日英・パラレル・コーパスの作成

田 中 康 仁
兵 庫 大 学
E－mail:yasuhito@humans－kc. hyogo－dai. ac. jp

〔１〕はじめに

自然言語の研究開発を進めるにあたって重要なことは、対象となる言語データをいかに集め、そろえるかにかかっている。しかし、言語コーパスを作成することは大変な努力と時間がかかるものである。だが、その割に評価されない面がある。

ここでは日英・パラレル・コーパスを作成してみた。その方法、コーパスの性質について述べる。

〔２〕日英・パラレル・コーパスの作成について

大学では多くの学生に情報処理教育の講義と演習を行っている。

タッチタイプによる文字入力もその一つのテーマである。この作業の確認として日本語、英語の入力を行っている。この成果を集めることにより日英・パラレル・コーパスを作成することを考えた。

次に学生への説明資料の一部を示す。

学生への課題

課題提示「用例文ファイル作成」

［課題］用例文ファイル作成

1 番号　（半角文字入力）［改行］
2 日本語文（全角文字入力）［改行］
3 英語文　（半角文字入力）［改行］

を一つのデータ単位とし、用例を調べ、データ件数300件をWordまたはメモ帳などで入力して下さい。ただし、各項目は¥（半角文字）で区切って下さい。（¥の前後も半角スペースを入力。）最終的に指定したファイル名でテキスト形式の保存を行い、2HD（1.44MB）のフロッピーディスクを提出して下さい。

［形式例］

¥1¥ 0001
¥2¥ 会員たちは不思議な経験について次々に語った。
¥3¥ The menbers told us about the strange experience by turns
¥1¥ 0002
¥2¥ 小動物が四方八方に走り去るのを見た。
¥3¥ I saw some small animals running away in all directions..

学生に説明しても統一した形式で入力するものではなく、様々な形式で入力する。これについては集まったデータを整理するなかで統一することを考えた。また、入力データの中には誤字や誤ったスペリングが見受けられるが、これはWordのスペルチェック機能を使うことを指示している。しかし、それでも発生する誤りデータはデータ整理の中で修正した。

このデータ入力形式は非常に単純でわかりやすい形式である。特別なソフトウェアに依存するものではない。

(1) 同一データの発生

学生達とこのような方法で文を集めたが、実際には多くの重複が発生した。この原因としては次のようなことが考えられた。

ｉ）学生が入力作業を怠り、他人のデータを流用する。

ii）同一文、同一文章が使われている。

- good morning　おはようございます。
- What time　is it now?　今何時ですか。

基礎的な慣用表現は多くのテキストの中で使われている。

iii）同一出版社の本の中には出版社の持っている例文を用いるため、同一文がしばしば現れる。同一著者の出版物にもこのようなことがある。

それ故、ただ単純に文を集めればコーパスができるというものではない。

(2) 形式の不統一と誤りについて

多くの学生の入力形式をみると、空白の使用、大文字、小文字の使い方、改行の入れ方、タブ記号の使用等色々な問題がある。これらは少しずつエディターを使って統一するより方法がない。

誤り文字等もWordの機能と１件、１件を確認し、修正せざるを得ない。

コンピュータにより発見できる誤りについては十分調べ、修正した。

(3) データについて

学生達が大学入試で用いた参考書、教材等の中から何でもよいとした。著作権の問題も考えなければならないが、単文ごとに著作権が発生するものではないと考えた。

このため同一の参考書を使用することによる重複もある。

(4) コーパスの収集量

１回に学生が入力するデータは300文であり、１年間で約200人～250人程度の学生を教えて、データを収集した。

1年目	5万文	1997年（準備の年）
2年目	合計12万文	1998年
3年目	合計16万文	1999年
4年目	合計21万文	2000年

４年間で21万文　日本語・英語のパラレル・コーパスを作成することができた。

学生達とこのように文を集めたが、実際にはこの倍程度の文を入力した。文の重複（訳文も含めて）が多いことが分かった。

Japanese － English Parallel Corpus
Yasuhito Tanaka
Hyogo University

〔3〕データ収集の考え方

学生を動員してデータを集めることについて次のような方針を考えた。データをただ単純に集めるということから発展して、ただ単純に事例を集めておもしろいという段階から、さらに研究を進めてみたい。

1) 少量ずつ集めても積もり積もれば山となる。量的変化は質的変化をもたらす。
2) 言語学習、文化、歴史の相異の研究に発展する。
3) 現代文の入力を基本とした。我々の日常生活で使われている文を対象とした。
4) 文単位の訳はすでにできた訳文を利用し、入力した。
5) 一つの例文に対しての多くの訳文を豊富に集めた。これにより色々な例外事象の発見につなげたかった。
6) 数量的分析
   数量的分析のためのコーパスとしてはどのようなものか研究が必要であるが、一つのコーパス事例 として考えたい。このコーパスが数量的分析に十分であるとは考えていない。
7) 異分野の研究者の交流
   異分野の研究者の交流の場としては共通に利用できるものが必要である。

このような考えに基づいてデータを集めた。

〔4〕日英コーパスの量と内容について

集まったデータの形式と誤りの修正、重複の削除を行い、2001年3月末現在約21万2千文の日英対訳データができあがった。

次にこの内容を示す。

1) Data総数	211,997文
2) 英単語の総数	1,636,461単語
3) 1文中平均単語数	7.72単語
4) 中央値（英単語数）	7単語
5) 1文中最小英単語数	1単語
6) 1文中最多英単語数	45単語
7) コーパスの総バイト数	約19MB

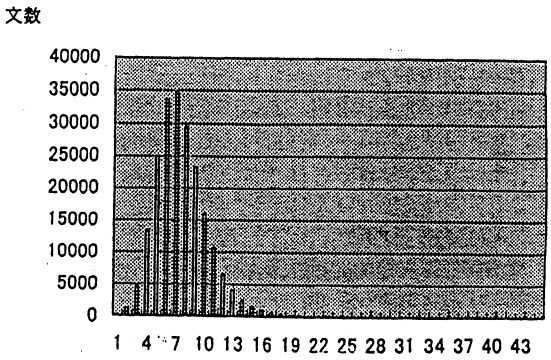
文の長さの分布を調べるために英単語による文数の頻度を調べると次のようになる。

1	2	3	4	5	6	7	8	9	10
178	1,314	4,745	13,451	25,099	33,726	34,907	29,551	23,369	16,092

11	12	13	14	15	16	17	18	19	20
10,719	6,611	4,175	2,528	1,693	1,102	715	506	314	278

21	22	23	24	25	26	27	28	29	30
199	150	116	99	71	55	42	33	38	18

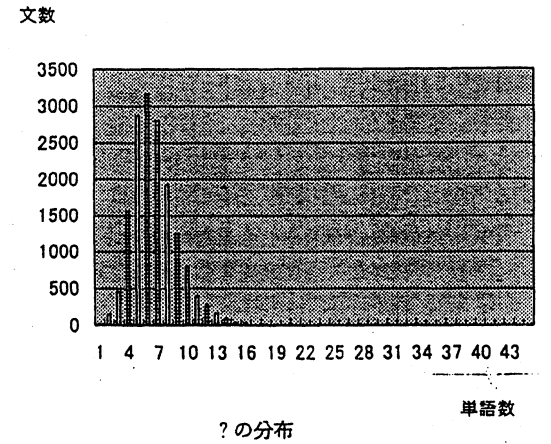
31以上	合 計
103	211,997



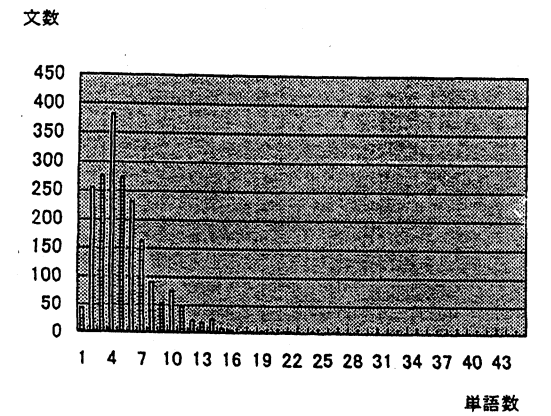
文の分布

次に？と！記号の分布を調べると次のようになる。

1) ?記号の数	16,154記号
2) ?の中央値	6英単語文
3) ?の使用文数の割合	7.64%
4) !記号の数	2,011記号
5) !の中央値	4単語文
6) !の使用文数の割合	0.95%



?の分布



!の分布

このコーパスには次のような特徴を持つことが分かった。

1）短い文が多い（平均7.72英単語）

収集量が増大するにつれ英単語数の平均が長くなった。これは日英ともに同一文を削除したことによるものである。

2）人称代名詞（I, you, He, She, They, We…）等が主語となる文が多い。

I	38,735文	
You	6,376文	
He	24,111文	
She	10,453文	
They	3,908文	
We	7,250文	
計	69,115文	→ 32.6%
That	1,286文	
This	4,906文	
It	6,990文	
計	13,182文	→ 6.22%

これらは文字列による数値なので厳密な数ではない。He とShe を比べるとHe の方が2倍強ほど多い。これは男性社会の一面がわかる。

3）Please を使用した文も多い。27,346文もある。約12.90% であった。依頼文が多いことも分かる。

4）疑問文の例文を調べてみると次のようになる。

（i）	What	3,392
（ii）	Where	600
（iii）	When	1,280
（iv）	Who	636
（v）	Which	279
（vi）	How	2,395
（vii）	Do you	1,533
	Does	219
		10,334

5）新聞や専門分野の文とは異なっている。

日本語の表記は一つであるが英語の表現では幾つかあるもの、英語の表現では一つであるが日本語の表記のゆれ、丁寧表現等で数種類の表現があるものがある。これらの中には、漢字の表記が仮名書きになった程度のものもある。しかし、ここではそれらについては削除していない。

例1　こんにちは。　1）Bonjour!
2）Hello, Nice to see you:
3）Hi.

Thank you for your kindness.
1）いろいろご親切にありがとう。
2）ご親切ありがとうございます。
3）ご親切にありがとう。
4）ご親切誠にありがとうございます。
5）親切にしてくれてありがとう。
6）大変お世話になります。

等6つの日本語の表現があった。

このようなことから、日本語だけ、英語だけにすると文の数はもっと減少する。

〔5〕パラレル・コーパスの応用分野

パラレル・コーパスの利用分野としては次のようなものを考えている。

(1) 機械翻訳システムのテストデータとして使用する。これは非常に有効であり、実際に実施して効果をあげている。

(2) 機械翻訳システム用の定形文の抽出

例えば

Good morning ←→ おはようございます。

このように我々が常に使う短い文で、一定のきまりきった文は色々な解析をするのではなく、そのまま対応した訳文を出力する。このような文はパラレル・コーパスを英単語数別に分類することにより、簡単に集めることができる。1英単語から10英単語までの文の中に多く含まれている。

(3) 機械翻訳システムの文型パターン抽出用文例として使用する。

機械翻訳システムではこなれた翻訳結果を出すために多くの文型パターンを使用している。このために使用する。

(4) 機械翻訳システムの訳語選択のセンスワードの選択や専門用語抽出のための文例として使用する。

(5) 外国語教育のための教材作成用文例

外国語教育においては、多くの学生に少しずつ変化に富んだ内容の文例を与えなければならない。教師がこれら全てを考え出すことは大変である。このための文例データ・ベースが必要である。学生毎に学習度合に合った教材が重要である。

(6) マルチリンガル・パラレルコーパスの作成

日・英パラレルコーパスを発展させて、中国語、韓国語、仏語、独語、ロシア語、スペイン語等を付け加えてゆき、各国語との対応を作る。

(7) 日・英パラレルコーパスを基にして日本語や英語の語の意味の研究や概念の研究に役立てる。

語用論の研究資料とする。

(8) 形態素解析、構文解析、意味解析

形態素解析、構文解析、意味解析用のテストデータとして用いる。多量のデータによる解析に耐えうるものにしなければならない。

(9) 自然言語解析のデータとして

自然言語解析の一つの道具として使用することができる。

自然言語研究によって各種規則が出来てくる。この規則を検証するための道具の一つとして使える。例外の発見にも役立つ。

〔6〕今後の計画

今後この日英コーパスをどのように発展させるかについて述べる。

1) 量について

次のような計画でデータを増加させる。

2001年	26万文
2002年	31万文
2003年	35万文
2004年	40万文
2005年	45万文

重複データが増えるので、これを削除しながら増加させてゆく計画である。文章も少し長いものが増え、平均英単語は8～9単語になるであろう。

2) 質について

日英対訳コーパスの中には誤りや. 、。の入っていないものが見つかる。これを修正し、品質の高いものに修正してゆきたい。コンピュータ処理によって見つかる誤りは極力見つけ、修正する。改訂版を出してゆく計画である。

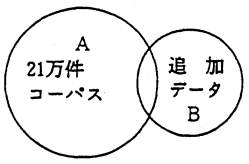
3) 今後の研究として

① 文の変換

このコーパスに含まれる文で、例えばHeのものをSheにするとか、その逆を行ってみるとか、Iの文をWeにしてみることなどを考える。全ての文が自動的に変換できるものではないが、何かおもしろい結果も出るであろう。

② 文の追加と文の長さの変化

今後もコーパスを増加させたい考えている。しかし、どのような長さのものが増加するのか興味あるところである。



AUBの作成
A∩Bの作成
B－Aの作成

AUB、A∩B、B－Aの各データの分析を今後の追加データについて検討する。

4) 応用分野

新しい応用分野を考え出し、使用する。

① 多言語化

英語を軸に多言語のコーパスに発展させたいと考えている。中国語、韓国語、ロシア語、フランス語、ドイツ語、スペイン語……等世界には主要な言語がある。これらに拡大してゆきたい。

② 自然言語処理への新しい応用

言語の意味処理を考えるにあたって、多くの言語に日本語を照らし合わせる中で、日本語の意味、概念が明確になるのではないかと考える。

③ 機械翻訳システムへの新しい応用

機械翻訳システムも日英、英日システムから多言語システムへ発展しようとしている。このための基礎的なコーパスとして使えるものする。

5) 今後このデータの公開について

このデータを公開すべきか否かについて各方面の方々と相談し決定したいと考えている。

6) 各文に特徴のマークを付けて文の分類を行う。

例えば：買物の文

：あいさつ文

⋮

目的別に分類し、利用方法を考える。

〔7〕おわりに

日英パラレルコーパスの作成、性格について述べた。これは一つの試作にすぎないが、今後このようなパラレルコーパスが多く作られることを望む。又、ここではパラレルコーパスの応用分野についても検討してみた。これら分野も拡大し、ただ単純に自然言語処理ばかりでなく、マルチメディア、インターネット、外国語教育等人と人のインターフェイスの分野で研究、応用されることを望む。

〔8〕参考文献

1) Yasuhito Tanaka, Kenji Kita
Multilingual Parallel Corpus of Major Asia Language
TKE'99 (Terminology and Knowlegde Engineering)
Insbruck (Austria) 23－27 August1999 Infoterm
2) 斉藤俊雄、中村純作、赤野一郎
英語コーパス言語学　研究社
3) 鷹家秀史、須賀　廣、
実践コーパス言語学　桐原ユニ
4) 田中康仁　機械翻訳用のテストデータ
情報処理学会　第61回全国大会　2T－1
PP2－127～128　2000年9月