中国語のコーパスの作成

1R−1

田中康仁
兵庫大学
E−mail:yasuhito@humans−kc. hyogo−dai. ac. jp

〔1〕はじめに

我々にとって中国は隣国であり古くから文化交流のある地である。中国との交流をはかるには言葉の問題がある。そこで、ここでは中国語のコーパスを作ることを考える。

〔2〕何を研究材料とするか

中国語のコーパスとして多くの研究で「人民日報」が取りあげられる。しかし、ここでは中国共産党の雑誌である「紅旗求是」を材料とした。これを対象とすると内容が政治的であると嫌がる人もいるが、次に述べる方法を採用して処理するとそのようなことはなくなってしまう。

〔3〕コーパスの処理手順

次のような処理手順でコーパスを作成する。

(1) 紅旗CD−ROMから文の抽出
↓
(2) 文字数別に文を集めファイルを作成
↓
(3) 同一文の抽出、統合化と頻度付
↓
(4) 先頭文字順に分類
↓
(5) エディターにより文のみの抽出編集、検査
↓
(6) 編集済ファイルの作成、各種統計の作成
↓
(7) 各種応用分野の研究

1つの文の形式は次のようにする。

×××, ×××, ××, ××××××××
番号　文字数, 頻度,　中国語（中文）

文末の判定は次の記号をもちいた。

. ！ ？

である。

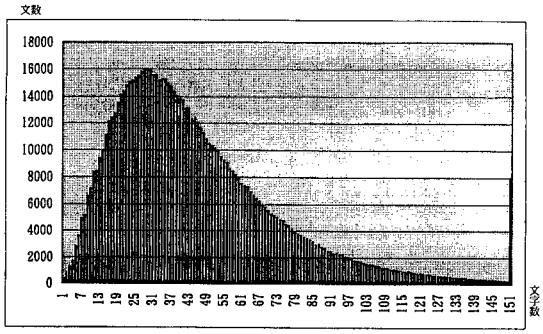
How do we make Chinese Corpus?

Yasuhito Tanaka
Hyogo University

〔4〕コーパスの性質

紅旗は中国共産党の理論誌であるため文章が固い文が多いし、長文が多いことが特徴としてあげられる。

文字別統計をグラフにすると次のようになる。



紅　旗

美しいグラフになる。29文字あたりが文数のピークであり、長い文では150文字以上の文もある。

抽出できた文数は約80万文である。そのうちデータとしてよいものだけを選び約70万文を整理した。

さて、この中で同一文がどの程度発生しているか調べてみた。次の表を参照されたい。

文字数1の文で説明すると次のようになる。

文の種類 …… 94文
重複文の種類 …… 40文
延べ文数 …… 402文
重複文の延数 …… 348文
重複文の種類／文の種類(%)　42.55
重複文の延数／延べ文数(%)　86.57
重複文の延数／重複文の種類　8.70

重複文は頻繁に使用される文であることがわかる。

紅旗データ

文字数	文の種類	重複文の種類		延べ文数	重複文の延数		
s	a	b	(b/a)×100	c	d	(d/c)×100	d/b
1	94	40	42.55	402	348	86.57	8.70
2	328	76	23.17	817	565	69.16	7.43
3	615	131	21.30	1290	806	62.48	6.15
4	1135	139	12.25	1669	673	40.32	4.84
5	2262	285	12.60	2843	866	30.46	3.04
6	3029	248	8.19	3604	823	22.84	3.32
7	4234	313	7.39	4932	1011	20.50	3.23
8	4685	250	5.34	5154	719	13.95	2.88
9	5917	295	4.99	6566	944	14.38	3.20
10	6597	265	4.02	7154	822	11.49	3.10
11	7868	255	3.24	8370	757	9.04	2.97
12	7900	237	3.00	8444	781	9.25	3.30
計	44664	2534	5.67	51245	9115	17.79	3.60

1文字から約10～12文字のデータには重複して使用される文が多いことがわかった。この性質は機械翻訳システムの例文を集めることにおいても重要なものである。

この紅旗のデータ量について

1958年～1995年までの38年間の電子総合訂本が紅旗出版社から出されている。

このCD－ROMは520MBの容量である(但し写真等を含めて)。

〔5〕中国語コーパスの利用方法

次のような6つのことがあげられる

(1) 機械翻訳用のテストデータとして使う。

中国語・英語または中国語・日本語等の機械翻訳テストデータとして有用である。

(2) 中国語の形態素解析、統語解析用データとなる。

(3) 中国語の読上げデータの材料となる。

(4) 中国語教育の材料として有用である。

特に短い文、1～12文字程度の文の中にはよく使われるものが多数みうけられるので有用である。

例えば

謝謝你。←→ あなたに感謝します。

您好 ←→ こんにちは

不客気。←→ どういたしまして

(5) 文型パターンの抽出

文型パターンの抽出用データとしても有用である。

例えば

我姓 田 中。←→ 私は田 中です。

我姓〔A〕。←→ 私は〔A〕です。

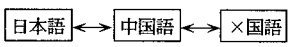
(6) 中国語文の形態素解析、統語解析

これらの研究はかなり進んでいる。形態素解析では北京大学の俞士汶教授と中国富士通研究センターが協同で人民日報一年分の形態素解析を行っている。その精度は機械処理と人間による修正で99.9%の精度を得るものが作られている。

(7) その他

〔6〕多言語のコーパスの作成について

筆者は英語が世界の各国語との媒介言語としての役割が大きいと考えたこともあるが、中国語を話す人々の数は多いし、中国料理店が各都市にあるように、中国人は世界の都市で活躍している。また、中国語は漢字を使用し、意味を表わす文字を使用している。これは日本語も同じである。(但し、中国語の漢字の意味と日本語の漢字の意味には異なっているものも多い。)



一つの可能性として検討すべきである。

〔7〕おわりに

我々日本人は、中国語のコーパスなどを作成することなどは不可能なことだと思っていたが、紅旗のCD－ROMを入手することにより中国語の初歩的分析の段階を行うことができ、約70万文程度の例文を集められたことは大変よろこばしいことである。さらに研究に役立てたいと考えている。

〔8〕データの入手先

(1) 中国語のCD－ROM、書籍

東京神田の内山書店は、中国語の書籍を中心に取り扱っている。CD－ROMも取り扱っている。

㈱内山書店 情報サービス部

〒101-0051 東京都千代田区神田神保町1－15

TEL 03－3294－0671

FAX 03－3294－0417

E-mail:XLN04226@nifty.ne.jp

中国で買う値段よりはるかに高いが、日本国内で入手できる点が便利である。

研究用として使用する場合は問題ないが何かの応用としてアプリケーション・プログラムに組込むにあたっては著作権の問題を解決しておかなければならない。

(2) 中国語の新聞

人民日報は次のホーム・ページで見ることができる。

http://www.people.com.cn

http://www.peopledaily.com.cn

中国の次の住所が人民日報の本社である。

中国北京市金台西路2号

100733 (Postalcode)

Tel: (8610) 65003109

人民日報社

(3) 日本で得られる中国語の新聞

中文導報

東京都品川区五反田7－13－6

SDI 五反田ビル5F 〒141－0031

Tel 03－5434－3177, 03－5434－3186

Fax 03－5434－3055

E－mail:chubun@gol.com

http://www.chubun.com/

週刊新聞である。

(4) 一般用語辞書

日中、中日の用語辞書は小学館、岩波書店から出版されている。

(5) 専門用語辞書

日中英の専門用語辞書も各社から出版され始めている。

- 朝倉書店：物理、生物・生化学、エレクトロニクス、電気、機械、土木、医学
- 工業調査会：英日漢工業技術大辞典

その他の出版社からも色々出されている。もはや、中国語処理も大変な時代ではなくなった。

〔9〕参考文献

(1) 内田慶市・野原康宏 マックで中国語
ひつじ書房 1996.7

(2) 俞士汶 現代漢語語法信息詞典詳解
清華大学出版社 1992.4

(3) 田中康仁 中華民国第二屆計算語言学研討会参加報告と研究所訪問報告(台湾の言語活動の報告)
情報処理学会自然言語処理75－13
情報処理学会1990.1

(4) 田中康仁・北 研二 中国の自然言語処理について
情報処理学会自然言語処理124－5
情報処理学会 1998.3

(5) Yasuhito Tanaka Kenji Kita
JCKE Multilingual Corpus Major Asian Languages
5th International Congress on Terminology and Knowledge Engineering (TKE' 99) 1999. Aug.
Infoterm