『国際社会における日本語についての総合的研究』
第３班「日本語表記・音声に関する実験言語学的研究」（代表者：賀集寛）

表記・表現に関する計算機による実験的研究

国立国語研究所　　中野　洋

１．研究目的

本研究では、コンピュータによって生成される日本語および英語を自然な表記・表現とするための基礎的研究を行う。すなわち、日英翻訳において生ずる不自然な表記・表現を収集・分析し自然な表記・表現に変換する方法をさぐる。

具体的には、以下の３つの部分によって研究をすすめる。すなわち、その内容・目的に応じた日本語表記の変換実験、日英両語の翻訳分析にもとづいた計算機による修正実験、日英翻訳における訳語選択・省略照応研究による機械翻訳の高度化実験である。

これらは、すべて各種のデータを解析し、それを生成する規則を構築した上で、機械辞書を作成する。そして、これを用いたプログラムを作成し、計算機で実験しその妥当性を検討するという手順で行う。

２．研究組織と分担

・研究分担者　中野　洋（国語研）
　　自然な日本語表記・表現の生成
・国語学関係協力者
　　分類語彙表の増補・機械辞書検討評価
林大，石井久雄，石井正彦,大島資生，山崎誠（以上国語研），宮島達夫（大阪大），鶴岡昭夫（山口大），高木一彦（大東大）
・情報処理関係協力者　　機械翻訳の高度化
Juin'ichi Tsujii, Natsuko Holden（UMIST)
　　日英翻訳分析と計算機修正実験
John D. Phillips（山口大学）
　　機械翻訳のための訳語選択・省略照応研究
吉村賢二（福岡大)、平井誠（松下電器）
　　検討評価

３．本年度の研究成果

①表記変換のための機械字書とプログラムを作成する。来年度評価を行う。
②各種情報を付き機械辞書を作成した。これは来年度の準備である。
③分類語彙表を増補した。増補用の検討資料（略称 671 本　２分冊計 322 頁）を７冊印刷した。
④機械翻訳の高度化について研究計画を立案した。

４．研究内容

4.1 日本語の表記変換実験

4.1.1 表記変換の方法

4.1.1.1　日本語の表記

以下で本研究における表記変換の考え方を中野(1993)を引用しながら述べる。本研究ではそれを実現する。

日本語は４つの文字体系を持つ言語である。したがって、漢字だけでも平仮名だけでも片仮名だけでも、またローマ字だけでも書くことができる。もちろん、現在は、漢字仮名混じり文で書いている。これらの書き方はすべて正しい。しかし、それゆえ表記のゆれが生じる。

また、文字使用の約束事以外にそれぞれの文字に異なったイメージが付与される。すなわち、漢字は重要なもの、正式なもの。平仮名は、日本語的、やさしい感じ。片仮名は、外国的、強調。ローマ字は外国語をそれぞれ表す。それゆえ、それらのイメージを利用した新しい用字法も生まれる。たとえば、NHKは外国語だと思っている人は多い。しかし、それは日本語で「日本放送協会」の略語である。逆に、「倶楽部」は漢字で書かれているため日本語だと思われているが、それは英語である。

このように、同じ文章でも使う漢字の多少で受け取る感じが異なる。漢字がおおいほうがよりフォーマルで、書いてある内容も重要であると思われる。

しかし、漢字化にゆれがあれば書き手の教養を疑われる。また、漢字化はそれぞれの語によって異なる。その要因は、国が示した漢字化のめやす（常用漢字）という外的な要因、その文字の本質的な機能（漢字は表意文字であり、かなは表音

文字である）という言語内の要因、慣例にしたがうという社会的な要因、さらにはその文字が好きか嫌いかという個人的な要因などさまざまである。

本研究では、①このような日本語の表記がどのような要因によって変化し、また変わりえるものであるかを各種の資料によって分析する。次に、②日本語での表記をいろいろなパラメータによって変換するプログラムを作成する。

4.1.2 表記変換用字書・辞書

表記変換実験のために、常用漢字・教育漢字・学年配当情報を持った漢字字書（斎藤秀紀作成）を利用する。さらに、品詞・語種・文体情報をもった辞書（６万語収録）を作成した。

4.1.3 本年度の成果

文字単位の変換プログラムを作成し、実験を行う。一貫処理の仮名漢字変換プログラムに（中野 1991）これを組み込み、以下の変換を実現する。

漢字使用の文章を、指定にしたがって、常用漢字の範囲、教育漢字の範囲、４年生配当漢字までの範囲での表記などに変換する。

以下に、現時点での実験例を示す。

指定した漢字の字種を用いる。それ以外は音読みはカタカナで、訓読みはひらがなで表記する。プログラムはまだ完成していない。

［漢字字書例］　フォーマット：漢字, 文字種

化, 3,　加, 4,　我, 6,　会, 2,　海, 2,　開, 3,
外, 2,　害, 4,　拡, 6,　活, 3,　関, 4,　企, J,
危, 6,　議, 4,　協, 4,　強, 2,　況, J,　業, 3,
近, 2,　係, 3,　懸, J,　拳, X,　言, 3,　後, 2,
口, 1,　向, 3,　更, J,　国, 2,　今, 2,

［入力文］（時の動き 7.2.1 一部表記変更し引用）

我が国の拳銃対策-拳銃の摘発強化への取り組み-

拳銃の押収状況を見ると、その大半は密輸入されたものであり、国際化の進展に伴い、今後、更に海外からの流入の増加が危惧されます。また、ここ数年、暴力団以外の者からの押収が増加しており、拳銃の一般社会への拡散が進むことも懸念されます。

［処理結果１］　４年生配当漢字まで使用

わが国のケンジュウ対サク－ケンジュウのテキ発強化への取り組み－　ケンジュウのオウシュウジョウキョウをみると、その大半はミツユ入されたものであり、国サイ化の進テンニともない、今後、さらに海外からの流入のゾウ加がキグされます。また、ここ数年、ボウ力ダン以外の者からのオウシュウがゾウ加しており、ケンジュウの一パン社会へのカク散が進むこともケ念されます。

［処理結果２］　学習漢字まで使用

我が国のケンジュウ対策－ケンジュウのテキ発強化への取り組み－　ケンジュウのオウ収状キョウをみると、その大半はミツ輸入されたものであり、国際化の進展に伴い、今後、さらに海外からの流入の増加が危グされます。また、ここ数年、暴力団以外の者からのオウ収ガ増加しており、ケンジュウの一パン社会への拡散が進むこともケ念されます。

［処理結果３］　常用漢字まで使用

我が国のケン銃対策－の摘発強化への取り組み－ケン銃の押収状況をみると、その大半は密輸入されたものであり、国際化の進展に伴い、今後、更に、海外からの流入の増加が危グされます。また、ここ数年、暴力団以外の者からの押収が増加しており、ケン銃の一般社会への拡散が進むことも懸念されます。

4.2 分類語彙表の増補

4.2.1 目的

国立国語研究所資料集６『分類語彙表』が昭和３９年３月に刊行されていらい、現在２８版をかさねる。研究所の刊行物の中ではもっとも発行部数が多い。一般の表現辞典としての利用が多いためだろうが、言語研究への利用も少なくない。宮島達夫・小沼悦は「言語研究におけるシソーラスの利用」（国立国語研究所報告 104、平成４年３月）で『分類語彙表』を言語研究に利用した論文１１９例を集めて解説している。そこに掲載されなかった論文の他、直接研究の対象や手段にはならなかったが、参考、目安として使われた研究など、『分類語彙表』を直接間接に利用した研究はこの何倍、何十倍にのぼると思われる。

『分類語彙表』の収録語数はおよそ３万２千６百である。これらの語は国立国語研究所報告 21『現代雑誌九十種の用語用字』第一分冊の語彙表に掲げる使用率の高い語、さらに阪本一郎氏の『教育基本語彙』など日常生活でより基本的な役割をはたしている語である。これを研究に用い、あるいは詞藻辞典として用いるには語が少ない。そこでこれを増補し、収録語数を約６万語とする。（中野（1993）から引用）

『分類語彙表［フロッピー版］』（国語研、1994）は、形態素解析や構文解析など各種の言語情報処理に用いられている。さらに本研究で行うよりよい表記・表現の生成の研究にはなくてはならない基礎データであるので、さらに増補しす

る。

4.2.2 本研究以前の研究経過

科研費を受けて作成した「『分類語彙表』形式による語彙分類表」（中野 1989）のデータ数は５万２千弱だった。次の科研費「言語研究におけるシソーラスの利用法」（代表：宮島達夫、平成元年度から平成２年度）では 60,784 語を得た。さらに国語研内の課題「分類語彙表の増補」を経て現在次表のように 83,407 語が増補の候補となっている。

	体	用	相	他	合計
抽象的関係	12810	8320	4228	127	25485
人間活動の主体	7178	0	0	0	7178
人間活動	19279	9491	3561	471	32802
生産物	7980	0	0	0	7980
自然	7613	1311	1038	0	9962
合計	54860	19122	8827	598	83407

4.2.3 本年度の研究成果

(1)検討資料の作成

これらの語をもとに、増補用の検討資料（平成６年７月１日本、略称 671 本、600 頁）をゼロックスコピーによって７冊作成した。

(2)検討項目

①現在、分類語彙表の大項目と小項目には見出しがついている。しかし、中項目には見出しがなくその表示に困ることがある。これを検討することとした。

②分類語彙表の番号によって中国語と日本語の語彙の対照研究を行った（中野 1995）。このような利用法について検討した。

③専門語・擬声語・擬態語の増補候補について仮番号をふる作業をしている。

④分類語彙表データの日本語処理への利用について意見を求めるため言語処理学会で、研究発表することとした（中野　1995）。

⑤慣用句の増補中である。

⑥分類項目間の調整を行っている。

(3)白表紙本の作成

平成元年に印刷してから６年たち、語数も３万語も増えた。これらの作業用に、あるいは関係者の意見を聞くためにも印刷する必要がある。内容と頁数は以下の通り。

第１分冊　解説：10 頁、本表：290 頁、付録：30 頁（「言語研究での分類語彙表の利用」）

第２分冊　索引：330 頁

4.3 機械による翻訳の高度化

4.3.1 目的

機械翻訳の処理結果や非母語話者が作成した日本語または英語をより自然な表現にするための基礎的な研究を行う。日本語のより自然な表記の生成については、 4.1 で研究を進める。本年度は、以下の研究計画を立てた。

4.3.2 日英翻訳分析と計算機修正実験 (中野洋・Jun'ichi Tsujii ・ Natsuko Holden)

人間が作成した日本語または英語を計算機を用いてより自然な表現にするための調査および実験研究を行う。

この研究の基本的な方法については、中野 (1993)、Holden(1989) に従う。

英語ができる日本語母語者の「天声人語」英語訳と、日本語ができる英語母語者の英字新聞の随筆日本語訳を収集する。これらをより英語らしい、または日本語らしい表記・表現にするためにどんな修正が必要か、また計算機によってどこまで修正できるかを明らかにすることを目的として調査研究を開始する。今年度は、準備調査として日本人が作った英語、英語使用者が作った日本語を収集し、以下の点について分析を開始する。

1) 忠実度　①語彙選択、②文法、③意味
2) 自然さ
3) 文体

4.3.3 機械翻訳のための訳語選択・省略照応などの研究　(John Phillips ・中野洋)

機械翻訳において生成される文をより自然な表現にするための基礎的研究を行う。研究内容は以下の通りである。

現在の機械翻訳システムは、原文と同じ構文でかつ元の言語の単語を個々に目的言語の単語に翻訳するという傾向がある。実際、このような方法だとある訳語が文章全体に合わなかったりして不自然な逐訳語が生れる可能性が常に存在する。ところが、機械翻訳の研究者は、このような適切な翻訳を生成するための研究はほとんどしていない。

個々の文で、ある語句をより適切な語句に訳すためには、文脈のどの面をてがかりにするのが有効であるか、それをいかに機械翻訳に適用するのかを研究することが重要である。

ここでは文脈を２つ分ける。ひとつは言語的文章（とくに省略や照応研究のために）、ひとつは記述された場面によってもたらされた文脈である。

以前、機械翻訳の文脈における意味表現から文章を生成する研究をしていたが、それはこの課題と一致する。文脈の分析と生成の研究法は、訳語選択、いくつかの可能な翻訳から、一つを選択するためには多分ちょっとした変更だけでこの研究に適用できるだろう。

私が関心を持っている二つの研究領域は、表現と参照表現と補助動詞である。西洋言語と日本語との翻訳における参照表現は、特に難しい問題である。もし代名詞が代名詞に、名詞が名詞に翻訳されるのなら、翻訳は常に理解不能であるが、自然な訳とはならない。普通、英語の代名詞は日本語では省略されるかまたは名詞句に訳される。省略するのと名詞句に訳すのとどちらが適当なのか。もし、名詞句にするならそれは言語的文脈に大きく依存する。つまり、何を参照するのか、それは視点か否か、最後に参照されたものをいかに得るかなどである。まさしく、どのように選択するかは研究されなければならないテーマである。

動詞の場合、英語あるいは西洋語ではしばしば単純な動詞はもっとも自然に日本語に翻訳される。たとえば、

She sent me a present ― プレゼントを送ってらった。

Did you walk here? ― 歩いてきましたか。

Mr. Nakano told me ― 中野さんが教えてくれた。

It rained ― 雨が降ってきた。

意訳する理由の一つは、テンスとアスペクト（「くる」と「いく」のような）である。しかし、もう一つの理由は、記述されたイベントおよびそのタイプにおける文脈で翻訳するからである。他の例で、好まれる翻訳である意訳は、補助動詞よりむしろ語形変化表で成される。たとえば、英語の自動詞はしばしば他動詞に訳される。能動詞は受動詞または使役動詞として訳される。すなわち、

The earthquake destroyed a lot of houses ― 地震で多くの家が壊れた。

Mr. Ishii told me ― 石井さんに聞いた。

これまでの機械翻訳におけるアルゴリズムの研究は、同じ内容の二つの文を作ることであった。今必要とされているのはどんな文脈においても、いくつもの翻訳の中から一つの翻訳を選択すること、その裏にひそんでいる要因を記述することである。

今年度は、次のことを行う。

①訳語選択がかかわる情報表示のためのロジックの計算機への導入

②日英両語の文法を得て訳語選択についてのアイデアを試みること。英語はＦＴＰで、日本語は奈良先端科学技術大学院大学の松本研究室から得る予定。

③日英の２ヵ国語テキストを比較することによって逐語訳ではない参照表現がどのような状況で行われるかを分析する。

（以上　John Phillips,　日本語訳：中野）

参考文献

国立国語研究所(1994)：『分類語彙表［フロッピー版］』（言語処理データ集５，秀英出版）

中野洋(1989)：『分類語意表』形式による語彙分類表（昭和 61 年度～昭和 63 年度科研費「大量データの収集と処理の研究」代表者：野村雅昭）

――(1991)：パソコンによる語の認定処理（国立国語研究所報告 103　研究報告集 12）

――(1993)：『分類語彙表』の増補（平成４年度国立国語研究所年報）

――(1993)：機械翻訳に望む日本語の質（電子情報通信学会研究会資料）

――(1995)：分類語彙表の増補とその利用（言語処理学会第１回大会）

中野洋　他(1995)：中国における流行歌の語彙（計量国語学 19 巻 8 号掲載予定）

Natsuko I. Holden(1989): Interferefce software for Japanese Writers of English (CCL report, UMIST)

John Phillips, Hiroshi Nakano (1993): Structural Change in Translation between Japanese and English (情報処理学会自然言語研究会報告 93－NL－94)

John Phillips(1993): Choosing the Right Word - Lexical knowledge & context in machine translation (Proceedings of PACLING)

John Phillips(1992): Lexical Choice in Machine Translation (IEICE NLC92-37)