は多量の語を含むのが一般的であるから，文献全体をそのまま分析するのは困難をきわめる。文献全体をグラフ化することもできるが，それは文献の複雑な構造をそのまま反映し，きわめて複雑なグラフになる。そこから，有意義な構造だけを取り出すことも不可能ではないが，それよりも有効なのは分析語彙を限定することである。

分析語彙は分析を開始するための基準点であり，出力される文脈グラフには分析語彙しか含まないということにはならない。分析語彙をふくむ文脈を抽出することで，結果的には分析語彙以外の多くの語がノードになる。

このように分析語彙を外挿するのは，検索エンジンの入力クエリーの問題と同様に，結果が分析者の意図や興味に限定されるという側面がないではない。しかし，文脈グラフは，思わぬところで「つながり」を示すことがあり，また，はっきりとグループを形成することもある。したがって，この点は単なるキーワード検索からでは得られない構造を得ることができる。

また，分析語彙の限定には，ある程度文章の性質を考慮する必要がある。文学作品であれば，作者が比較的自由に使える形容詞に作品の特徴がでやすいと推測できる。したがって，形容詞をノードとすることは有効な分析方法となるだろう。しかし，新聞記事では，客観的な報道のため使用可能な語彙が決められており，とくに形容詞はまれである。したがって，新聞記事の構造を視覚化する場合には，名詞や動詞からノードを抽出せねばならない。

4.文脈グラフの描画過程

前節でのべた点をふまえ，本稿で行った文脈の視覚化過程を示すと，以下の通りである。

1 テキストデータを入手する。
2 テキストデータにあるノイズ（空白・括弧・ルビ・ダッシュなど）を削除する。なお，この段階では句読点は削除しない方がよい。以下の過程で文を越えた遷移を抽出してしまうことを防ぐためである。
3 前項の結果に形態素解析を実行する。
4 前項の結果から機能語（助詞や助動詞）を削除し，内容語の基本形及び句読点を抽出後，一文ごとに改行し，「分析用テキスト」を作成する。「分析用テキスト」では語の順序を壊さないようにし，

語の句切り情報として，dot 言語の関係記述子（"->"）付加しておくと，後の段階で語の句切り情報を変換する必要がない。

5　分析語彙を決定し，「分析用テキスト」から分析語彙を含むものだけを抽出する。ただし，「分析テキスト」が少量の場合は全体をグラフにしてから，分析語彙を決定することも可能である。

6　前項の結果から，dot 言語ファイルを記述し，GraphViz に読み込ませ，グラフを画像として出力する。

```
こんな->夢->見る;
腕組->する->枕元->坐る->いる;
仰る->向->寝る->女;
静か->声->もう->死ぬ->云う;
女->長い->髪->枕->數く;
輪郭->柔らか->瓜実顔->その->中->横たえる->いる;
真白->頬->底->温かい->血->色->ほどよい->差す;
唇->色->無論->赤い;
うつ->いる->死ぬ->そう->見える;
しかし->女->静か->声;
もう->死ぬ->判然->云う;
```

図 2：分析用テキストのデータベース（『夢十夜』第一夜冒頭[10]）

図 3：GraphViz 実行画面

5.実例

前節の過程をへて出力した文脈グラフの実例を示す。原文テキストの基礎統計は以下のとおり。

	品詞	語数	語種
内容語	名詞	2,696	918
	動詞	1,759	449
	形容詞	259	73
機能語	助詞	3,170	48
	助動詞	950	17
	副詞	307	147
	接続詞	131	22
	連体詞	116	8
	接頭詞	46	16
	未知語	44	33
	感動詞	17	12
	フィラー	3	2
	記号	1,245	13
	総計	10,743	1,758

表 1．夏目漱石『夢十夜』の語彙

原文テキストは約一万語で構成されており，1700 種の語彙で成り立っている。内容語の総計は 4714 語（43.8%），1440 種(81.9%)である。機能語のうち「フィラー」とは発話と発話の間をつなぐ時に挿入される語のことである。このうち，「第一夜」の部分をすべて文脈グラフにすると，次のようになる。

図 4：『夢十夜』第一夜の全体グラフ

図 4 で矢印が集まっている点は，頻度の高い動詞（「する」「来る」「いる」等）や，「女」や「自分」といった登場人物である。これらのうち，以下に登場人物と色彩語の関係をみてみる。

《分析語彙》

・色彩語：黒・真黒・黒い・白・真白・白い・赤い・青い

・登場人物：自分・女

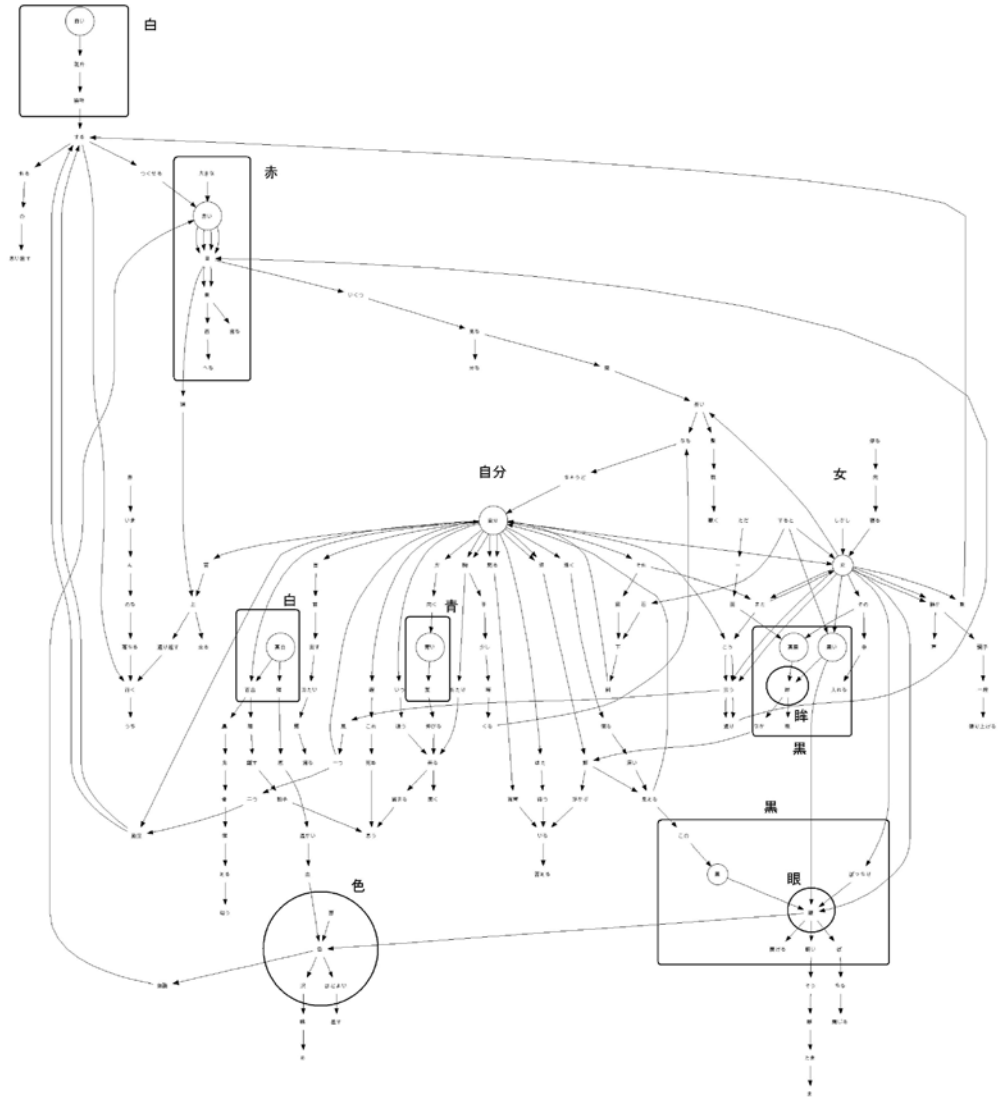

図 5：夏目漱石『夢十夜』第一夜における色彩語と登場人物の文脈グラフ

図５の細線円は「分析語彙」であり，それ以外のノードは文脈を形成している索引語である。太線部分はグラフを描画後に筆者が説明のために挿入した。中央部は「自分」がふくまれている文脈であり，中央右が「女」の文脈である。「黒」は「女」と関係がふかく，一部では「自分」の文脈と交差している。この文脈交差のようすを以下に示す。

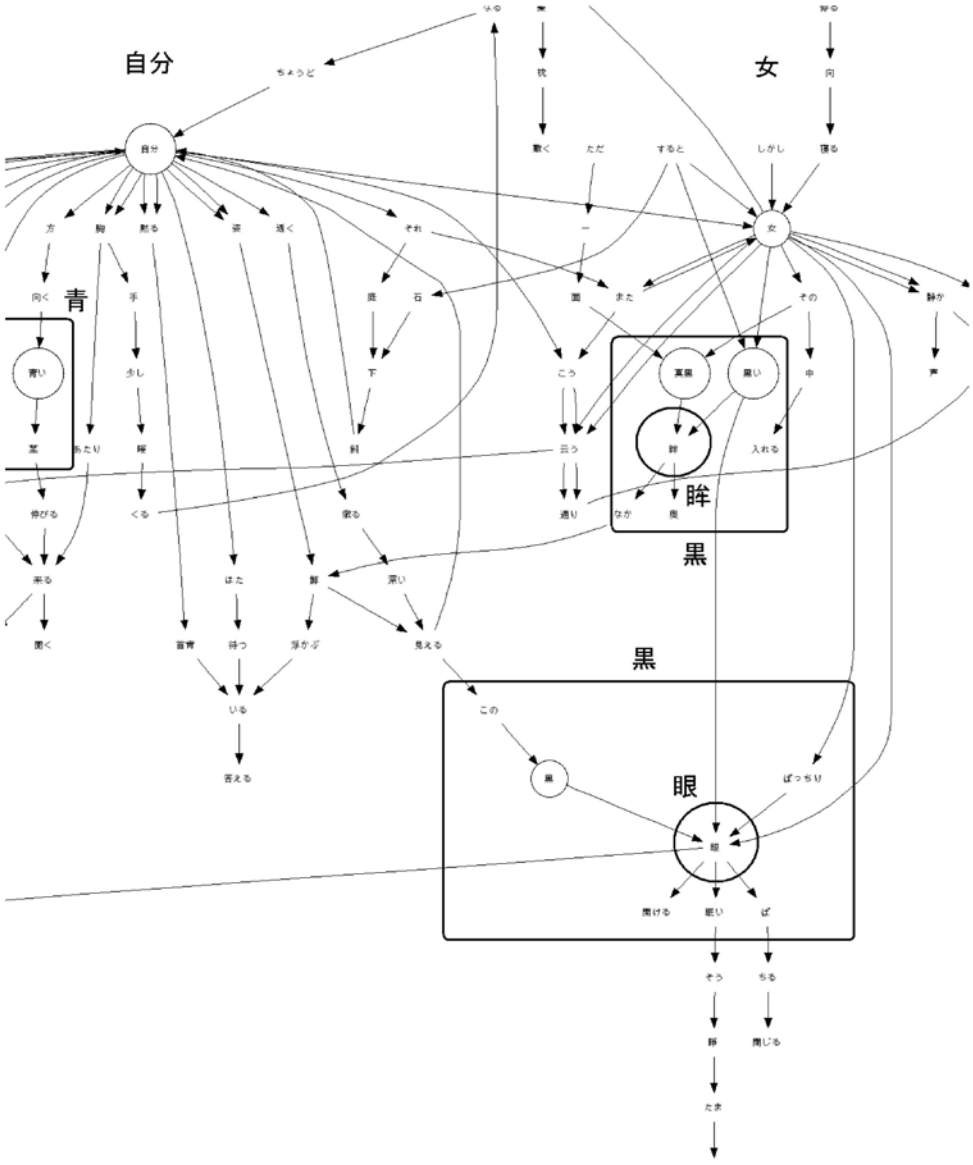

図６：文脈交差の例（図５の拡大図）

6.考察

図 6 では，「自分」をふくむ 8 語の文脈と，「女」をふくむ 3 語の文脈が，「黒」に連なる「眼」というノードで交差していることが確認できる。

原文を参照すると，この構造の源として，以下の二つの文をみつけることができる。

1　女はぱっちりと眼を開けた。大きな潤のある眼で，長い睫に包まれた中は，ただ一面に真黒であった。その真黒な眸の奥に，自分の姿が鮮に浮かんでいる。

2　自分は透き徹るほど深く見えるこの黒眼の色沢を眺めて，これでも死ぬのかと思った。

上の二つの文を，語の遷移関係であらわすと以下になる。

1’　女→ぱっちり→眼→開ける

2’　自分→透く→徹る→ほど→見える→この→黒→眼→色沢→眺める

この中で分析語彙と 1’・2’に共通するノード（交差ノード）は，つぎの 4 語である。

3　（分析語彙）女・自分・黒（共通語彙）眼

この 4 者の関係を有向グラフで示すと以下になる。

3’　女(1)→眼(1・2)←黒(2)←自分(2)

上の関係が文脈の交差点として，図 6 で表現されているのである。

以上，文脈交差を原文と比較した結果，注目すべき点は以下の 2 点であると思われる。

・1 の文には女の「眼」が「黒」であるとは書かれていない。「長い睫に包まれた中」が「真黒」であり，また「真黒な眸」という描写があるが，「眼」については，厳密には書かれていない。

・2 の文には，「黒眼」が誰のものか書かれていない。

もちろん，1 の文から，「女」の「眼」が「黒」であることは，人間の読解を導入すれば容易に把握することはできる。2 の文についても「黒眼」が「女」のものであることはすぐに把握できる。

だが，図 6 のネットワーク構造の抽出過程では，人間の読解を導入しなかった。また，「眼」と「眸」が同じ範疇に属する語であるという言語知識データベースも，ネットワーク抽出の過程に外挿しなかった。さ

らに，この小説には「自分」と「女」しか登場せず，したがって「自分」が「眺めた」のは「女」の「眼」にちがいないといった作品知識も導入されていない。

こうした，言語に関する知識や作品に関する知識を何も与えなくても，純粋に語彙の関係だけで，図6は「自分」と「女」を「眼」というノードによって結びつけている。

この構造は誰が行っても同様に再現する構造であり，その意味では人間の解釈を離れた客観的構造である。こうした客観的構造をふまえることは，テキストの分析を刺激するツールとなるだろう。客観的構造からいえば，図6において，「自分」と「女」を結びつける「眼」という語の存在を無視できない。

この文脈の客観的構造から意味をくみとるのは，人間の読解力が導入されるべき分野であり，その意味では，視覚化された文脈も変形されたテキストにすぎない。

7.おわりに

以上，最も基本的な意味における「文脈」の視覚化について，その方法を紹介した。

このような処理をコンピュータで行う場合の利点として，一般に網羅性があげられるであろう。網羅性は本稿でみたような小品の分析では，ありがたみを感じられないかもしれないが，大きな文献の分析では威力を発揮する。網羅性はうらをかえせば多量のノイズを出力する根源でもあるが，それをうまく除去できれば，見落とされていた事項を発見することができるという可能性そのものでもある。この可能性を掘りおこす手法，すなわち「データマイニング」の点では，まだ問題は残っている。たとえば，頻度の高い動詞（「する」「いる」）を文脈の交差点とする部分は，多くのエッジが集中するが，人間の読解ではふるいにかけられる部分であり，他の構造の理解を妨げるノイズとなる。こうした問題はさまざな文献で実験し，その解消方法をさぐっていく必要があるだろう。

本稿では文学作品を扱ったが，たとえば町内会誌やある一定期間の新聞などに，文脈の視覚化の手法を導入すれば，ある話題がどのような語を中心に語られ，またどのような話題を表現する語へ結びついていくのかという構造も視覚化することが可能である。この構造を単純化できれば，話題のマップを作成することも可能となるだろう。

また，語のネットワーク構造は行列で表現することも可能である。カリフォルニア大学アーバイン校のフリーマンらが開発したUCINET（ネットワーク分析ソフトウェア）では，ネットワークを行列で表現する方式をとっている。このようなネットワーク分析ソフトウェアは，単にネットワークの構造を視覚化できるだけでなく，各ノードにおけるさまざまな中心性や，ノードからノードへの到達可能性なども求めることができる。このような機能をつかえば，話

題を表現する中心的な語はどれなのか，また話題の内部でどのような語の住み分けがなされているのかといった問題も追究可能である。

こうしたネットワークの特性の部分については，本稿では扱えなかったが，いずれにしろ形態素解析とネットワーク分析ソフトウェアの組合せで実現できる処理である。語と語のネットワークの分析はこうしたさまざまな可能性を秘めているのである。

本稿で紹介した手法は非常に素朴な方法であるが，さしあたり，まずは読解を刺激するツールとして，このような視覚化を行う意味もあると思われる。

注・文献

- 安田雪『ネットワーク分析』新曜社 1997 年
- 土橋喜『情報視覚化と問題発見支援』あるむ 2000 年
- 中尾浩・宮川進悟・赤瀬川史朗『コーパス言語学の技法〈1〉テキスト処理入門』夏目書房 2002 年
- 伊藤雅光『計量言語学入門』大修館書店 2002 年
- 辻幸夫『認知言語学への招待』大修館書店 2003 年

1 漢文における試みは、拙稿「『老子』の聖人と玄徳」(『漢字文献情報処理研究』第 8 号好文出版 2007) を参照。古典漢文の場合はコンピュータによって、形態素解析を行うことが困難なので、Ngram インデキシングを用い、頻度の高い 4 字句を起点とし、原文における 4 字句遷移をとりだした。

2 「遷移」は確率文法の書物にでてくる。詳細は、北研二『確率的言語モデル』東京電機大学出版会 1999 を参照されたい。本稿では、「遷移」とはただ前の語に後ろの語が続くことを指す。

3 著作権保護期間の問題、および分量の適切さなどを考慮し、夏目漱石『夢十夜』を選んだ。テキストデータについては、「青空文庫」(http://www.aozora.gr.jp) で公開されているものを使用し、ルビは GREP をつかい削除した。

4 『広辞苑』岩波書店第 5 版

5 いわゆる「ボリンジャーの法則」とよばれるもの。詳細は参考文献、辻幸夫 2000 を参照。

6 奈良先端科学技術大学院大学が開発した形態素解析ソフトウェア「茶筌」は、http://chasen-legacy.sourceforge.jp/から配布されている。なお本稿では形態素解析に使用した辞書はデフォルトのままとした。

7 AT&T が開発したネットワーク視覚化ソフト Graphviz は以下から配布されている。

http://www.graphviz.org/

8 dot 言語の詳細については、2007 年現在、いくつか解説サイトが存在している。その中では以下が網羅的である。

http://homepage3.nifty.com/kaku-chan/graphviz/

9 「使用語」と「索引語」については、参考文献、伊藤雅光 2002 を参照。

10 3 行目、「仰る->向->寝る->女」の部分は、「仰向」(あおむき) を誤って形態素解析した例である。本来は「仰ぐ->向」と解析すべきだろうが、デフォルトの辞書を使用したためこのようになった。形態素解析の精度の問題は、Chasen の辞書をカスタマイズすることで対応可能である。