テキストマイニングを用いた小説の教材化
CREATING TEACHING MATERIALS FROM NOVELS IN JAPANESE USING TEXT MINING

王　伸子, 専修大学
Nobuko Wang, Senshu University

1. はじめに

本研究は、テキストマイニングの手法を用いて小説等の文章を分析し、読解の授業で用いるという試みを行った実践報告の一部である。分析に用いたソフトは KH-Coder で、形態素分析をおこない、それを応用して小説から頻出語彙を取り出して語彙表を作成し、小説を読む前に語彙表を学習者に提示し、学習者の日本語のレベルによって教師が工夫できるようにするということを試みた。

学習者のレベルは、今回は日本の大学の学部に在籍する上級者を対象におこなったため、上級を前提としているが、今後の取り組みでは初級段階でも使用できるようにしたいと考えている。この取り組みは、初級学習者であるから簡単なものしか読めない、簡単なものを教材として使用するという既存の考えを離れ、成人学習者には読みたいものを読ませたい、ということを到達目標として考えたものである。2004 年から 2005 年に筆者はニューブランズウィック州のマウントアリソン大学（Mount Allison University）で大江都先生の授業を拝見した。Creative Writing を掲げ、初級学習者であっても、書きたいものを書かせたい、言いたいことを言わせたいというコンセプトでスピーチコンテストに学生を送り出す授業の取り組みと、その指導を受けた学習者の書き上げた日本語に感銘を受け、日本でも同様のコンセプトで、「読む」指導をおこないたいと考えた。そう思いながらなかなか具体的には取り組めなかったが、数年前から、大学に入学した留学生に対して、専門書だけでなく、日本語で読書を楽しもうという授業をいくつかおこなってきた。昨年、形態素分析等も含め、大量のデータを分析できる KH-Coder というソフトを知り、小説も含め、日本語の文章をさまざまな側面から分析できる可能性があることがわかったので、それを利用し、研究だけでなく、教材作成にも取り組んでみた。そうしたことを教材作成という側面から報告したい。

2. 研究の背景

2.1 授業での取り組み

数年前から、「一般日本事情」科目において、専門書以外の読書をして発表するという課題を与え、授業を進めてきた。評論や実用書ではなく「小説」を読むことを課題としたが、そこでわかったのが、留学生が読みたい小説は、日本語の中上級学習者向け授業でもしばしば取り上げられる、夏目漱石や川端康成、芥川龍之介などの定番の名作ではなく、現代作家の東野圭吾や村上春樹、三浦しをん等の作品だということである。その理由は、それらの作家の作品というのは、英語、中国語、韓国語などに翻訳され、学生の母国でも多く読まれているということと、ドラマや映画等で取り上げられている作品でもあるので、日本に来てから

は、テレビドラマやネット上で簡単に閲覧できるので、そうした作品を読みたいと希望する学習者が圧倒的に多いということがわかった。本研究でもそうした一連の作品の中から、教師自身が読んでみて面白く、ストーリー性がある東野圭吾の作品を取り上げることとした。

2.2 学習者への調査と分析

学習者に対し、日本語で読書をするかどうかを調査した結果、文学部日本文学文化学科の留学生以外は、ほとんど読書をしないということがわかった。日本語ではしない、という学習者と、母語でもほとんど読書はしない、という学習者がいることも判明したが、本来は読書が好きなのに、日本語ではほとんど読書をしないのはなぜかを質問した結果、以下のような回答が得られた。学習者のグループとして、漢字圏出身の学習者と、非漢字圏出身の学習者では回答が異なっている。

今回の場合、漢字圏出身者の出身国は、中国、韓国であり、非漢字圏出身者の出身国は、ベトナム、ラオス、カンボジア、カナダである。

漢字圏出身者の理由：

① 日本語の書籍は縦書きなので読みにくい。何度も同じ行を読んでしまいそうになる。（母語の書籍は、現在では横書き）
② 人名がわかりにくい。読みながら覚えられないのでストーリーが頭に入ってこない。どの物語も、犯人はすべて「田中」、主人公は「山田」などであれば区別しやすいが。

非漢字圏出身者の理由：

① とにかく、漢字が多すぎて、新しい語彙もなにもかも読む気が失せる。
② 縦書きは読みにくい。

等々であった。漢字圏出身の学習者のうち、中国語を母語とする学習者の場合、全部読んでも話が覚えられないという者もいるが、これはおそらく漢字の語句を目で追い、ひらがなは目に入っては来るが、きちんと読んでいないために、モダリティなどを表す語が用いられた場合、きちんと読んでいないからではないかと推測される。

つまり、いずれにしても頻出語彙、あるいは品詞別の語彙表を与えることにより、読み仮名、意味、語句等を事前に調べ、当該小説をカバーする語句を頭に入れることにより、ボトムアップ方式で読んでいっても、内容理解が進むと期待される。

また、品詞別により分類できれば、人名を予め把握することが可能になるので、メモも取りやすくなると考えられる。

以上のことから、今回は、有効な語彙表を作成し、事前に与えて読ませるという試みをおこなった。また、語彙表の作成には、形態素分析機能を持つ、テキストマイニングのためのソフト、KH-Coder を使用した。

3. 先行研究

いわゆる「読書」が理解力を向上させるというのは、決まりきったことのように思われるが、もっとも影響するのはどのくらい語彙を知っているか、つまり既知の語彙数によると考えられる。しかし、既知の語彙数が多いと理解力が進むのか、読書をすることによって語彙数が増えるのか研究するためにはその因果関係にも着目しなければならない。その観点から先行研究にあたると、猪原（2014）では、語彙について、受容語彙（receptive vocabulary）と表出語彙（productive vocabulary）に分け、受容語彙の獲得という点から述べている。また、その際、読書量が増えることによって語彙力が増え、それに基づく文章理解力が向上するのか、語彙力があるから、多くの読書をする力があるとするのかについては研究が必要であるとしている。

一方、文章の体裁であるが、これについては学習者からは横書きではない日本語を読むのが負担であるという声も聴取している。神長（2014）は、これについては、分を読むのに要する眼球運動は、その停留時間が短ければ、相対的な時間は短くなるとしている。眼球の移動をサッカード（saccade）と言い、その間に生じる停留時間は、新密度の低い語、文字数の多い語ほど長くなるとしている。また、サッカードは横書きのほうが速いという実験結果も出ているということで、学習者の言う縦書きが読みにくい、ということも、これまでの読書週間の違いだとは一概には言えないということになろう。

4.「読み」についての考察

いずれにしても、文章理解力のためには、既知語彙の数が大きく関係しているということは否定できない。だからこそ、漢字圏出身者は母語と目標言語に共通する漢字語彙があると、学習する前から認識可能な既知語彙数あるため、非漢字圏出身の学習者に比べ、文章の認知と理解力は圧倒的に高くなる。口頭表現として用いる場合には、発音等も関係するので、運用能力が全般的に高くなるとは言えないが、認知に関しては有利である。漢字圏出身学習者にとっては、その分、共通する語であっても 100％意味が同じでないものもあるので、その点は、しっかり学習しなければ本当の理解力の向上にはならないので、教師も学習者もこの点については知っておかなければならない。

また、知らない語が多いと「読み」が進まず、速度が落ちるということになり、非漢字圏出身の学習者は、同じページ数でも読書にかかる時間が長くなってしまうということになろう。漢字圏出身の学習者の場合、既知の漢字語彙が多いということは、逆に、読みの負担を無意識的に減らすため、漢字語彙から漢字語彙へとサッカードがおこり、語に後続する助詞等はあまり認識せずに読書を進め、正確な理解ができない場合も往々にしてあると言えよう。中国語を母語とする学習者が、普段でも極端に助詞の習得が劣っていると感じるのは、こうしたところに原因があるとも考えられる。

以上のようなことを考えると、読書を進めるにあたっては、まず、語彙を意識させるということが、漢字圏、非漢字圏出身にかかわらず、学習者にとって必要なことであると確認できた。

5. テキストマイニングによる語彙表の作成

5.1 教材対象作品の選定

今回は、前述の東野圭吾著『手紙』を取り上げ、語意表を作成した。この前年、同書を語彙表等を作らず上級クラスで読んだが、漢字圏学習者と非漢字圏学習者の読みのスピードがかなり異なったため、非漢字圏出身の学習者が内容の確認やディスカッションへの参加が難しく、読書がかなり苦しいものになってしまったので、今回は、再度同じ作品を対象とし、分析ソフトの力を借りた補助教材の効果を見ることとした。

5.2 作成作業

まず、本文をスキャナで取り込み、それを KH-Coder で分析するためにテキストファイルに変換した。スキャナは富士通の ScanSnap を使用し、取り込んだものを OCR 変換ソフト、e-typist で取り込み、さらにフリーソフト TeraPad によってテキストファイルを編集したものを KH-Coder にかけて語彙の分析をおこなった。品詞別語彙表を作成するためには、形態素分析が必要であるが、これは、KH-Coder に同梱されているソフト「茶筅」（wincha）で行い、相関図などを作成するためには、同じく、KH-Coder 同梱の統計ソフト「R」を使用する。

KH-Coder で分析するさい、データにタグを付けることによって、章別、節別にも語彙が取り出せるので、OCR データからテキストファイルにするときに識別タグを付す作業もおこなう。

以上の作業によって、以下の語リストの抽出がおこなえる。

① 全章 / 章別 いずれかの頻出語彙 150 語
② 全章 / 章別いずれかの品詞別語彙リスト
③ 章別特徴語彙各 10 語

① の頻出語彙 150 語は以下のように抽出される。

表 1　頻出語 150 語リストの一部

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	抽出語	出現回数		抽出語	出現回数		抽出語	出現回数		抽出語	出現回数		抽出語	出現回数
2	直	1246		人	86		兄さん	56		立つ	42		いつ	34
3	思う	480		私	84		前	56		意味	41		違い	34
4	自分	289		男	83		送る	55		一緒	41		下げる	34
5	見る	270		時	80		朝美	54		閉じる	41		課長	34
6	由実子	234		緒	87		今日	53		子供	41		見せる	34
7	剛志	226		兄	86		事件	52		心	41		見つめる	34
8	顔	195		仕事	84		作る	51		最後	41		公園	34
9	目	176		話す	80		動く	51		待つ	41		好き	34
10	今	166		兄貴	79		窓	51		一つ	40		初めて	34
11	手紙	163		出す	78		ドア	50		読む	40		切る	34
12	入る	163		言える	78		貴雪	50		軍隊	40		店	34
13	出る	158		帰る	76		寺尾	50		警察	39		機	33
14	前	154		見える	76		表情	50		中条	39		歌う	33
15	行く	153		持つ	75		違う	49		信長	39		座る	33
16	知る	149		実紀	74		考え	49		刺事	38		質問	33
17	書く	136		少し	74		読む	49		利恵所	38		大きい	33
18	手	123		店	74		ほか	48		変わる	38		理由	33
19	話	122		言	73		使う	48		開ける	37		連絡	33
20	考える	120		相手	70		気づく	47		関係	37		変える	32

② の品詞別語彙リストは以下のように抽出される。

表２　品詞別語彙リストの一部

名詞		サ変名詞		形容動詞		固有名詞		組織名		人名	
自分	42	話	29	嫌	7	剛志	17	イマジン	3	貴	223
手紙	37	電話	12	急	6	寺尾	5	毎日	2	由実子	108
公園	24	結婚	9	大変	6	前山	2	サン	1	実紀	74
言葉	23	意味	8	元気	5	ディズニーランド	1	安藤	1	緒方	28
平野	23	返事	8	好き	5	葛西	1	新星	1	町谷	19

③ の章別特徴語彙 10 語は以下のように抽出される。

表３　章別特徴語彙 10 語

	序章		第一章		第二章		第三章	
3	剛志	.064	貴	.072	貴	.077	貴	.085
4	窓	.027	剛志	.046	寺尾	.029	朝美	.039
5	老婦	.024	思う	.033	倉田	.028	思う	.039
6	入る	.023	兄	.022	自分	.026	自分	.026
7	家	.022	店長	.019	音楽	.019	見る	.022
8	手	.020	教諭	.018	大学	.017	朝美	.018
9	金	.020	顔	.017	入る	.017	李文	.017
10	緒方	.017	声	.017	寺尾	.016	顔	.016
11	ドライバー	.017	梅村	.016	顔	.016	話す	.014
12	引っ越し	.017	目	.014	コータ	.015	前	.014
13	第四章		第五章		終章			
14	思う	.036	貴	.071	手紙	.049		
15	由実子	.029	由実子	.048	弟	.027		
16	手紙	.027	思う	.038	兄貴	.025		
17	書く	.024	実紀	.035	今日	.024		
18	自分	.022	見る	.023	読む	.021		
19	会社	.021	今	.017	寺尾	.021		
20	知る	.021	手紙	.017	観客	.020		
21	人間	.016	目	.015	イマジン	.020		
22	社長	.016	書く	.014	貴	.020		
23	行く	.015	話	.013	外	.019		

5.3 授業で用いる教材として

以上のような表から、必要なものを加工し、教材として使用した。上記②の品詞別語彙リストを加工し、語彙数を任意の数に制限し、表４のように、品詞も、形容詞→イ形容詞、形容動詞→ナ形容詞と修正し、表中の品詞の場所も学習者が見やすいと思われる順番に配置した上で、メモを書き込める箇所をエクセル上で増やし、学習者が、新出の語、読み方がわからない漢字語等に意味を書き込んだりふりがなをふったり、人名には、小説の中での役割を書き込んだりするのに使用できるようにした。また、上級学習者の場合は、知らない語にのみ書き込みをすれば、自分が把握していなかった語が一目で区別がつく。次の表４は、終章を例とした教材用語彙表の一部である。

表４　教材用語彙表（終章）

名詞			人名			イ形容詞			ナ形容動詞		
手紙	19		貴	29		激しい	2		気楽	2	
兄貴	8		寺尾	8		小さい	2		結構	2	
観客	5		緒方	6		暗い	1		からから	1	
自分	5		サー	1		楽しい	1		ショック	1	
ライブ	4		ジョン	1		嬉しい	1		阿呆	1	
刑務	4		震	1		極まりない	1		異質	1	
コンサート	3		忠夫	1		近い	1		健康	1	
ユニット	3		由実子	1		細かい	1		幸せ	1	
会場	3					若い	1		困難	1	
思い	3					少ない	1		自分勝手	1	
自己	3					情けない	1		重大	1	
人間	3					深い	1		大丈夫	1	
グラウンド	2					辛い	1		適度	1	
マイク	2					大きい	1		必要	1	
灰色	2					熱い	1		不安	1	
気持ち	2								不快	1	
気配	2								無事	1	
客席	2								陽気	1	
言葉	2								冷酷	1	
控え室	2										

6. おわりに

今回は、KH-Coder の機能のうち、形態素抽出機能を利用した品詞別語彙の抽出表を加工して授業の教材作成をする試みについて記した。実際に、授業を進める中で、どのようにディスカッションに発展させたらよいかということを考えて臨んだが、学生たちが積極的に取り組み、小説の場面ごとに感想を述べながら、お互いに意見交換を進め、案じていたより授業が順調に展開していった。例えば、『手紙』のストーリーの登場人物について、主人公を応援してくれる高校の担任教師に焦点を当て、本当は関わりたくないのに役目上、しかたなく接しているのではないか、とか、よく考えてみると、この話を通じて、ほとんどの人が犯罪者の家族をかわいそうだと思いながら、関わりを持つのをさけているようで歯がゆい、などと感想を述べ合ったりするなど、思った以上に内容に没頭して読み進めていたようであった。

今後は、素材として長編ではなく、もう少し短いものを選び、教材化を試みたいと考えている。また、KH-Coder 自体も、まだ使っていない機能、加工できる部分もあるので、それについても使い方に習熟し、別の試みもしてみたい。KH-Coder はすでにテキストファイル化されている「青空文庫」の作品なら、簡単に扱えるので、そうした点も鑑み、新しい語意表作成も試みたい。

参考文献

猪原敬介（2014）「第 4 章　読書からの語彙獲得」『文章理解の認知心理学』誠信書房

河元由美子（1998）「読解で長編小説を読む―速読から鑑賞へ―」『講座日本語教育第 34 分冊』　早稲田大学日本語教育研究センター

佐野香織、李在鎬（2007）「KH-Coder で何ができるか：日本語習得、日本語教育研究利用への示唆」『言語文化と日本語教育』33、御茶の水女子大学日本言語文化学研究会

樋口耕一 （2004） 「テキスト型データの計量的分析 ―2 つのアプローチの峻別と統合―」 『理論と方法』（数理社会学会）
樋口耕一（2014）『社会調査のための計量テキスト分析』ナカニシヤ出版
樋度 他（1983）「日本語文章に対する注視と認知」 『テレビジョン学会誌 Vol.37, No.11』
山川 他（2006）「縦書き・横書きの弁別と大脳半球差に関する脳磁場反応」『電子情報通信学会総合大会発表原稿集』
李在鎬、石川慎一郎、砂川有里子『日本語のためのコーパス調査入門』くろしお出版
T・マケナリー、A・ハーディ 石川慎一郎訳（2014）『概説コーパス言語学』ひつじ書房

教材用使用書籍
東野圭吾（2003）『手紙』（文春文庫）