現代中国語のテキスト処理について

——魯迅「故郷」を例に、ngram・形態素解析・正規表現を使って——

An Introduction to the Text Processing of Modern Chinese e-texts

非常勤講師　齊藤正高

1．中国語コーパスについて

「言語資料」という意味のコーパス(corpus)は、現代中国語では「語料庫」という。中華人民共和国では1970年代末から各大学で「語料庫」の構築が開始されている。その主なものを示すと、以下になる。

名　称	制作年	字数（万字）	事業主体
漢語現代文学作品語料庫	1979	527	武漢大学
現代漢語語料庫	1983	2000	北京航天航空大学
中学語文教材語料庫	1983	106	北京師範大学
現代漢語語料庫	1983	180	北京語言学院
漢語新聞語料庫	1988	250	山西大学
北大漢語語料庫	1992	500	北京大学

表1.中国語コーパス（馮志偉『計算語言学基礎』商務印書館2001 p.63-64）

これらの大部分はインターネットの普及以前に構築されたものであり、主に大学内で使用する運用方針をとり、2005年5月までの所いずれもインターネット上で公開されていない。

インターネットの普及後には文学作品の電子テキストを中心にしたサイトが構築されている。このようなサイトとして、以下が代表的である。

・亦凡公益図書館　http://www.shuku.net
・新語絲　http://www.xys.org

コーパスにはほかにも、口語資料の語料庫や、新聞記事に品詞情報を付加した「標注語料庫」などがあり、多様な広がりをみせている。

以上のさまざまなコーパスのなかで、とくにインターネット普及後に作成された文学作品の語料庫は基本的に万人に開かれた形で現代中国語の言語資料を提供している点で貴重であると思われる。

もしこれらのテキストデータを十分に活用できるなら、中国語の用例収集および教材作成などに有効であろう。以下では、その手続き・分析ツール・分析例を紹介したい。

2．　著作権

上にあげた「亦凡公益図書館」を閲覧して気づく点は、著作権保護期間中の作品もテキストデータとして公開している点である。巴金(1904-)や高行健(1940-)の作品も見られるし、なかには『ハリーポッター』の中国語版など

もある。これらのテキストデータは、著者の権利保護という点から大きな問題[1]をふくんでいる。最近、インターネット上の著作権侵害に関する国家版権局の実施規則[2]ができ、著作者がインターネットにおける著作権の侵犯を申し立てるための手続きが整備されたので、テキストデータを提供しているサイトが、データの掲載を自粛する傾向も生ずるかもしれない。しかし、一方で、アメリカのローレンス・レッシグらが提唱し、団体として活動をはじめている Creative Commons が、インターネットおける新たな著作利用許諾のスタイルを提示しようとしている。これはあくまでも著者に著作権の内容を選択する余地をのこすものであり、著作権そのものを放棄することではないが、新たなこころみといえるだろう。この Creative Commons（創作共用）は中国においても活動[3]を行っている。

このような動きのなかで、インターネットに公開されているコンテンツの著作権が今後どう変化するのかは予測が難しい。ともあれ、こうした文学サイトからテキストデータを利用する場合、その著者の権利保護期間を一応確認しておく必要がある。

中華人民共和国著作権法においては、著作権の保護期間は作者の死後 50 年であり、50 年目の 12 月 31 日に満了となる。サンフランシスコ対日講和条約における著作権保護期間の戦時加算は、条約に不参加であった中華人民共和国に対しては加算されない。また、日本も中国も「万国著作権条約パリ改正条約」に加盟しているので、中華人民共和国と同じ保護期間が日本においても適用される。したがって、日本で中華人民共和国の著作を用いる場合、著作者の死後 50 年を経過するまでは著作権は保護される。

以下に、「亦凡公益図書館」や「新語絲」にテキストデータが公開されている現代作家のうち、著作権の保護期間が経過したものの一部を示す。老舎（1899-1966）、茅盾（1896-1981）などの著作は、保護期間内であるから、インターネット上でテキストデータを公開することに問題がある。

もちろん、個人使用や引用などは著作権の制限項目に入る。したがって著者の許諾がなくても、引用することは自由である。だが、著作権侵犯のファイルの存在を前提にして、その文献のテキスト処理を語ることは、それ自体に問題がある。

本稿では、こうした問題を回避するため、著作の保護期間が満了した作家の作品を使ってテキスト処理の技法を紹介したい。

このような作品として魯迅の「故郷」を例にした。

作家名	生卒年	出身	サイト
魯迅	1881-1936	浙江	亦・新
叶紫	1910-1939	湖南	亦
穆時英	1912-1940	浙江	亦
許地山	1893-1941	台湾	亦
郁達夫	1896-1945	浙江	亦
章衣萍	1900-1947	安徽	亦
朱自清	1898-1948	江蘇	新

表 2.著作の保護期間が満了した現代作家

3． ngram による反復部分の抽出

3.1. 機械的分割

ngram とは、シャノンの提唱した概念で、ある一定の長さで文字列を切り取ったものをいう。とくにｎが１のものをユニグラム、２のものをバイグラム、ｎが３のものをトリグラムという。

例えば「豊橋市町畑町町畑」を長さ 2gram で分割すると表３の分割結果が得られる。こ

の例では、星印をつけた「町畑」だけが２度使われている。

ngram の特徴は語の認定を行わず、文を機械的に分割する点にある。これは分かち書きの習慣のない日本語や中国語にはある程度有効な分割方法である。分割された ngram のうち頻度の高いものはそれなりに意味のまとまりを示し、グラム数が多くなれば、単なる語彙ではなく、作者がその作品でくり返し用いる比較的長い言い回しを抽出することができるのである。

豊橋
　橘市
　　市町
　　町畑　*
　　町町
　　町畑　*

表 3.バイグラム分割の例

3.2. ngram 分割ツール

前節の ngram 分割を行うツールに、花園大学の師茂樹（もろ・しげき）氏が開発したフリーウェア、morogram がある。[4]このソフトはユニコードに対応しており、テキストを UTF-8 のエンコードで保存すれば、中国語の文献を ngram 分割できる。

操作は基本的にコマンドプロンプトで行い、主に以下のスイッチを使う。

□　--f=min,max
min に最小頻度の数値、max に最大頻度の数値（省略可能）を指定する。
□　--g=min,max
min に最小グラム数、max に最大グラム数（省略可能）
□　--p　区切り文字の削除

例えば、頻度が１回しかでてこない 4gram をきりだすには、コマンドプロンプトから、morogram をインストールしたフォルダーに切り替え、以下のようにコマンドを実行する。
moro –f=1,1 –g=4,4 分析 file 名>出力 file 名

3.3　2gram 分割の実例

つぎに魯迅「故郷」を ngram 分割し、集計を行った結果を示す。

2gram 分割については、以下の morogram のコマンドを実行した。

moro –f=2,0 –g=2,2 –p guxiang.txt > g2.txt

これにより出力された g2.txt を表計算ソフトで頻度順に並べ替えると、以下になる。

頻度	2gram	頻度	2gram
26	我们	12	只是
26	*闰土*	11	什么
25	没有	11	来了
20	母亲	10	东西
18	知道	10	故乡
16	了我*	9	到了
14	一个	9	*宏儿*
13	他的	9	是我*
13	我的	9	*水生*
13	时候	以下略	

表 4：魯迅「故郷」の 2gram 頻度（上位）

頻度２以上の 2gram はおよそ 540 種類ある。その上位においては、基本的に現代中国語のある一定の単位として認められるものである。ただし星印をつけた「了我」と「是我」は意味をもつまとまりではない。また、斜体で示した「闰土」「宏儿」「水生」は登場人物の固有名である。

3.4. 4gram 分割の実例

では、つぎに比較的ながい 4gram の分割を示す。2gram と比較すると、作品の内容に関係する表現が抽出できる。

実行したコマンドは以下のとおり。

moro –f=2,0 –g=4,4 –p guxiang.txt > g4.txt

これにより出力された g4.txt の出力を表計算ソフトで頻度順に並べ替えると、以下になる。

頻度	4gram
5	我们这里
4	我的母亲
4	没有见过
3	他的父亲
3	是我自己

表 5：魯迅「故郷」の 4gram 頻度（上位）

この中から第 1 位の「我们这里」を原文に即して検討してみると、以下になる。

1.（我们这里给人做工的分三种…）
2. 你夏天到我们这里来。
3. 我们这里是不算偷的。
4. （我们这里煮饭是烧稻草的…）
5. （这是我们这里养鸡的器具…）

5 例のうち、3 例は作者が括弧書きで故郷の様子を説明する部分に出てくる。2 と 3 の例はどちらも「闰土」のセリフである。直訳すれば「私どものところ」「わたしたちのところ」となる。故郷を語る上で重要な言い回しであり、第一位であるのも首肯できる。

4. 形態素解析による品詞情報の付加

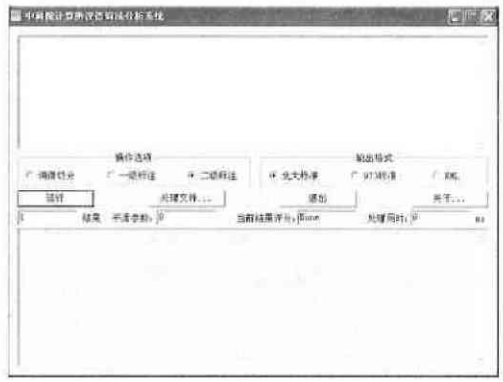
4.1.形態素解析ツール

形態素(morpheme)とは「意味をもつ最小の言語単位」(『広辞苑』) である。文を形態素に分けるには電子的に構成された辞書をつかい、大まかな構文の解析が必要となる。

中国語の形態素解析ソフトとしては、中国科学院計算技術研究所が制作した、ICTCLASS がある。[5]このソフトウェアはウェブ上でも公開されており、200 文字程度は解析ができる。提供されている中国語版 OS(中文 XP など)を搭載したコンピュータにインストールすれば、大きなファイルでも一括処理できる。その際、データを前もって中華人民共和国の文字コードである GB コードに変換する必要がある。

ICTCLASS の Readme ファイルによれば、形態素分割の精度は 97%に達し、品詞解析もある程度行うことができる。[6]

図 1:ICTCLASS の中文 OS 版　実行画面



4.2.分析例

ICTCLASS は、「故郷」の最後の文を以下のように「分詞」する。

（分析対象文）

这正如地上的路；其实地上本没有路，走的人多了，也便成了路。

（出力結果）

这/r 正/d 如/v 地上/s 的/u 路/n ；/w 其实/d 地上/s 本/d 没有/v 路/n ，/w 走/v 的/u 人/n 多/a 了/y ，/w 也/d 便/d 成/v 了/u 路/n 。/w

スラッシュ以下のアルファベットは前の形態素の品詞情報であり、つぎの略号を用いている。

印	品詞	印	品詞	印	品詞
ag	形语素	j	简称略语	r	代词
a	形容词	k	后接成分	s	处所词
ad	副形词	l	习用语	tg	时语素
an	名形词	m	数词	t	时间词
b	区别词	ng	名语素	u	助词
c	连词	n	名词	vg	动语素
dg	副语素	nr	人名	v	动词
d	副词	ns	地名	vd	副动词
e	叹词	nt	机构团体	vn	名动词
f	方位词	nz	其他专名	w	标点符号
g	语素	o	拟声词	x	非语素字
h	前接成分	p	介词	y	语气词
i	成语	q	量词	z	状态词

表６：ICTCLASS の品詞タグ（ICTCLASS 添付ファイル）

説明の必要な部分は以下にのべる。

- 副形词　→　直接「状語」（副詞性修飾語）を作る形容詞
- 名形词　→　名詞の機能をもつ形容詞
- 语素　→　多くの語素が合成語を作ることのできる「語根」のこと
- 副动词　→　直接「状語」を作る動詞
- 名动词　→　名詞の機能をもつ動詞

ICTCLASS をつかい品詞タグをつけたファイルは、品詞タグの後ろで改行するなどの処理を、ワードプロセッサでほどこした後、表計算ソフトで集計することができる。ただし、ICTCLASS はGB コードで品詞情報を付加するので、日本語アプリケーションにデータを移行するには、UNICODE か UTF-8 で保存しなおす必要がある。

4.3　分析結果の集計

「故郷」を ICTCLASS で品詞情報付加処理後に、集計処理を行った。その動詞部分を示す。

頻度	動詞	頻度	動詞
43	是/v	7	可以/v
34	去/v	7	起/v
34	说/v	7	起来/v
32	有/v	7	吃/v
18	知道/v	7	坐/v
18	到/v	7	如/v
13	来/v	6	站/v
12	出/v	6	卖/v
12	要/v	5	回/v
12	见/v	5	想/v
10	叫/v	5	像/v
9	看/v	5	拿/v
9	走/v	5	圆/vg
8	管/v	5	觉得/v
8	没有/v	以下省略	

表 7:魯迅「故郷」の動詞頻度（上位）

ICTCLASS の「分詞」（形態素分割）は 100%の精度で成功するわけではない。だから、出力ファイルをよくみれば、２文字の形態素を１文字づつに区切っている部分もある。魯迅の「故郷」は約 6000 字ほどであるから、ICTCLASS が公称する 97%の精度を出せたとしても残りの 3%の 171 文字ほどが誤って「分詞」される計算になる。

したがって、言語の精密な研究にはまだま

だ、人間が時間をかけて原文を分析しなければならない。しかし、大まかに分析の方向を探る程度には、形態素解析ソフトを用いる余地があると思われる。

ここで、得られた動詞の頻度は、ほかの作品と比べてみることも可能であり、その差異が本文の内容にもとづくのか、それとも別の理由（作者の用いる用語の変化など）によるものなのかを考察することも可能である。

このような分野として、計量文献学や計量文体学があり、日本語においては計量国語学会を中心に蓄積がある。中国語の計量文体学については調査中であり、べつに論じたい。

5．正規表現によるパターン抽出

5.1.正規表現とそのツール

正規表現(regular expression)は、パターン一致を行う場合に使われる半角文字であり、以下がその代表的な記号である。

半角記号	機能
.	任意の１文字
+	直前の１回以上の反復
*	直前の０回以上の反復
[～]	～のいずれか
[^～]	～のいずれも含まない
（ ）	カッコ内をひとまとめに
|	左辺か右辺かどちらか
＼1	１番目の一致と同じパターン
＼2	２番目の一致と同じパターン

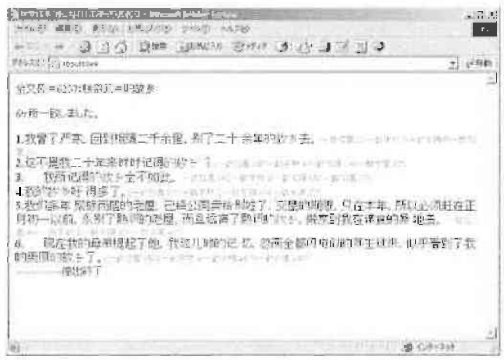
表 8:常用正規表現（一部）

正規表現を使えるツールとしては、秀丸やEmEditorなどに内蔵されているGREP(抽出プログラム）があり、perl や JavaScript、VBScriptなどのスクリプト言語、SQLでも使用できる。いずれも、中国語を使用する場合はユニコードに対応していなければならない。

筆者はVBScriptを使い、正規表現の練習用として、ウェブ上でうごく“grep on the web”[7]を作成した。このスクリプトは、テキストデータを貼り込み、別のウィンドウに結果を抽出する方式であるから、複数のファイルから一度の操作で用例を抽出することはできない。しかし、一度に一本の小説を全編貼り付けて動作させることもできる。[8]そのほかに、最も近い句読点を計算して、一文や一節を取り出す機能もつけてある。

図２：grep on the web 実行画面

図３：grep on the web の抽出画面



5.2.正規表現による中国語のパターン抽出例

上記の「grep on the web」をつかい、正規表現によって「故郷」の表現パターンを抽出した。以下に実例を示す。

例1)　正規表現：　**我.了**
(日本語訳)「我」の後ろに何か1文字あって、その後ろに「了」があるパターンに一致せよ。
(一致例)
- □ 我冒了严寒
- □ 第二日清早晨我到了我家的门口了。

例2)　正規表現：　**我..了**
(日本語訳)「我」の後ろに何か2文字あって、そのうしろに「了」があるパターンに一致せよ。
(一致例)
- □ 我愕然了。
- □ 哦，我记得了。

例3)　正規表現：　**我.+了**
(日本語訳)「我」の後ろに1文字以上の文字あって、その後ろに「了」があるパターンに一致せよ。
(一致例)
- □ 我冒了严寒，回到相隔二千余里，别了二十余年的故乡去。
- □ 我所记得的故乡全不如此。我的故乡好得多了。

例4)　正規表現：　**我.+?了**
(日本語訳)「我」の後ろに1文字以上の何か文字あって、その後ろに「了」があるパターンに最小範囲で一致せよ。
(注記) ?は最小範囲を表す。
(一致例)
- □ 我冒了严寒，回到相隔二千余里

例5)　正規表現：　**来.*了**
(日本語訳)「来」の後ろに0文字以上の文字あって「了」があるパターンに一致せよ。
(注記)来と了の間に文字が入っていなくてもよい。
(一致例)
- □ 我的心禁不住悲凉起来了。
- □ 但要我记起他的美丽，说出他的佳处来，却又没有影像，没有言辞了

例6)　正規表現：　**我.+[了过着]**
(日本語訳)「我」の後ろに1文字以上の文字あって「了」か「过」か「着」のいずれかがあるパターンに一致せよ。
(一致例)
- □ 我吃了一吓
- □ 我还抱过你咧
- □ 宏儿和我靠着船窗

例7) 正規表現：　**又[^。，]+了**
(日本語訳)「又」の「。」と「，」以外の文字が1回以上あり、その後ろに了があるパターンに一致せよ。
(一致例)
- □ 天气又阴晦了
- □ 这些人又来了

例8)　正規表現：**(我的|他的)(母亲|父亲)**
(日本語訳)「我的」あるいは「他的」があり、その後ろにすぐ「母亲」あるいは「父亲」があるパターンに一致せよ。
(一致例)
- □ 我的母亲很高兴
- □ 我的父亲允许了
- □ 所以他的父亲叫他闰土

例9)　正規表現：　**(.)\1**
(日本語訳) 1文字目に何か1文字があり、そ

の１文字目と同じ文字がすぐうしろにつづくパターンに一致せよ。
(注記)「\」はコンピュータの言語環境により、「¥」（半角）で同様に機能する。
(一致例)

□ 冷风吹进船舱中，鸣鸣的响
□ 这不是我二十年来时时记得的故乡？

例10） 正規表現： **(.)(.)\1\2**
(日本語訳) １文字目に何か１文字があり、２文字目にも何か１文字があり、3 文字目は１文字目と同じ文字であり、4 文字目は２文字目と同じ文字であるパターンに一致せよ。
(注記)「\」はコンピュータの言語環境により、「¥」（半角）でも同様に機能する。
(一致例)

□ 阿呀阿呀，真是愈有钱，便愈是一毫不肯放松，

とくに、「**\1**」や「**\2**」は、沈国威氏がすでに指摘しているように形容詞のくり返し型である ABAB 型や ABB 型などの抽出に用いることができる。

5.3 品詞情報と正規表現の組合せ検索

4 節に紹介した ICTCLASS により、形態素分割処理をほどこしたファイルを、正規表現で検索することによって、品詞の組合せ検索ができる。

例11） 介詞＋方位詞

（正規表現）/p.*/f
（日本語訳）/p「介詞」（前置詞）のタグがあり、その後ろに 0 文字以上の空白があり、その後ろに/f「方位詞」のタグがあるパターンに一致せよ。
（一致例）

□ 从/p 蓬/v 隙/ng 向/p 外/f 一/m 望/v
□ 我/r 先前/t 单/d 知道/v 他/r 在/p 水果店/ns 里/f 出卖/v 罢了/y
□ 自从/p 我家/n 收拾/v 行李/n 以来/f

例12） 数詞＋量詞＋名詞

（正規表現）/m[^/]+?/q[^/]+?/n
（日本語訳）
/m「数詞」のあとに/以外（つまり他の品詞タグなし）の文字がつづき、最小範囲で/q「量詞」があり、同様に「/」以外の文字があって/n「名詞」へつづくパターンへ一致せよ。
（一致例）

□ 又/d 买/v 了/u 几/m 件/q 家具/n
□ 手/n 捏/v 一/m 柄/q 钢叉/n
□ 我/r 扫/v 出/v 一/m 块/q 空 地/n 来/f
□ 都/d 有/v 青蛙/n 似的/u 两 /m 个/q 脚/n
□ 总/d 要/v 捐/v 几/m 回/q 钱/n

例 13） 接続詞の連鎖

（正規表現）/c.*/c
（日本語訳）/c「接続詞」の後ろに/c「接続詞」があるパターンに一致せよ。
（一致例）

□ 时候/n 既然/c 是/v 深冬/t ；/w 渐/d 近/a 故乡/n 时/ng ，/w 天气/n 又/c 阴晦 了 ，/
□ 我/r 的/u 母亲/n 很/d 高兴/a ，/w 但/c 也/d 藏/v 着/u 许多/m 凄凉/a 的/u 神情/n ，/w 教/v 我/r 坐下/v ，/w 歇息/v ，/w 喝茶/v ，/w 且/c 不/d 谈/v 搬家/v 的/u 事/n 。

例 14）動詞の後の名詞

（正規表現）看[^。，]*/n

（日本語訳）看（みるという動詞、結果補語がつくものも含む）のあとに句点以外の文字が続き、うしろに/n、つまり名詞があるパターンに一致せよ。

（一致例）

- 看/v 鸟雀/n 来/f 吃/v 时/ng
- 同/p 看/v 外面/f 模糊/a 的/u 风景/n
- 我/r 只/d 觉得/v 我/r 四面/f 有/v 看/v 不/d 见/v 的/u 高/a 墙/n
- 他们/r 都/d 和/c 我/r 一样/u 只/d 看/v 见/v 院子/n 里/f 高/a 墙上/s 的/u 四/m 角/q 的/u 天空/n 。
- 现在/t 我/r 的/u 母亲/n 提起/v 了/u 他/r ，/w 我/r 这/r 儿时/t 的/u 记忆/n ，/w 忽而/d 全都/d 闪电/v 似的/u 苏/j 生/v 过/u 来/f ，/w 似乎/d 看到/v 了/u 我/r 的/u 美丽/a 的/u 故乡/n 了/y

6．まとめ

以上、中国語のテキストデータを処理するための様々な方法を紹介した。しかし、ここで紹介した方法には、まだ問題点が多い。

インターネット上の著作権の問題は管理規則ができたばかりであり、ngram は形態素を認知できないという点で弱点があり、形態素分割は精度に問題がある。正規表現は実装している環境によって表現がことなり、また実際に正規表現で検索や抽出を行うには慣れが必要である。本稿で紹介した例も、とくに形態素解析の部分に、コンピュータによるテキスト処理の現状での限界を示している例も少なくない。

だが、様々な問題点を差し引いても、テキストじたいから、語彙や表現に関する統計的情報を取得できる点、また、文法的構造を抽出できる点などは、その限界を見極めたうえで使用するならば、中国語研究の粗調べに役立つ点もあろうかと思う。

（了）

参考文献

- 沈国威『電脳による中国語研究のススメ』白帝社 2000 年（ただし、添付 CD 所収データの著作権問題で版元切れ）
- 漢字文献情報処理研究会編『漢字文献情報処理研究』好文出版、第 2 号 2001 年は、古典・現代の中国語における Ngram の応用を専門に論じている。

参考ウェブサイト

- 社団法人著作権情報センターのウェブサイト、(http://www.cric.or.jp/) には著作権関連のコンテンツが多い。とくに世界各国の著作権法を紹介している部分は興味深い。

注

[1] 中華人民共和国著作権法は 1990 年に発布され、2001 年に改訂された。1990 年では著作者の権利は 5 項目であったが、改訂後は 17 項目に増えた。そのなかで「情報ネットワーク送信権」が設定されたのは興味深い。(国家版権局 http://www.ncac.gov.cn)

[2] 2005 年 5 月 30 日に「互联网著作权行政保护办法」が公布された。その第五条は「著作者がインターネットにおける伝播の内容についてその著作権を侵犯していることに気づき、インターネット情報サービス提供者に通知を送付したなら、インターネット情報サービス提供者はただちに関係する内容を削除する措置をとり、著作者の通知を 6 ヶ月間保存せね

ばならない。」というもので、著作者による著作権侵犯の訴えができるようになった。サービス提供業者は「削除を実施せずに、社会公共の利益を害した場合」（第十一条）にのみ処罰の対象となる。（国家版権局、信息産業部 2005/4/30）。

[3] 創作共用 http://www.creativecommons.cn/ を参照。

[4] http://sourceforge.jp/projects/morogram/からダウンロードできる。ただし、msvcr71.dllが同じフォルダーに必要である。

[5] http://mtgroup.ict.ac.cn/~zhp/ICTCLAS.htm を参照。

[6] ICTCLASS の readme ファイルの情報によった。

[7] 以下で公開している。
http://taweb.aichi-u.ac.jp/saitom/chuugokugo/hanyu/grep_on_web.html

[8] 日本語の作品では、青空文庫が提供している夏目漱石の『こころ』を全編貼り付けてみて機能が実行できた。