マルチリンガル・コーパスの作成

田中康仁

兵庫大学

E-mail:yasuhito@humans－kc.hyogo－dai.ac.jp

〔0〕はじめに

自然言語の研究分野を拡大して、一つの言語から複数の言語に移そうとするとその基礎となるデータが必要である。それらは辞書であったり、各種のツールであったり、コーパスである。

ここではマルチリンガル・コーパスの作成について考えてみる。多くの言語の理論家は数十の文でことたりるかもしれないが、実際の応用物を自然言語処理において、特に多言語を基礎として作成しようとする際には、マルチリンガル・コーパスはなくてはならない。ここで述べるマルチリンガル•コーパスは単に複数の言語の同一内容のデータがあるというだけではなく、文単位に異なった言語で表現され、それが簡単にコンピュータで処理できるものである。マルチリンガル•コーパスであり、パラレル•コーパスである。

〔1〕マルチリンガル・コーパスの開発の意義

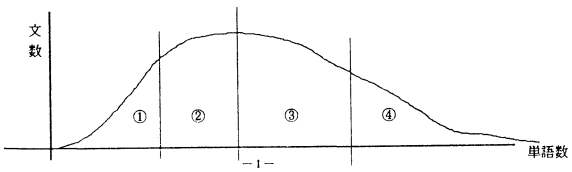
言語の研究では、学者の頭脳の中に蓄えられた多くの例文や、反例などによって研究を進められてきた。そのため、学者の頭脳の中に多くのものが蓄えられるまで良い論文が書けないことになっていた。しかし、現代では計算機の進歩により、複数の言語を取り扱うことができるフォントや、コードが統一化されてきた。また入力も複数言語で同時に可能になってきたし、入力されたデータを画面上で自由に取り扱うこともことも可能になってきた。このような計算機環境の整備により、マルチリンガル・データの入力、蓄積が可能になってきた。言語学の新しい研究もマルチリンガル・データを基礎とした研究が主流になり始めている。工学者と言語学者の提携で新しい計算言語学が興隆してきた。

若い言語学者でもマルチリンガル・コーパスがあれば充分研究が出来る状況になっている。今後はこのような状況をもっと発展させるため、「紙に書かれた知識の電子化」を進めていかなければならない。

ルネサンス文化はギリシャやローマで発展した文化（ラテン語文化）の翻訳から始まった。

我々は今までの紙に書かれた膨大な知識を電子化し、再構築し、利用できるものにしてきた。マルチリンガル・コーパスはまさにその第一歩と考えてよいものである。

目前の研究目標は機械翻訳の精度向上である。機械翻訳は二つの技術が融合しようとしている。一つは例文を基本とした機械翻訳であり、もう一つは統語解析による機械翻訳である。特に統語解析の機械翻訳では意味解析の文型パターンの抽出と拡大が重要である。しかし、これらについての具体的研究方法論が確立していない。我々は英文を単語数別に分析すると一般的に次のような分布図になる。



－1－

考え方だけがあってもそれを支えるデータがなければその技術は役に立たない。

この①の部分の文は例文を基本とした機械翻訳の例文として使える。

例えば　Good morning.　のようなものである。

②の部分はほぼ単文である。これは例文翻訳や一部可変な文として利用できるし、単文から意味解析の文型パターンの抽出ができる。

③の部分は複文、重文などの少し複雑な文の構造を研究するために用いる。

④の部分は長文の分割や文を部分構造として分解して利用するための材料として研究できる。

このような研究と同時に、複数の言語との対照研究も進めることができる。これにより一つの言語では見えてこなかった現象がはっきりとしてくる。

さらに、機械翻訳のテスト文としても利用できる。機械翻訳では正しい訳が無いと正しいか否かの判定が専門家に頼らざるをえない。しかし、マルチリンガルのパラレル•コーパスがあれば、誰でも簡単に判断が付くようになる。

機械翻訳システムも一つの工業製品である。この工業製品にも製品としてのテスト方法を確立しなければならない。マルチリンガル•コーパスはそれに一石を投じるものである。

〔2〕マルチリンガル・コーパスの作成について

ここでは一般的にコーパスを作成するにあたって要求される一般的条件と問題点について考えてみる。

コーパス作成に際して要求される要件

1）安い、すなわち入力作業やデータの校正作業に手間がかからない。
2）速く。
3）品質が良い。
4）著作権の問題がない。
5）色々な分野から選ばれた文であり、現在使われている文を考える。

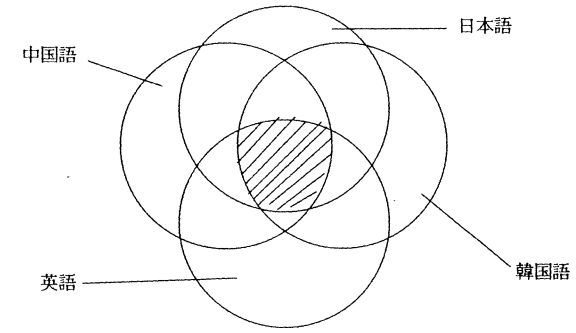
次に考えなければならない点は量の問題である。コーパスとして使用できるものにするためには最低でも数万文以上必要であり、数十万文程度のものへ拡大しなければコーパスを作成しても応用範囲が限られてしまう。

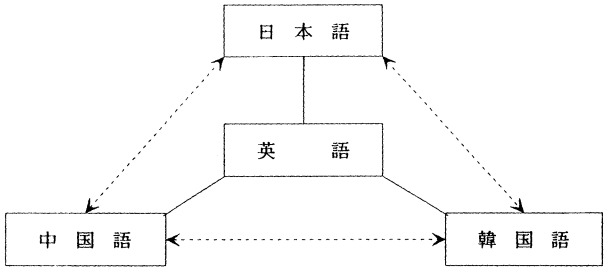
さらに考えなければならない点は、どのような言語のマルチリンガル・コーパスを作成するかということである。

ここでは日本語、英語、中国語、韓国語を最初の目標とし、順次拡大できる方法を考える。

各言語を考えるにあたって、言語はその言語を使う人々、民族の歴史や、文化を背景に持っている。そのため正確に相手の言語に翻訳できないという側面がある。しかし、ここではそのような面も考えながら試行してみることにする。

英語は世界の共通語的な言語でもあるので、日本語⇔英語、中国語⇔英語、韓国語⇔英語の組を考え、それらの組を相互に結合する中で翻訳を考えてゆくことにした。





もちろん、英語以外の言語と相互に翻訳可能な人にも充分データのチェックを行ってもらうことにした。

〔３〕マルチリンガル・コーパスを作成する方法

マルチリンガル・コーパス作成方法として、次の五つの方法を検討する。

(1) 機械翻訳を応用する。
(2) 多くの人達による文の選定と入力を行う。
(3) 専門家に依頼して作成する。
(4) ｗｗｗ上のデータを利用する。
(5) 各種辞書（ＣＤ－ＲＯＭ）からの例文抽出

〔３〕－１　機械翻訳を応用する方法

単一の言語のコーパスとか、ある特定分野の言語データをＣＤ－ＲＯＭにまとめたものがある。これから文を抽出し、整理する。その後機械翻訳システムにかけて翻訳してみる。

英語→日本語の機械翻訳システムでは約７０％の正解が得られるので、この方法は有力な方法である。

次に具体的に実験した結果について述べる。

日本電子化辞書㈱の作成した英文コーパスは約12万文ある。これら各文を単語数別に分類すると次のようになる。

単語数	1	2	3	4	5	6	7	8	9	10	11
文の数	0	19	460	1,889	5,030	6,798	7,791	8,416	8,698	9,114	9,018

12	13	14	15	16	17	18	19	20	21	22	23
9,176	9,050	8,815	8,446	8,245	7,466	6,559	5,283	3,645	562	464	321

24	25	26	27	28	29	30	31	32	33	合　計
235	142	77	46	23	12	2	0	0	1	125,803

この中で２～７単語で構成されている21,854文を対象とした。Ａ社の機械翻訳システムにかけ、その結果を人手によって分析すると次のようになった。

英日翻訳システム評価結果

評点	5	4	3	2	1	合計	平均値
2　単　語　文	15	1	1	1	1	19	4.47
3　単　語　文	336	97	16	11	0	460	4.65
4　単　語　文	1,371	355	156	25	2	1,889	4.61
5　単　語　文	3,655	808	510	53	4	5,030	4.60
6　単　語　文	4,742	1,379	595	81	1	6,798	4.58
7　単　語　文	4,471	2,331	870	118	1	7,791	4.43
合　計	14,588	4,952	2,149	289	9	21,987	4.53

このうち正しく翻訳できたものをコーパスとして採用する。また一部修正を加えればよいものは、順次修正を加えながら翻訳し、追加する。

この方法は入力文のうち７０％しか正解が得られないが、日本語の入力を行わなくて済むので、大変良い方法である。しかし、正しい文か否かの校正作業に時間がかかるという問題点がある。

原文の作成者に著作権があるので、この点については十分注意を払わなければならない。

〔３〕－２　多くの人々による文の選定と入力

大学では情報処理についての基礎的講座がある。そこで学生達はコンピュータの基礎について学ぶと同時に、コンピュータの使用方法も習得する。タッチタイプの方法も習得する。その成果としてデータの入力を行わせている。付録の作業指導書はその例である。このようにして日本語と英語の対を入力させた。学生達のタイピングの練習成果物の再利用としてマルチリンガル・コーパス（データ・ベース）を作成することを考える。

具体的実験結果を次に示す。

入力に参加した学生	182名	
学生の入力データ総文数	54,606文	100.0%
英語と日本語が重複していない文	42,889文	78.5%
英語の重複していない文	37,547文	68.7%

英語と日本語が両方とも完全に一致したものが11,717文あった。これは学生同志のコピーによるものか、たまたま同一のテキストを使用したために起こったものかである。５万４千文強の学生の入力データの内１万２千件の重複が見つかった。これは同一の教材を使ったことも考えられるし、慣用表現という特定表現の中では同一の表現を幾つかの本、教材の中で使用しているという点がある。このため同一データの作成がなされたものである。他人が入力したものをそのまま使用した学生は２人（１組）だけであった。このことからも同一表現は５万４千文強の内２割程度はあることがわかった。１０万文～１５万文程度にすると３～４割程度の慣用表現の入力は既に入力したものと一致するものと考えられる。さらに慣用表現の文を調べてみると同一点が多くの資料にあらわれる。

例えば

(1) How do you do?
(2) What time is it now?

これらはある資料から孫引きされたものではないかとも考えられる。別の考えとして英語文化圏の特定表現形式であり、文化であって孫引きではないという意見がある。筆者は後者の考えである。著作権の問題は柔軟に考えてゆきたい。

この約4万3千文のデータを単語数で分析すると次のようになる。

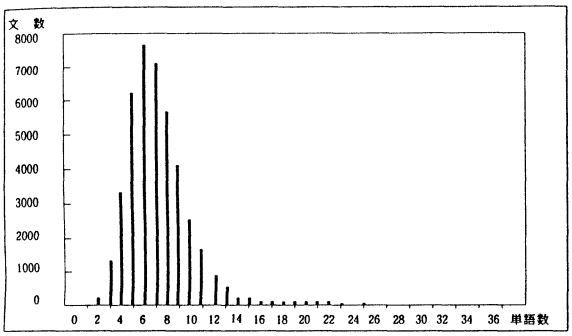
単語数	1	2	3	4	5	6	7	8	9	10	11
文　数	7	254	1,216	3,342	6,291	7,800	7,266	5,842	4,093	2,754	1,631

12	13	14	15	16	17	18	19	20	21	22	23
996	530	276	213	115	81	43	40	25	18	20	10

24	25	26	27	28	29	30	31	32	33	34	35
4	9	1	5	1	1	0	0	0	0	0	0

36	37	合　計
0	1	42,885

グラフにすると次のようになる。



これは1年間の学生達の入力成果であるが、これを数年間継続すれば十数万件のコーパスになる。

学生達にはデータの校正は充分行うように注意しているが、この点については今後の問題点である。著作権についても多少問題がある。

学生達の入力環境には次のような問題がある。

学生達は入学時に購入したノート型パソコンと大学のパーソナル・コンピュータを主に利用している。しかし、その他の機器としては独自に持っているワープロ等もある。学年毎に使用しているＯＳがWindows３．１，Windows９５，大学の機器はWindows ＮＴとさまざまである。入力用のソフトウェアもWindowsのワード９５，ワード９７，ワード９８，メモ帳、一太郎等さまざまである。これらを一つにまとめ統一的なデータにまとめあげることも一苦労であった。だが、この程度の多様さは認めなければならない。

〔３〕－３　専門家に依頼して作成する。

英語の教師に日本語と英語の対になっているデータの作成を依頼した。

約１，２００文の日本語と英語の対になった文を作るのに約６ケ月かかった。また費用は一対の文当たり２００円程度かかった。費用がかかり文数は少なかった。具体的には英語検定文の時制や人称代名詞をおきかえて作成してもらった。著作権は問題ないが、費用がかかるわりに量が少なかった。

〔３〕－４　WWW上のデータを利用する。

WWWを作っている多くの日本の企業や団体の中には画面の一部に日本語、英語の選択ボタンがあり、そのいづれかを選択すると日本語の説明文、英語の説明文が現れるようになっている。これを利用し、英語、日本語の対になった画面を選択し、その中から対応する文を自動的に抽出する方法が考えられる。これは機械的に多くの部分が処理できるので良い方法である。しかし、最終的には人手による検査を行わなければならない。著作権についても少し検討しなければならない。この方法については、具体的に検討中であり、実験中である。WWWから日本語と英語の対応文を抽出する方法は大変興味あるものであるが、WWWは宣伝文が多いので少し普通の文ではない面がある。また文の区切りが明確でない点や、画面上での表現に重点がおかれている側面が強くあらわれている点に注意しなければならない。

〔３〕－５　各種辞書（ＣＤ－ＲＯＭ）からの例文抽出

この方法はＣＤ－ＲＯＭを読み、日本語と英語の例文が抽出できれば可能であるが、辞書作成会社にマルチリンガル・コーパス作成の許諾を得なければならないし、中国語や韓国語を附加した場合の著作権についても、契約を取り交わしておかなければならない。これは大変重要なことである。このような事を行わずに作業をしてはならない。このためこの方法については、具体的に実施していない。

〔４〕具体的方法の評価

〔２〕で述べた幾つかの方法を検討した結果

(1) 機械翻訳を応用する。

(2) 多くの人達による文の選定と入力を行う。

(3) WWWのデータを利用する。

この三つの方法が実用的な方法ではないかと考えている。機械的処理を多くすることにより、品質の良いデータが得られる。

著作権の問題は常に考えなければならない。しかし、多くのデータを混ぜあわせ、量を非常に多くする。しかも、英語、中国語、韓国語を追加する。単語数別に分類してファイルを作成する。最初は主として自然言語の研究を目的として活用する。このようにすれば我々独自の著作権、編集権が生じると考えている。

〔５〕マルチリンガル・コーパスに向けて

我々は多くの国々と自由に交流を行うことができるし、インターネットで情報の交換も行えるようになってきた。また色々な国の留学生や長期滞在の外国人、国際結婚した人達、外国に長期滞在して帰国した人達がいる。このような人達と協力して翻訳を行えば、マルチリンガル・コーパスは作成できる。さらに中国語と韓国語については、次のような友人が協力してくれることになっている。

・中国語……北京大学　俞士汶　教授

・韓国語……韓国先端科学技術大学　崔　杞鮮教授

この二人の協力者は、中国語⇔英語、韓国語⇔英語のコーパスを作成している。

そこで我々は、それぞれコーパスを交換することにより、マル

チリンガル・コーパスを作成することができる。このようなことについては具体的な話合いが行われている。将来の発展に期待して頂きたい。

この研究組織は、日本語、中国語、韓国語、英語を取り扱うので、ＪＣＫＥプロジェクトと名前を付けた。

マルチリンガル・コーパスの例文を示す。

(1)
日本語：その工場は昨年数千台のトラクターを生産した。
英　語：The factory turned out thousands of tractors last year.
中国語：工厂去年生产了数千辆拖拉机。
韓国語　그 공장에서는 작년 트랙터를 수천대 생산했다.

(2)
日本語：あなたはそこにいる２人の男の人の名前を知っていますか。
英　語：Do you know the names of the two men over there?
中国語：你知道那边的两个男人的名字吗?
韓国語：당신은 저기에 있는 2명의 남자의 이름을 알고 있습니까?

〔６〕マルチリンガル・コーパスの応用

マルチリンガル・コーパスの応用としては、次のような分野が考えられる。

１）自然言語処理

特に機械翻訳の基礎処理として役に立つ。

２）辞書作成

機械可読辞書の見出し、語の分析抽出や例文抽出等に有効である。

as soon as、at onceというような例文を抽出しようとすると６６文、１２４文集まった。サンプルを付属資料として添付する。

３）言語教育、学習

日本語、中国語、韓国語の教育や、それら言語の学習等の教育素材として有効である。

４）音声研究

データにもっと附加情報を付け、音声研究の素材として利用する。

文章の読み上げや、音声認識等の言語情報となる。

５）その他

マルチリンガル・コーパスも対応する文だけの情報では応用文野が限定されるので、次のような処理が必要である。

(1) 分かち書き
(2) 形態素解析
(3) 統語解析（曖昧性の除去）

Tag付きコーパス、統語解析済みTreeバンクが必要である。

〔７〕収集されたコーパスについて

今年度収集されたデータは次のようなものである。

(1) 機械翻訳により収集したもの　　約１４，０００文
(2) 学生達の入力文（慣用表現文）　約４３，０００文
(3) 専門家により作成された文　　約　１，２００文
合計　約５８，０００文が集まった。

１９９９年、２０００年も同程度収集し、約１５万文程度にしたいと考えている。これらを中国、韓国の友人と交換し、翻訳してもらいマルチリンガル・コーパスとしたい。

これらは我々の研究用としても使うが広く公表して研究者には公表してゆきたい。一方、企業の人々には我々の作成費用の一部を負担してもらう費用で公表したいと考えている。２００１年の２１世紀の初めから公表を行いたいと計画している。

〔８〕今後の課題

今後のデータの収集分野としては

(1) 会話文

旅行、買物、病院、ホテル、恋人同志、商売、友人同志、学校での会話、教師と学生、親と子の会話も考えている。

(2) 手紙文

恋文、挨拶文、慶弔文、誕生日の文、商業文、ＦＡＸ文

(3) 標語

工場、病院、乗物、交通表示、等で多くみうけられるもの簡単に表現されたものがある。そして一定の意味を伝達するものがある。

これらを収集してゆきたい。

この研究の広がりとデータ・収集作業の拡大

幾つかの方法でマルチリンガル・コーパスの作成方法について考えてみた。学生達との共同作業を行っても、データの整理をうまく行えば良いマルチリンガル・コーパスが作成できる。同様の方法で全国の大学、研究所で行われることを期待する。その結果、特色のある内容のマルチリンガル・コーパスが作られる。

〔９〕おわりに

マルチリンガル・コーパスの作成、応用上の色々な事柄について述べた。

大量のマルチリンガル・コーパスを実現し、応用分野の拡大を計ってゆきたい。どのような分野の研究を行うにあたっても、基礎となるデータが充分にないと良い研究はできない。これらデータは広く研究者へ利用できるように考えてゆきたい。

〔10〕参考文献

(1) 田中康仁　対訳付慣用表現の収集について　情報処理学会自然言語処理研究会報告集121－14　１９９７年９月
(2) 斎藤俊雄、中村純作、赤野一郎編「英語コーパス言語学」研究出版　１９９８年３月
(3) Branimir Boguraev, James Pustejousky "Corpus Processing for Lexical Acquisition" A Bradford Book, The MIT Press Cambridge, Massachusetts London, England １９９６.
(4) Susan Armstrong "Using Large Corpora" A Bradford Book, The MIT Press Cambridge, Massachusetts London, England １９９４.
(5) Yorick A. Wilks, Brian M. Slator and Louise M. Guthrie "Electric Words" A Bradford Book, The MIT Press Cambridge, Massachusetts London, England１９９６.