

パラレルコーパスからの専門用語の抽出

田 中 康 仁  
愛 知 淑 徳 大 学

専門用語の抽出の色々な問題点を調べると同時に、専門用語抽出の技術的方法 コーパスからの専門用語の抽出、特に用例文の抽出とその評価について述べた。また専門用語を体系化し整理すること、その応用等についても述べた。最後に日本の企業における専門用語についても述べた。

Extraction of Technical term using Pallarel corpus

YASUHITO TANAKA  
Aichi Shukutoku University  
9-Katahira Nagakute Nagakute Aichi  
〒480-11 JAPAN

Technical term Bank has a many difficult problems. One of problem is how to extract technical term using Text corpus. Sample sentence is impotant role of technical term Bank and How to extract Sample Sentence and How to evaluate these sentence.

Another problem is systematization of technical term and its applications. Finaly author described that Japanese company and technical term Bank.

## 1. はじめに

専門用語の収集、整理は単純な作業のようであるが、色々問題があり簡単なものではない。何が専門用語か？どのようにして効率よく集めるか？対応付ける外国語の妥当性、全体としてどの程度集めたらよいのであろうか？人間が使用する辞書であるのか（概念語中心か）それとも機械処理用に網羅的に集めるか？など、色々検討しなければならないことが多い。ここではパラレル・コーパスの中から機械処理用専門用語の抽出を考えてみる。

## 2. 研究の意義

### 2-1 用語と定義

我々は一般用語や専門用語についての定義をあまり明確にしないまま使っている。

同一言語、同一文化、同一専門分野の人々が言葉や論文を使ってコミュニケーションする場合はあまり問題が発生しないが、異った文化の中では言葉の定義を明確にしておかなければ共通の議論の場がゆらぐことになる。1つの例を示しながら考えてみる。

私は1990年4月からイギリスのGuildfordにあるサリ大学へ3ヶ月間留学した。その時週末には旅行した。その中でイギリスの青年と話していると「貴男の奥様はイギリスにきているのか」「いいえ、私の妻は飛行機に乗るのが嫌いでイギリスには来ていないよ」「それではボートで来ればよいではないか」なぜ彼はボートと言うのであろうかと思った。私の感じ（定義）では小さい小船がボートで、大きな船が"ship"であると思っていた。これは日本人、アメリカ人の感じであって、ヨーロッパ人はボートは人が乗る船であり、shipは貨物に乗せるものであるというように使われているのである。これはイギリスの研究者から説明されてはじめて判ったことである。我々の感じが全て他国の人と同じ言葉の定義ではないのである。ロンドンでは地下鉄はundergroundであり、subwayは地下通路である。日本、アメリカと全く異っている。これは有名なことであるので別に問題ないが、時に言葉の定義まで逆のばらなければならぬことに合うことがあった。この意味から用語は定義を明確にしなければ単に対訳語集を作っただけでは意味がない。

### 2-2 専門家と用語

専門用語は専門家にまかせておけばよいのであるが、専門家は彼ら自身の研究や、仕事に忙がしく、用語を集め体系化することの意義は判っても、行うとはしない。また他の分野の人々がその専門分野のことを学習しようとしても長期間時間がかかるし、

効率的でない。一方で言語を機械で処理する要求は強く、日本語ワープロの仮名漢字変換の精度向上、機械翻訳システムの精度向上、専門用語集の出版等がある。

そこで、情報を取り扱う者としては専門家の労力を最小限度にとどめ、専門用語を収集し、機械可読辞書の作成方法、実現を考えなければならない。

### 2-3 企業の壁

専門用語は企業内部で使われている。しかし、これらのドキュメントを使用するにあたっては著作権の問題が常に発生する。また企業は利益を追求する集団であるため、専門用語集の公表はなかなか行わない。

このような中で、何か公開された資料の中から専門用語の抽出を行わなければならない。専門分野に精通していない人が行ってもできるためにはパラレル・コーパス等を利用する方法しかない。

### 2-4 専門用語の公表と一般化

専門用語の収集を考えると色々問題が多いことが判る。しかし専門用語を収集し、整理し、定義づけを明確にしてゆくことは、その分野の学問体系を整理し、明確化することである。

この意味から専門用語の収集は学問の体系化といっても過言ではない。専門用語の収集整理は機械翻訳システムや人間による翻訳等にも大きな助けになる。これは専門分野のより一層のコミュニケーションの円滑化や国際化にも結びつくことである。

このためにも専門用語の収集公表と一般化は重要な意味がある。

## 3. パラレル・コーパスの利用

ある分野の専門家以外の人々が専門用語を取り扱うには専門分野のドキュメントを入手しなければならない。しかも、それは2つ以上の言語で同一の内容が書かれ、信頼のおけるものであり、公表されており著作権についてもあまり問題なく、安価に入手できるものでなければならない。

ドキュメントを階層的にみると次の図1のようになる。

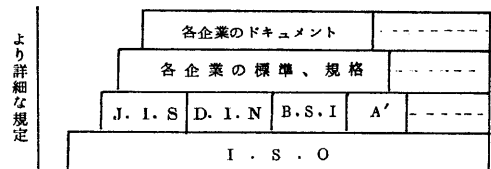


図1

つまり I. S. O で世界的に定められたようなものがあり、その上に各国の規格書があり、各国の協会で作成されたドキュメントがあり、企業のドキュメントがある。

そこで、ここでは日本の規格書（日本語版と英語版）を使用して専門用語の抽出を考えることにする。

日本の規格書がどの程度翻訳されているか調べてみると次の表 1 のようになっている。

◎ JIS 全収版

部 門	収録規格数 (規定を含む) (7桁以内)	冊 数	定価(円)(本体10%)
A 土木及び建築	559	17	209,060(202,000)
B 一般機械	1,266	34	601,250(584,000)
C 電子機器の取組	793	20	412,900(400,000)
D 防 護 軍	359	8	143,170(139,000)
E 汽 油 機	221	5	74,160(72,000)
F 鉄 道 機	555	11	194,970(189,000)
G 鉄 道 機	326	10	178,130(171,000)
H 非 鉄 金 属	396	8	139,820(134,000)
K 化 学	1,552	31	602,640(605,000)
L 電 機 機	312	6	110,210(107,000)
M 鉱 山	227	8	124,590(121,000)
P パルプ及び紙	103	1	15,480(15,000)
R 製 紙	237	4	71,670(69,000)
S 工 用 車	260	7	120,310(117,000)
T 医療安全用具	263	3	76,290(74,000)
W 汽 車	99	4	74,180(72,000)
X 情報処理	140	6	118,450(115,000)
Z その他	702	18	319,300(310,000)
計	8,489	196	3,046,370(2,978,000)

(上記価格は、平成2年3月31日まで有効です)

◎ 英訳 JIS 全収版価格表

部 門	規格数 (7桁以内)	冊 数	定価(円)(本体10%)
A 土木及び建築	337	9	846,810(828,000)
B 一般機械	306	10	756,110(737,000)
C 工具工作機械	251	7	499,290(475,000)
D 一般機械	384	10	794,120(771,000)
E 電気・磁気材料	262	7	557,220(541,000)
F 電機・風動	168	7	424,360(412,000)
G 電 子	149	6	395,320(384,000)
H 汽 油 機	224	5	427,450(415,000)
I 鉄 道 機	80	2	94,780(92,000)
J 非 鉄 金 属	201	8	446,850(435,000)
K 汽 車	240	6	360,500(350,000)
L 工業用・石油 ・油類・塗料 ・プラスチック ・セラミック・金属材料	247	6	529,420(514,000)
M 紙	334	10	629,330(611,000)
N 製 紙	84	2	133,900(130,000)
O 山 岳	82	2	191,380(186,000)
P 用 車	185	4	294,560(286,000)
Q 情報処理	58	2	177,160(172,000)
Z その他	413	11	782,800(760,000)
計	4,005	114	8,129,190(7,899,500)

(上記価格は、平成2年3月31日まで有効です)

そこで、この中で筆者の専門分野とは異っているが、自動車を取りあげてみることにする。自動車を取り上げた理由はサリ一大学で自動車の専門用語のターミノロジー・バンクを作成していたからである。又、友人のポーランド人 Dr. Zenon Grabaczyk が関心を持っているからである。

#### 4. 専門用語の抽出方法

専門用語の抽出を色々な方法で行っている研究を検討してみる。

##### 4-1 基礎的用語を用いる方法

一般に語は基礎的用語と専門用語に分類することができる。つまり次の式がなりたつ。

$$\text{語} = \{ \text{基礎的用語} \} + \{ \text{専門用語} \}$$

我々が使用している語の中から基礎的用語を取り除けば専門用語が抽出できる。しかし、基礎的用語

とはなにかを明確にしなければならない。

一般的には各分野のドキュメントを分析し、その中の高頻度語を基礎的用語としている。また、小学校、中学校程度の生徒が学習する用語で約 6,000 語程度である。英語の基礎的用語を作るとすれば、アメリカン・ヘリテージ社の語彙分析表が有用であろう。

しかし、基礎的用語を組合せた用語の中にも専門用語がある。

例 black hole

##### 4-2 統計的方法を用いる方法

統計的方法により専門用語を抽出する。

これは色々な分野のドキュメントを分析した基礎的用語の頻度分析データを用いて行っている。専門分野の用語の頻度分析データと基礎的用語の頻度分析データとを比較し対象としている分野の特出している用語又は語の組合せ等の頻度やある係数を分析して専門用語を抽出するものである。

これらの方法は機械的処理が中心で、人間の判断に頼らなくても専門用語の抽出ができる。

しかし、最終的な検討は人間が判断しなければならない。

##### 4-3 既存辞書の整理

我々の身のまわりには色々な目的で作られたハンド・ブック、辞書、専門用語辞書がある。又、専門書の最後に付いているインデックスを利用するのも一つの方法である。これらをまとめ専門用語のデータ・バンクを作ることができる。

このようにして収集した専門用語をもとに各種ドキュメントに照合し、例文、説明文、定義文の抽出に役立てることができる。このようにして専門用語のタームバンクを充実させることが可能である。

##### 4-4 品詞列による専門用語の抽出

専門用語の品詞列の構成は形容詞と名詞の組合せである。前置詞を含むものもあるが、それらはごくまれである。そこで形容詞、名詞の連続している部分を抽出し、しかも、それらが基礎的用語でないものを抽出すると、残りのものはほぼ専門用語である。最終的には人間の判断が必要である。

4-1 から 4-4 まで専門用語の抽出方法について述べたが、これはあくまでも大量のドキュメントから専門用語らしきものをマークするためのものである。最終的判断は人間の検討が必要である。

また、ドキュメントの中に書かれている語形が専門用語とはならないので一部修正しなければならないものもある。

## 例

- ①複数 ②定冠詞の削除 ③大文字

また、機械可読辞書に追加するためには、アクセント、発音、詳細な品詞、分野区分……を追加しなければならない。

### 5. 専門用語の収集整理について

専門用語の収集について別の視点から幾つかの問題点をながめてみよう。

#### 1) 既存の辞書を基礎に作成する方法

- ① 著作権の問題を解決しなければならない。
- ② よく知られた用語、高出現の用語には良い、単純な専門用語が収集できる。
- ③ 複雑な専門用語、複合語が収集できにくい。

#### 2) 文章、コーパスからの専門用語の収集する方法について

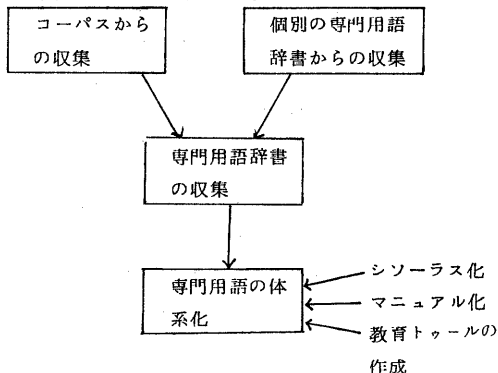
- ① 専門用語のサンプル文が得られる。
- ② 頻度情報が得られる。
- ③ 実際に使われている専門用語、最新の用語が収集できる。
- ④ 網羅的に収集できる。
- ⑤ 分野を限定して収集することができる。
- ⑥ 専門用語が実際の文章で使われている状況を知ることができる。

複数形、修飾語の結合、その他……を知ることができる。

#### 3) 文章、コーパスからの専門用語の自動抽出にあたっての重要な点を述べる。

- ① 自動抽出の各種の Tool の作成
- ② 専門用語の最終確認のための人間-機械系の会話 Tool の作成 Editor 機能、search 機能 File 機能等がある。
- ③ コーパスの作成、又は入手方法著作権の問題  
コーパスはパラレルコーパスが望ましい。

これらを総合し専門用語の抽出は次のような方法を体系化することが望ましい。



### 6. 用例の抽出

専門用語辞書は専門用語と属性と対訳語だけでは不十分である。その専門用語がどのような状況で使われているかを示す用例を集めねばならない。この用例について検討してみる。

#### 6-1 用例の抽出

用例の抽出にあたっては対象となるコーパス（文の集合）と対応させ、コーパスの中から用例を抽出するようにしなければならない。

人間が作為的に考えた例文は良いものではない。コーパスはパラレル・コーパスであれば対訳語の抽出、対訳語の妥当性の検証に役立つ。

#### 6-2 用例の抽出、整理

用例の抽出、整理は多くの場合、カード等へ書き出して行われているが、これらの作業も、機械処理を考えるべきである。ディスプレイ画面上で簡単な操作で用例の検索抽出、整理が行われるようにすべきである。

#### 6-3 用例の評価方法

専門用語の用例が 1 例しかみつからない場合は評価付けの値を付けてもあまり意味がないが、複数の用例が見つかった場合その中のどの用例を採用するか、判断する基準を作成しなければならない。ここでは、幾つかの基準を考えてみる。

##### 1) 文の種類

専門用語を含む文が用語の定義文であるか、説明文であるか、特性や、属性を説明しているものであるか、ただ単純に専門用語を含む文かを調べる。

これは専門用語の位置と動詞の種類や表現形式等で判別することができる。

##### 2) 文の長さ

用例は簡潔であることが望ましい。長文の最後に専門用語が出てくるのでは、長文を読んだり理解するために時間がかかる。それ故短文が望ましい。又、辞書を作成する場合はデータの総量に影響する。

##### 3) 文中の専門用語の位置

用例中の該当する専門用語は文の最初の方に出現することが望ましい。もし、文中の最後の方に出現すると読解に時間がかかる。

##### 4) 用例文中の専門用語の数

一つの用例の中に数多くの（3 つ以上）専門用語が出現すると、文を理解することが難しくなる。

## 5) 意味別の用例文

専門用語は一つの意味しか持たないのが通常であるが、複数の意味で使われるのであれば、その意味別に用例を集めなければならない。

## 6) 品詞別の用例文

専門用語が複数の品詞を取り扱う場合は、品詞別に用例を収集すべきである。

## 7) 用例文の数

用例の数を何回まで採用するか検討しなければならない。

最後にこれら 1)~6) の 6 つの項目にウェイト付けを行いある評価関数を作り、これにより用例を評価し、評価の良いものから順に並べ、その中から用例を決定する。これにより人手を減らし、用例評価と選択を迅速に行うことができる。これらの項目はサリ大学で学んだことである。

## 7. 専門用語とシソーラス体系

### 7-1 概念体系の整理

専門用語を集め、語彙を増加させることは重要なことであるが、これを何からの体系化することを考えなければならない。1つの体系としては概念体系への整理である。別の整理の方法は人間の理解を助ける本(ハンドブック、解説書、詳細なドキュメント)として、体系化することであろう。又、情報検索用のシソーラス体系であるとか、機械処理用の体系等が考えられる。

### 7-2 属性の付加

専門用語は見出し語だけではなく、色々な属性が備っていないなければならない。それについて述べる。

- |                     |               |
|---------------------|---------------|
| ① 見出し語              | ⑤ 用例文         |
| ② 読み                | ⑥ 意味          |
| ③ 品詞                | ⑦ シソーラス体系への対応 |
| ④ 対訳語(英語、<br>ドイツ語…) | ⑧ 分野区分        |
|                     | ⑨ その他         |

等が考えられる。

これら属性も系統的に集めなければならない。

## 8. 専門用語の統一

### 1) 専門用語の発生

学問を推進している研究者や技術者はどのようにして専門用語を作り出しているのだろうか。

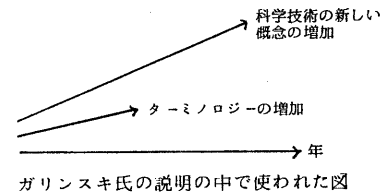
研究者達は自然現象や社会現象を観察し、データを集め、分析を行い、仮説を立てて、各種の現象を理論的に説明する方法を作成し検証しているのである。その中で今までとは異なった概念や物質が新しく見つけ出されたり、作り出されている。これらに

ついて新しい名前が付けられているのである。これが専門用語である。それゆえ、今までにない新しいものであるため旧概念や物質に付けられていた専門用語を幾つか結合させて新しい概念や物質に付けることがしばしば行われる。また旧概念の内容を変更し、専門用語はそのまましてしまう場合もある。先端分野では研究が十分進まないままに概念や物質に用語が付けられるため、いろいろな混乱が発生する。

さらに研究者の所属を考えると大学等の公的な研究機関で研究している人々と、企業の研究者とでは少し性質が異なってくる。企業では多くの資金を投入し研究を行っているため、その成果を独占し、実用化をはかりたいという願望が働く。このため競争会社の専門用語とは異なった専門用語を作るとか、特許文の専門用語のように特許の請求範囲を拡大できる専門用語を付けておく等さまざまな力が働く。

このため専門用語の統一や概念の規定とは逆方向の力が働く場合もしばしば見受けられる。このため同音異義語、同形異義語、同一概念に複数の専門用語等さまざまなことが発生し、後で混乱の原因となっている。このように混乱の発生する原因を充分把握しておかないで専門用語の議論をすることはできない。

専門用語の増加と科学技術の新しい概念の増加の関係をガリンスキ氏は1987年ドイツのトリアーの会議で次の図で説明している。



科学技術の進歩はめざましいものがある。これにつれて新しい概念が発生して来ている。しかし、ターミロジー活動がこれにともなって充分なされていなければ、同音異義語、同形異義語、多義語が発生する。これは社会に混乱を起すことであり、正しい情報伝達ができなくなってしまふ。この意味からも専門用語の整理、造語活動、造語の普及に力を入れてゆかねばならない。

### 2) 専門用語の統一

専門用語が発生する初期にはいろいろな混乱も起るが、次にあげるような活動の中で少しずつ統一が行われる。

### (1) 研究者間のコミュニケーション

研究内容は論文として、また、報告書として発表される。この時研究者が勝手な専門用語を新しい概念や物質に付けていたのでは相互の意志疎通が行えない。このため相互理解のために専門用語の統一化が行われる。

### (2) 商品や技術の販売

企業は自分達が開発した商品や技術を広く利用してもらい利潤をあげようとする。このためには業界で概念を定義し、専門用語の統一、規格の統一を行っている。しかし、これは利潤をあげる側面では強く働くがその他の面では専門用語の不統一をまねいている。我々はA社の商品とB社の商品で専門用語の意味が異なっていて使い方に不便をきたすことがある。通産省の下部機関である日本規格協会を中心に専門用語の統一化が行われている。しかし、完全に網羅的に行われているわけではない。国際的にはISOを中心に専門用語の統一化、定義が行われている。また、規格の統一も行われている。もしこの規格に合っていない商品やマニュアルを作成しても世界の市場から受け付けられない。

### (3) 科学技術教育

科学技術を広め利用してゆくためには大勢の科学者や技術者、職人等の人々が必要である。これらの人々に統一的な専門用語を教育し育てておかなければ、一つの専門用語に対しある人はaという物を想定し、また別の人はbという物を想定し行動するということが起こる。これでは生産現場や工事現場や実験で混乱が生じてしまう。特に多民族の集団で多言語の集団をコントロールする場合には大変である。日本では方言はあるものの標準語が学校教育で教育されているし、文化や習慣が大幅に異なるという集団もあまりない。この点では共通の基盤が有るといえる。

専門用語に関しては文部省を中心として学術用語集の制作が行われ、基礎的概念語の整理が行われている。しかし学術用語集に対して一つ注文を付けてと専門用語(日本語、英語対訳)、読み、ローマ字表記だけである。専門用語の定義、解説、説明といったものが無い。今後はこれらを追加し充実させてほしい。学術用語集は基礎的概念語に限っているため数が少ないが、今後は専門用語(概念)の数を増加させてほしいものである。いろいろ注文はあるが学校教育や学会での専門用語の統一、整理に大きな貢献をしている。

### (4) 科学技術文献の流通

科学技術の発見・発明・応用は個人の秘密としておくものではなく、論文、技術報告、特許という形式で公表される。これら論文等を書く際には、ある程度の知識のある専門家が持っている共通の専門用語で書かなければならないし、新しい用語は意味を定義して書かれる。このためにも専門用語は共通性を持たなければならないし、統一化への努力が行われる。もし独特の専門用語を使用しても、それが認めなければ論文の価値が減り、研究も発表しにくくなる。

### (5) 専門用語の統一と国際化

専門用語は日本だけで発展しているものではなく、世界各国の研究、技術開発とともに作られるものである。このため専門用語は世界各国の専門用語と同一概念で結びつけられる必要がある。専門用語の訳語を付ける必要がある。例えば英語(米語)、フランス語、ロシア語、中国語、スペイン語、アラビア語……等と結びつけていかなければならない。文献の翻訳、論文の流通のためにも訳語が必要である。

このようにいろいろな側面から専門用語の統一化が行われている。しかし、これは積極的に行うものではなく自然淘汰にまかされたものである。今後はこのようなものではなく、積極的に専門用語を取り扱い、普及・発展させてゆく方向をみつけ出さなければならない。

## 9. 試行システム

専門用語辞書を作ってみる中で、実際の問題点を整理しなければならない。そこで、自動車の規格書とその対訳から専門用語の抽出整理をカードを使って行っている。

又、筆者の留学したサリー大学では自動車メーカーと協力し、機械可読の専門用語データバンクを作成している。このプロジェクトは10名程度の研究員を使い英(米)語、フランス語、ドイツ語、スペイン語、ギリシャ語等のマルチリンガルの機械可読の専門用語データバンクを作成している。これは今後の研究開発の参考になるシステムである。

## 10. 専門用語データ・バンクの応用分野

専門分野の用語を網羅的に収集し、整理することはいかなる意味があるか、又、その応用分野を述べてみる。

1. 機械翻訳システム、人間と機械による翻訳システム
2. 概念辞書の作成

機械用、人間用

3. ある学問分野の整理と体系化

4. 用語学、知識工学等の発展

概念と用語の関係 概念、用語と文化、歴史

概念と知識の関係

等を研究するための資料となる。

## 11. 企業と用語

### 11-1 B. M. Wの場合

南ドイツの都市ミュンヘンにB. M. Wという自動車メーカーがある。

B. M. W では用語の管理を次のように行っている。

1. 専門用語についてドイツ語を主見出語として英語(米語)、スペイン語、スウェーデン語、フランス語、イタリア語、オランダ語(米語については英語と異なる場合のみ)の6ヶ国語の対訳語タームバンクを作成している。

役40,000語を収納している。このほかドイツ語→英語のサービスマニュアルの用語を管理している。また日本の翻訳会社と協同開発してドイツ語→日本語の自動車に関するタームバンクを開発している。

B. M. Wの用語の担当者は次のように言っている。「In business, knowledge means power in winning the race……」これは印象的である。

ヨーロッパにおいては国家と国家が隣接しており、多くの戦争や侵略を繰り返してきた歴史の中では、互に用語を明確に定義しておくことは相手を理解する上で重要なことである。相手の国の言葉にマニュアルを翻訳し、技術者を育て、商売をしなければ、商品が売れにくいのである。

これは1つの会社の例であるが、ヨーロッパの大きな企業では用語の管理にかなりの費用をさいている。

### 11-2 日本の企業と用語

日本の企業は戦後アメリカからの技術導入が続いたため、企業の顔がどちらかというとなアメリカに向けられている。このため専門用語のタームバンクを持っている企業でも、英語(米語)→日本語の専門用語のタームバンクで、しかも見出しだけである。用例や、用語の定義等はほとんど付いていない。

日本の企業では、英語は世界の共通語であるから、英語が話せ、英文のドキュメントがあれば世界に通じると一般に信じられている。

しかし、これは一面では正しいが、世界の全ての人々が自国語と英語を話すとは限らない。

また、第二外国語のレベルもさまざまである。日本の企業もヨーロッパに目を向け、日本語→英語(米語も含む)ドイツ語、仏語、スペイン語、オランダ語、イタリア語、ポルトガル語…等についての専門用語データバンクに関心を示していただきたい。これは日本が世界を相手に貿易や技術交流を行う上での重要な要素であろう。

専門用語データバンクを作成するにあたっては多くの費用と年月、労力を必要とする。しかし、作らなければならないものである。

## 12. おわりに

この研究は1990年4月から3ヶ月間英国サリー大学に留学中に学んだことと1990年7月から3ヶ月強ウインに滞在し、ヨーロッパ各地で開かれた国際会議に出席する中で考えたことである。

又、ポーランドの友人、Dr. Zenon Grabaczykと共同で自動車用語辞書の開発を行うと計画する中で考えたことである。

専門用語の抽出、整理、体系化は大変な問題であるが、この理論、研究の方法論、応用分野等を1つ1つ丁寧に研究開発してゆかねばならない。

## 13. 参考文献

1. 日本規格協会「最新海外規格ガイドブック」