自然言語処理 48－1
（1985. 3. 27）

VOX：英語語い抽出分析システム

末松 博　　寺本 雅則
（日本電気株式会社 ソフトウェア生産技術研究所）

1. はじめに

制限英語開発の自動化手法として、磁気化された英文マニュアル中の語いとその出現頻度を分析するVOXシステム（VOcabulary eXtracting system）を開発し、いくつかのマニュアルに適用したので、システムとその適用結果に関し報告する。

制限英語は、文章作成時に、作成者に対し、単語の用い方、ターミノロジー、構文等に一定の指針を与え、一定の範囲の語い、構文で書かせる手法である。制限英語により書かれたマニュアルは、平易な文章であり、英語圏以外の読者でも容易に内容を理解できることから、他言語への翻訳は不要である。従って、単に翻訳量の問題ばかりではなく、専門用語の訳語の考案および統一の問題や、翻訳によって起こる発刊時期のずれの問題、誤訳の問題等を避けることが出来る。この試みは、米国キャタピラ・トラクタ社において成功を納めており、コンピュータメーカや、電話交換機メーカにおいても試行されている。

制限英語開発には、次の手順が必要とされている［2、3］。

（1）英文マニュアルを読み、出現する各単語に対し動詞、名詞、形容詞等に分類する品詞分析を行なう。
（2）品詞毎に単語の表を作成する。
（3）各単語の使用頻度を計数する。
（4）対象としたドキュメント内で通常の意味とは異なる専門領域の単語や熟語（技術用語）の範囲を調査する。
（5）使用頻度の高い語とそれらの同義語から最も適した語を選ぶ。
（6）選んだ語を用いて文章を表現して、その語の妥当性を確認する。

VOXシステムは、（1）～（4）の処理をサポートする。また、品詞別語い曲線を提供して、語いの飽和状態を確認することにより、調査したマニュアルの量が十分であるかどうかが判断できる。（5）、（6）のサポートは今後の課題である。

次にVOXシステムの機能について、詳細に述べる。

2. VOXシステムの機能

VOXシステムの機能は主に次の3つである。

（1）品詞別語い表作成
　品詞ごとに、出現した単語を頻度と共にもれなくリストアップする。
（2）技術用語表作成
　単語および熟語形式の技術用語表を頻度と共にリストアップする。
（3）品詞別語い曲線作成
　品詞ごとに、語いの飽和を示す、品詞別語い曲線を提示する。

2．1 品詞別語い表作成

12の品詞別語い表を作成する（表1）。品詞は、動詞、名詞、形容詞、副詞である非機能語と、冠詞、助動詞、BE動詞、接続詞、前置詞、代名詞、所有名詞、従位接続詞である機能語に分割できる。非機能語は内容語とも呼ばれ、文章の内容を伝える役割が主であるが、機能語は主に文法的役割を果たす。

表1 品詞別語い表

非機能語	動詞語い表 名詞語い表 形容詞語い表 副詞語い表
機能語	冠詞語い表 助動詞語い表 BE動詞語い表 接続詞語い表 前置詞語い表 代名詞語い表 所有名詞語い表 従位接続詞語い表

各語い表は、小文字で表現されており、アルファベット順に単語が並び、その左には、使用頻度が示される。

動詞は、規則動詞、不規則動詞も全て原形化される。進行形、過去分詞形の形容詞や、動名詞も原形化されて、動詞語い表に加えられる。名詞は、複数形は単数形に置き換えられる。形容詞に語尾lyを付けて副詞としてあるものは、そのまま副詞語い表に登録した。形容詞には、比較級、最上級があるが、原形化は行っていない。

なお、名詞語い表に登録される名詞と、形容詞語い表に登録される形容詞は、単独に現れるものであり、形容詞＋名詞や名詞＋名詞のパターンは、熟語形式の技術用語表に登録する。

2．2　技術用語表作成

2つの技術用語表を作成する。

（1）単語形式の技術用語表

（2）熟語形式の技術用語表

この場合、小文字化は、行なわず、表は、アルファベット順にソートしてある。熟語形式の技術用語表は、複数形と単数形の両方が出現したときにのみ、単数形に直してある。これは、技術用語には大文字を含んだ表現が多く、かつ固有名詞的なので、できるだけ原形を保つのが良いと判断したからである。

単語形式の技術用語表は、ほとんどが、頭字語（SE，CPUなど）である。技術用語用の語いとして通常の単語を選ぶこともできる。しかし、現段階では、そのような単語が、文章中で通常の意味なのか、技術用語としての意味なのか、表面からは判定しにくく、この分析を行なっていない。つまり、技術用語用の語いは、現在、各品詞別語い表に分散して含まれている。

品詞別語い表および、技術用語表は、いずれもコンピュータにより自動出力されるものが3．1節の理由により、完全ではないため、容易に修正できる専用エディタが開発してある。このため、より人間の理想に近い品詞別語い表が作れる。

2．3　品詞別語い曲線作成

マニュアルを分析して行くと、初めのうちは、出現する語いが急速に伸び、次第にその伸びは緩やかになり、やがて、もう新しい単語がほとんど出現しない飽和状態になる（図1）。さらに、後述のように、各品詞について見ると、その現象は明確に現われる。接続詞や前置詞のような機能語は早く飽和し、動詞、形容詞、名詞、副詞といった非機能語は、飽和しにくいと言える。

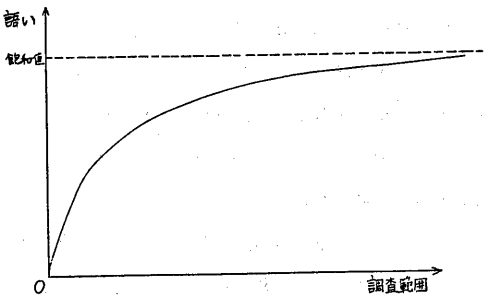

図1　語いの飽和

マニュアルの調査範囲がかなり大きかったとしても、さらに調査を進めると、まだ新しい動詞や形容詞が頻繁に出現して来るようでは、調査範囲が十分であったとは言えない。各品詞が飽和状態に達しているかどうかを見ることは、調査範囲を判断する目安となる。

また、制限英語で書かれたマニュアルとそうでないマニュアルを比較すると、前者は、早く飽和状態に達するはずである。英語が母国語である人が書いたマニュアルとそうでない人が書いたマニュアルを比較すればまた違った結果が現われるはずである。

さらに、各品詞の語い数の変化を比較することにより各品詞の言語に対する位置が明確になる。

このように、品詞別語い曲線を採ることにより、

（1）マニュアルの調査範囲が十分であるかどうか

（2）各種マニュアルの表現方法の違い

（3）各品詞の言語に対する位置

が判断できる。

品詞別語い曲線は、マニュアルの語数とその語数に至るまでに出現した各品詞の語いの累積を対応させたものである。

3．　VOXシステム基本構成

次に、VOXシステムの処理概要を図示し（図2）、以降、その中の自動処理部分における各手段について説

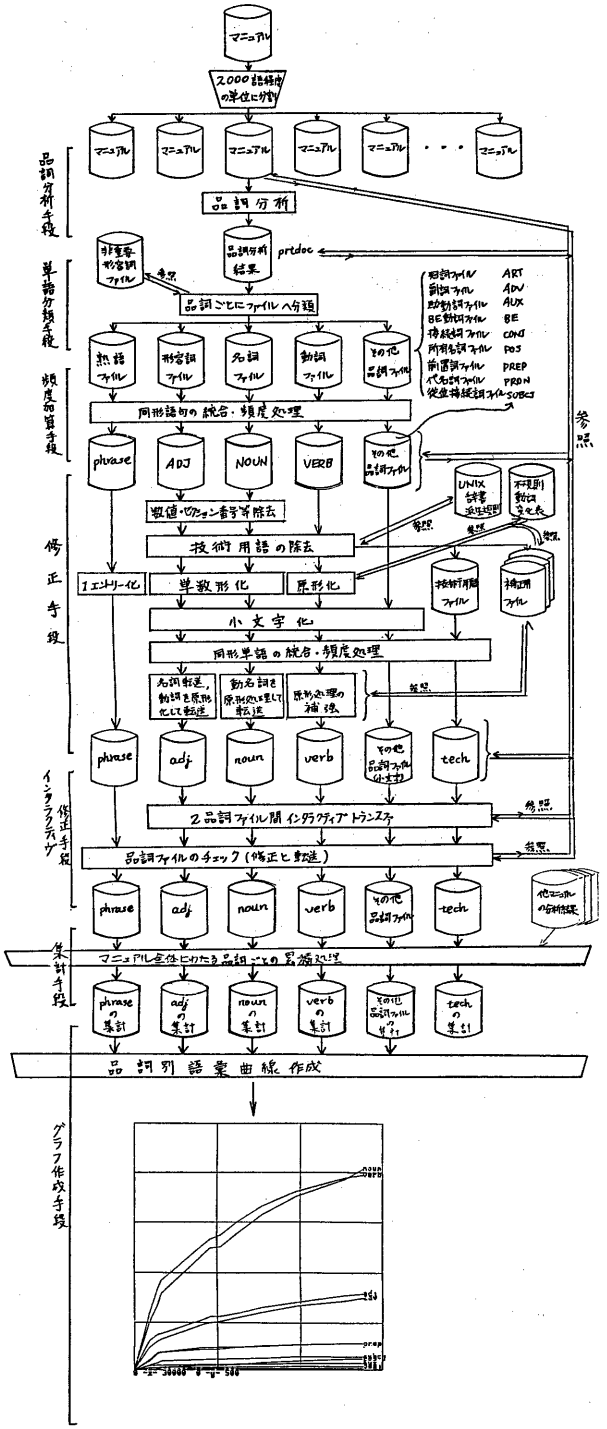

図2　VOXシステムの処理概要

明する。

3．1　品詞分析手段

入力されたマニュアルに対し、品詞分析を行なう。

ＵＮＩＸには文体に関して統計を出すstyleプログラムがあり、これは、partsという品詞分析プログラムを使用している［4］。ＶＯＸは、これを使用している。partsは、小さな辞書（機能語、不規則動詞）、接尾辞規則、英文法を備えており、単語の文における主語、動詞といった機能を見て品詞付けを行なっている。精度は、約95%である。

3．2　単語分類手段

partsの品詞付けに従って、単語を１２の品詞別語い表に分類する。同時に、熱語（技術用語）の抽出を行なう。

名詞と形容詞以外は、そのまま各品詞別語い表に登録される。しかし、形容詞（…）＋名詞のパターンは、技術用語の可能性が大きいので、熱語形式の技術用語表に登録される。この際、形容詞が、

（1）数詞（one，two，……）

（2）数詞的意味のもの（many，entire，……）

（3）感覚的表現（interesting，visible，……）

（4）文法的表現（following，such，given，……）

である場合は除外される。しかし、これらの形容詞を取り去っても後続して熱語がある場合、この熱語部分は登録される。上記のような形容詞や、独立に現れる形容詞、名詞は、それぞれ形容詞語い表、名詞語い表に登録される。

熱語として登録しない形容詞は、当社情報処理標準用語対訳集英和編（REG01-1）を参考にした。この対訳集では、上記のようなかなり基本的な形容詞を除いて、多くを技術用語の一部として認めている。ＶＯＸシステムも同様の方針を取った。熱語として登録したい、または、登録したくない形容詞がある場合、ユーザは、上記の形容詞が登録されているファイル（非重要形容詞ファイル）に対して、単語の追加、削除を行うことができ、技術用

語表への出力をコントロールすることができる。

3．3　頻度加算手段

語い表に生じた同形の単語や熟語を、1エントリー化し、その個数を使用頻度としてまとめる。また、既に、使用頻度が語い表にある場合で、同形の単語や熟語が存在する場合は、1エントリー化し、頻度同士を足し合わせる。

3．4　自動修正手段

a）不要単語除去部

数値、セクション番号（1.1.3など）等を、除去する。

UNIXspell辞書に登録されていない単語や、表面的に一般用語ではないもの（数字・カンマ・点を含むもの、子音のみから構成されるもの、等）は、技術用語と見なし、技術用語表（単語形式）に登録する。

b）原形化部

次のような原形化を行っている。

大文字の小文字化

複数形の単数形化

規則動詞の過去・過去分詞形の原形化

不規則動詞の過去・過去分詞形の原形化

進行形の原形化

動名詞の原形化

この原形化により、エントリの数を減少させることができる。原形化により生じた単語の品詞が元の品詞と異なる場合、単語は、自動的に、新しい品詞の品詞別語い表に転送される。

c）辞書部

parts用辞書

UNIXspellコマンド用辞書

非重要形容詞表

不規則動詞変化表

parts用辞書は、品詞分析プログラムpartsを実行する際に必要な辞書（機能語、不規則動詞）である。UNIXspellコマンド用辞書は、綴を調べるため、UNIXに備わっている辞書（派生規則も含む）である。技術用語の抽出に使用する。非重要形容詞表は、熟語形式の技術用語に登録すべきでない形容詞をリストアップしている。不規則動詞変化表は、動詞の過去形・過去分詞形に現在形を対応させたものである。不規則動詞の原形化に使用する。

d）派生規則部

UNIXspellコマンドの-vオプションは、派生規則を提供しており、VOXシステムは、これを使用している。この派生規則により、複数形、規則動詞の過去・過去分詞形、進行形、動名詞の語尾変化が判り、原形化を施すことができる。

3．5　インタラクティヴ修正手段

語い表を効率的かつ正確に修正するために、2つの専用エディタを用意した。

（1）語い表からマスターの語い表に載っていない単語を転送（move）

（2）語い表、技術用語表（単語、熟語形式）の修正、単語転送（trans）

（1）は、2つの品詞別語い表に相違する単語を抜き出して、それらの単語に対して（2）を適用しているので、基本的な機能は同じである。（2）は、品詞別語い表の最終的な修正に使用する。必要に応じて、マニュアル、品詞分析結果、初期の品詞別語い表、等をエディタ内部から参照しながら修正、転送を行なうことができる。単語の意味、その単語に対する品詞分析、原形処理、転送、等を確認でき、それに基いて修正できるので、正確に、効率的に語い表を作成することができる。

3．6　集計手段

いくつかのマニュアル（部分）に対して、できあがった品詞別語い表、技術用語表を、品詞ごとに集計する。同形の単語、熟語は、1エントリー化され、頻度が加算される。この際、品詞ごとに、各マニュアルの累積語い数が報告される。

3．7　グラフ作成手段

品詞別語い曲線を作成する。

UNIXのワードカウントプログラム（wc）により、マニュアルの語数を計測し、各品詞の累積語い数を自動

的に対応させて、グラフ用データを作成する。さらに、UNIXgraphコマンドをテクトロ4014上で使用して、品詞別語い曲線を作成する。

4. VOXシステム試用結果

4.1 適用対象マニュアル

次の3つのマニュアルに対して、適用した。

（1）マニュアルA……スウェーデン人作成の交換機関係のマニュアル

（2）マニュアルB……米国人作成の計算機関係のマニュアル

（3）マニュアルC……米国人作成の計算機関係のマニュアル

マニュアルAは制限英語によって書かれているとの情報があり、もしそうであれば、分析することにより制限英語用の語いが求まることになる。マニュアルBは、テクニカルライターによる英語である。マニュアルCは、UNIXコマンドの使用解説書である。これは複数のアメリカ人によって書かれた自然言語のマニュアルである。

なお、ランタイムは、総語数2万～3万語のマニュアルあたり、VOXシステム作成者本人で約1日の工数を費した。

4.2 マニュアル分析結果

4.2.1 品詞別使用語い数

表2 マニュアルA，B，Cの品詞別使用語い数

		マニュアルA	マニュアルB	マニュアルC
調査範囲（総語数）		27772	28479	26344
非機能語	動詞	395	342	415
	名詞	408	325	449
	形容詞	159	150	189
	副詞	151	105	152
機能語	冠詞	3	3	3
	助動詞	14	16	14
	BE動詞	8	8	9
	接続詞	7	11	10
	所有名詞	24	10	31
	前置詞	56	36	39
	代名詞	56	32	44
	従位接続詞	28	26	29
使用総語い数＊		1309	1064	1384
技術用語（単語形式）		292	274	396
（熟語形式）		988	1344	1058

＊技術用語を除く。

4.2.2 各品詞の占める割合

マニュアルの種類が違うにも関わらず、マニュアルA、B、Cは、極めて類似した結果を示した。

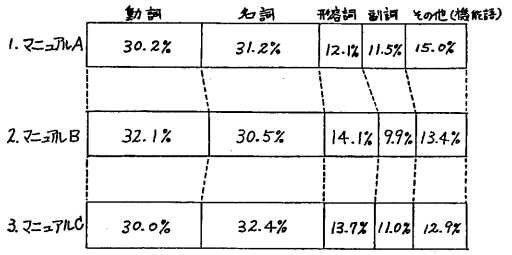

図3 3マニュアルにおける各品詞の占める割合

4.2.3 品詞別語い曲線

マニュアルAは、いずれの品詞もほぼ飽和状態にあり（図4）、従って、調査範囲は、ほぼ十分である。マニュアルBも緩やかな傾斜を見せ始めているが、弧の大きさが小さく、徐々にではあるが、未だ伸びる様相を呈している（図5）。しかし、マニュアルCは、いずれの非機能語も増加傾向にあり（図6）、調査をまだ行なわなければならないことを示している。

ここで興味深いことは、いずれのマニュアルの場合も、品詞が、語い数に関してよく似た動きを見せる3つのグループに大きく分かれたことである。各グループ内の品詞を比べるとその機能に共通性があることがわかる。

＜グループ1＞ ――― 動詞、名詞

名詞および動詞は、言語の中で基本的役割を果たしている。名詞は、表現する対象物の名前であり、動詞は、対象物の動作、関係を表わす。

＜グループ2＞ ――― 形容詞、副詞

形容詞および副詞は、上記のものに対する修飾の機能を果たす。

＜グループ3＞ ――― その他機能語

機能語は、文法的役割を果たす。

調査対象マニュアルがまだ3つであり、一般的だと捉えるには、まだ検証を必要とするが、各品詞の占める割合が、3マニュアルを通じてほぼ同じだったこと（図3）を考え合わせると、次のような仮説が成り立つ。

言語は、名詞、動詞の基本機能、形容詞、副詞の修飾機能、その他の文法機能の3つに大別でき、基本機能内

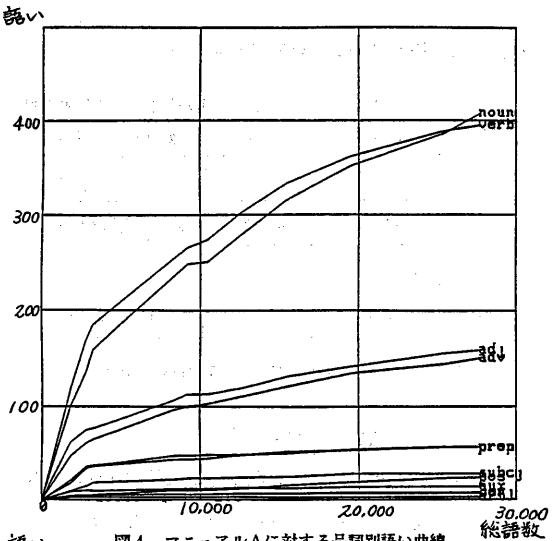

図4　マニュアルAに対する品詞別語い曲線

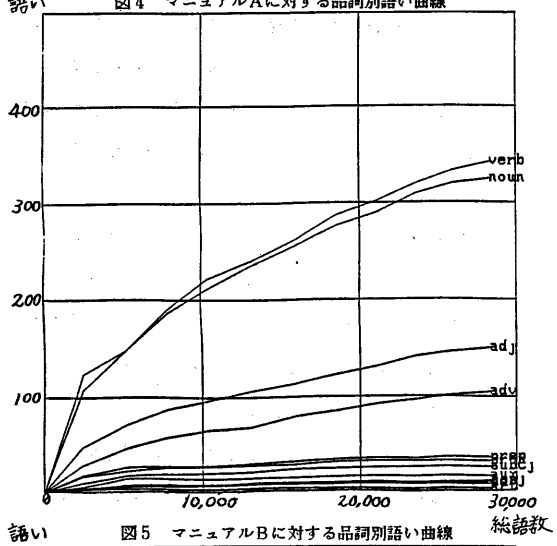

図5　マニュアルBに対する品詞別語い曲線

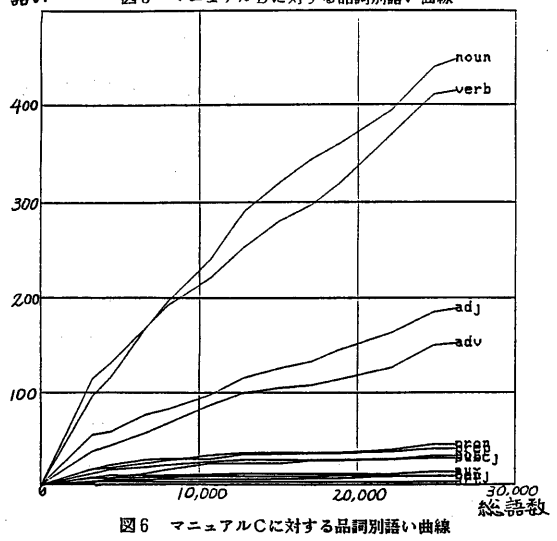

図6　マニュアルCに対する品詞別語い曲線

の2品詞は、ほぼ同じ割合で、また、修飾機能内の2品詞もほぼ同じ割合である。さらに、基本機能対、修飾機能対、文法機能の比率は、約6：2：1となる。

この仮説が真理であれば、これは、制限英語を設計する際に役に立つ。つまり、各品詞の比率がこれに合わなければ、まだ、削除したり、追加したりすることのできる単語が存在するということである。

しかし、飽和時における基本機能対、修飾機能対、文法機能の比率が一定であるということに対しては、まだ疑問がある。マニュアルCの品詞別語い曲線（図6）を見てわかるように、基本機能、修飾機能は、未だ増加傾向にある。従って、マニュアルの調査を続けると、文法機能の語い数はほぼ一定なので、6：2：1の比率が変わってくると思われる。最終的に、マニュアルCの基本機能と修飾機能の比率は、やや大きくなると考えられる。ここで得られた比率は、実は制限英語における比率ではないかと思われる。

3マニュアルは、ほぼ同じ調査範囲（総語数）であるが、マニュアルA、Bが緩やかな上昇傾向を見せたのに対して、マニュアルCは増加の勢いが衰えなかった。このことから、マニュアルA、Bは、何らかの形で制限されているのではないかと考えられる。

文献［1］には、交換機用の制限語いについて報告があり、技術用語以外の語いとして、約900語、技術用語の語いとしては、約400語設定したとある。この足した数は、マニュアルAの分析結果、約1，300語に一致する。このマニュアルは、制限英語で書かれ、求められた1，300語のうち400語は、技術用語用の語いであったのではないかと考えられる。

また、マニュアルAが飽和に達したことは、英語がスウェーデン人にとって、第二外国語であることに起因しているとも考えられる。そうであれば、自然言語で書かれたマニュアルCとの品詞の割合における類似性から、かなり進んだ英語教育を受けたと思われる。

マニュアルBをさらに調査したところ、6万5千語付近でほぼ飽和し、使用総語い数（技術用語を除く）が約

1,470語となった。この値は、マニュアルCと比較すると、低いと思われ、コントロールされているようである。また、先の比率は、6：2：1のままであった。

いずれにせよ、このデータは、電話交換機および計算機マニュアル用の制限英語開発に貴重な資料となる。マニュアルA、Bは、既に列記としたものとして出まわっており、このデータは、それを分析して得られた、制限英語らしきものであるから。

4．2．4　品詞別語い表および技術用語表

付録1に、3マニュアルに対する、動詞、名詞、形容詞、副詞の品詞別語い表を示す。また、付録2に、熟語形式の技術用語を示す。いずれも頻度順にソートし直した、上から十番目までの表である。

マニュアルA，B，Cの内容の違いが、名詞、技術用語に明確に現れている。動詞を見ると、マニュアルB，Cは、共通な単語が上位に多く昇っているが、マニュアルAは、異なっている。形容詞および副詞は、共通な単語が上位を占めているが、マニュアルBには、技術的表現も多く用いられている。使用頻度を見ると、マニュアルBは、特定の動詞、名詞が高く、内容が他と比べて専門的なことを伺わせている。

5．　おわりに

VOXシステム（VOcabulary eXtracting system）は、制限英語開発用のツールとして開発された。磁気化されたマニュアルを入力すると、自動的に品詞ごとに、語い表を作成し、同時に、技術用語表も作成する。各語句には使用頻度が伴われる。修正は、専用エディタを通じて、容易にかつ正確に行うことができる。また、品詞別語い曲線を作成し、語いの飽和状態を観察することにより、調査範囲が妥当であったかどうかを判断できる。このデータを基に、同義語の整理、最適語の選択、リライトを行なうことにより最終的に、制限英語用の語いが求まる。この部分のサポートは、今後の課題である。

VOXシステムをいくつかのマニュアルに適用した結果、いくつかの興味深い事実が見つかった。

（1）名詞と動詞、形容詞と副詞は、それぞれほぼ同じ語い数を示し、各品詞の占める割合もほぼ一定である。（熟語にのみ現れる単語は除く。）

（2）マニュアルA、Bは何らかの形で制限されている。

マニュアルA、Bの分析結果は、それぞれ電話交換機用および計算機用の制限英語を開発する上で、大いに参考になると考えられる。

謝辞

VOXシステムの仕様に関して、第二交換事業部ドキュメント部佐藤課長、五島氏に多くの意見を頂いた。C&Cシステム研究所村木主任からは、様々なアドバイスをして頂いた。これらの方々と、研究の機会を与えて下さった、ソフトウェア生産技術研究所藤野所長、祢津所長代理に対し、感謝いたします。

参考文献

[1] J. Kirkman, C. Snow, I. Watson: "Controlled English as an Alternative to Multiple Translations", IEEE Transaction on Professional Communication, VOL. PC-21, NO. 4 (1978).

[2] B. W. von Glasenapp: "Caterpilar Fundamental English", Proceedings of the 19th International Technical Communication Conference, Society for Technical Communication, U.S.A. (1972).

[3] 吉田：日本語の規格化に関する基礎的研究、昭和58年度科学研究費補助金一般研究（B）、研究成果報告書、九州大学工学部（1984）.

[4] L. L. Cherry, W. Vesterman: "Writing Tools - The STYLE and DICTION Programs", UNIX Programmer's Manual, Volume 2c - Supplementary Documents Seventh Edition, Virtual VAX-11 Version(1980).

付録1　品詞別語い表

マニュアルA	マニュアルB	マニュアルC
動詞		
124 connect	392 set	116 use
107 send	254 use	95 print
89 have	177 specify	59 give
76 use	174 write	58 set
69 set	156 process	57 follow
64 consist	111 issue	49 specify
59 see	75 open	48 match
55 select	60 create	45 cause
54 take	59 contain	44 read
48 detect	58 read	38 write
名詞		
151 subscriber	394 data	202 file
93 digit	175 record	130 line
92 block	120 set	100 command
91 exchange	113 figure	70 option
91 call	95 volume	61 name
64 example	77 system	61 character
58 way	77 buffer	60 address
58 program	76 address	55 error
57 device	71 block	46 number
52 software	63 byte	45 output
形容詞		
170 this	92 each	62 this
63 each	76 this	55 all
56 same	59 one	54 no
53 only	58 first	43 each
50 other	56 any	42 first
44 these	52 same	38 any
43 different	50 no	28 these
42 such	44 all	28 only
42 all	38 next	27 two
34 two	31 more	27 same
副詞		
111 not	143 not	80 not
81 also	36 also	69 then
54 then	33 only	32 only
46 thus	32 automatical	30 with
43 now	29 then	27 also
35 up	22 to	26 by
35 so	20 thus	20 so
34 only	19 however	17 normally
30 as	12 according	15 to
23 out	12 abnormally	15 n't

付録2　技術用語表（熟語形式）

マニュアルA	マニュアルB	マニュアルC
48 control system	58 macro instruction	31 regular expression
32 central software	58 data sets	26 standard output
24 group selector	48 dd statement	25 error message
23 ringing signal	43 operating system	22 addressed line
23 junctor circuit	39 buffer pool	17 standard input
22 subscriber stage	32 block prefix	14 current line
22 data area	29 logical record	13 named file
22 call set-up	28 dcb macro instruction	12 file name
20 data word	24 parameter list	10 last line
19 data store	21 synad routine	8 input file