JICST機械翻訳システム(1)

機械翻訳辞書の作成

芦崎達雄*

JICST machine translation system(1)

Production of machine translation dictionary

ASHIZAKI Tatsuo

〔著者抄録〕 JICSTにおける機械翻訳システムの開発は、1986年から開始した。この機械翻訳システムは1990年から英文データベース作成のために運用されている。機械翻訳システムは、複数の翻訳用の辞書より構成されている。翻訳辞書作成のための用語辞書として、最初に対訳辞書を作成した。JICST 科学技術文献ファイルから単語抽出を行い、対訳辞書を作成する。この対訳辞書を基に名詞の翻訳辞書を作成した。動詞の翻訳辞書は、やはり例文から動詞の用法がある語と名詞を基に作成した。なお英文データベース作成中に分かった未登録語を追加作成している。翻訳実行時辞書としては、オリジナル辞書、属性辞書、ルール辞書等を作成して運用を行っている。英文データベース作成工程の最終結果は、バイリンガルのコーパスとなっていて、イグザンプルベースの機械翻訳システム開発に当たって貴重なソースとなるであろう。この英語抄録はJICST機械翻訳システムを使用して作成した。

〔著者付与キーワード〕 機械翻訳, 翻訳辞書, JICST, 英文データベース, 名詞辞書, 動詞辞書, ターミノロジー

〔Author Abstract〕 The development of machine translation system in JICST started from 1986. This machine translation system is operated for English database preparation since 1990. The machine translation system has been composed of a dictionary for multiple translation. First as a terminology dictionary for the translation dictionary preparation, a correspondence translation dictionary was made. The word extraction is carried out from JICST scientific and technical literature file, and a correspondence dictionary is made. A translation dictionary of a noun was made on the basis of this correspondence dictionary. A translation dictionary of a verb was made on the basis of also word and noun from example sentence with the application of a verb. Still, proven unregistered word is added and is made under English database preparation. An operation is carried out by the preparation of original dictionaries, attribution dictionaries, rule dictionaries, etc.. Final result of English sentence database preparation process will become a source which is precious in machine translation system development of example-base in becoming bilingual corpus. This English abstract was made using JICST machine translation system.

〔Keywords by Author〕 machine translation, translation dictionary, JICST, English database, noun dictionary, verb dictionary, terminology

* 日本科学技術情報センター技術開発部 (〒100 千代田区永田町 2-5-2) Tel. 03(3581)6411
JICST Technology Research and Development Department
(5-2, Nagatacho 2-chome, Chiyoda-ku, 100)

1. JICST の機械翻訳システム

JICST の機械翻訳システムは1982～86年の Mu システムを引き継いで、1986～91年の5年間開発を行った。1990年夏より英文データベース (JICST-E) 作成に利用している (図1)。

英文データベースとは、日本で発生した科学技術論文約24万件に対して英語のタイトル、書誌的事項、キーワード、抄録を提供するデータベースである。機械翻訳を利用して作成する件数は、抄録1.5万件とタイトル7万件である。その他オリジナル論文の英語タイトル、抄録等を利用している。

現在、JICST-E ファイルは、約180万件蓄積されており、JOIS, STN Int を通じて世界中に提供されている (表1, 2, 3)。

機械翻訳システムの規模としては、以下の通りである。

①翻訳辞書

日本語側：名詞35万+18万語

動詞1.5万語

形容詞0.7万語

英語側：名詞41万語

動詞0.5万語

形容詞0.7万語

②文法規則

解析1,500, 変換500, 生成500規則。

機械翻訳システムは、今のところ、JICST の英文 DB 作成のためにしか使用していない。翻訳辞書や文法規則の更新は、約3か月に1回行っている。

2. 科学技術用語辞書

2.1 科学技術用語の種類

科学技術専門用語集は、『辞典事典総合目録』出版ニュース社発行 (1990) によれば数百種類にのぼる。

JICST で、JICST 科学技術文献ファイル作成の参考資料として、使用している用語集のうち、日英専門用語集は約200種にもなる。

2.2 科学技術用語の量

JICST 科学技術分類を基に、各分野の代表的な用語集40種について調査した結果、延べ語数は約100万語に達した (表4)。

また、JICST 科学技術文献ファイルの文献タイトルの切り出し語について調査した結果、5年間で200万抄録からの異なり語は、約300万語にも達した。

この中で頻度が3以上の語は約30万語になった。

2.3 対訳辞書の作成

JICST 科学技術文献ファイルから使用言語が英語の文献を抽出し、その日本語のタイトルに対して、対訳辞書作成管理システム (TERM-JET) を用いて単語分割を行う。そこで分割された日本語見出し語を検索キーとして、既作成の対訳辞書マスター (JICST シソーラス用語, JIS 用語, 学

図1 JICST 日英機械翻訳システム

表1 JICST-E ファイルにおける機械翻訳の割合

	1990	1991	1992
標題	147,123 (63%)	156,511 (67%)	147,065 (60%)
{ 原文中	82,649 (35%)	70,846 (30%)	73,448 (30%)
{ 人手	48,982 (21%)	—	—
{ 機械翻訳	15,492 (7%)	85,665 (37%)	73,617 (30%)
標題+抄録	87,035 (37%)	76,104 (33%)	95,790 (40%)
{ 原文中	77,607 (33%)	59,678 (26%)	82,459 (34%)
{ 人手	506 (0%)	—	—
{ 機械翻訳	8,922 (4%)	16,426 (7%)	13,331 (6%)
機械翻訳合計	24,414 (10%)	102,091 (44%)	86,948 (38%)
総合計	234,158	232,615	242,855

術用語等)と照合する。照合した結果、一致しなかった単語については、新規に対訳辞書を作成する候補語として、その語を含む日本語のタイトルと英語の原文タイトルを出力する。

次にそのリストから、日本語見出し語に相当する英語見出し語を選び、さらに出典として JICST 科学技術分類表に従った分野コードも付与する。このような方法で新規に作成した語と既作成の対訳辞書マスターをマージして、最終の対訳辞書マスターを作成した。約32万語の対訳辞書が完成した(図2)。

対訳辞書の持つ項目としては、以下の通りである。

①見出し語

JICST 科学技術文献ファイルのタイトルより抽出した語と、JICST で辞書類より収集した語。

②フリガナ

カタカナで表す。英数文字は、そのまま表記

表2 JICST-E ファイルの作成・蓄積件数

	作成件数	蓄積件数
84	18,454	
85	173,862	192,316
86	200,101	392,417
87	235,867	628,284
88	260,262	888,546
89	244,068	1,132,614
90	234,158	1,366,772
91	232,615	1,599,387
92	242,855	1,842,242

表3 翻訳辞書の語数

94.1

品詞	日本語解析辞書	日英変換辞書	英語生成辞書
名詞(科学技術)	353,796	353,940	294,240
名詞(医学)	182,765	182,765	119,306
固有名詞	(24,896)	(24,896)	(806)
代名詞	30	30	47
動詞	15,256	14,398	4,971
形容詞	6,794	6,794	7,088
形容動詞	451	451	2,123
限定詞	156	156	48
接続詞	87	87	34
後置詞	102	87	—
補助動詞	115	115	9
接辞	279	252	—
前置詞	—	—	158
冠詞	—	—	3
基数	—	—	30
序数	—	—	30
単位	—	—	5
合計	558,348	557,594	428,097

し、合成、上つき、下つき文字は、標準の文字にて表記する。同一表記で読みが異なる語、(例:「根」(コン、ネ)に対しては、複数付与する。

③対訳(複数)

表4 科学技術分野の代表的用語集

1993.09

名 称	和→英	英→和	その他	語 数	ふりがな	出 版 社	発行年
インタープレス社 科学技術35万語大辞典<和英>	○			350,000	かな	I P C	90
インタープレス社 科学技術35万語大辞典<英和>		○		350,000	かな	I P C	90
インタープレス社 新ビジネス18万語大辞典<和英>	○			180,000	かな	I P C	90
インタープレス社 新ビジネス18万語大辞典<英和>		○		180,000	かな	I P C	90
マグローヒル 科学技術用語大辞典	○	○		100,000		日刊工業新聞社	85
科学技術英大辞典	○	○		25,000		オーム社	87
J I S 工業用語大辞典	○	○		60,000	かな	日本規格協会	87
岩波理化学辞典	○	○	○	14,000		岩波書店	87
化学大辞典	○	○		70,000	かな	共立出版	63
化学大辞典	○	○		25,000		東京化学同人	89
分析化学辞典	○	○		12,000	かな	共立出版	71
岩波生物学辞典	○	○	○	13,700		岩波書店	83
生化学辞典	○	○		12,500		東京化学同人	90
分子生物学辞典	○	○		15,000		化学同人	92
畜産用語辞典	○	○		3,000	かな	養賢堂	77
農業土木標準用語辞典	○	○		2,500		農業土木学会	74
食品工業総合辞典	○	○		8,000	かな	光琳	79
総合水産辞典	○	○		15,000	かな・ローマ	成山堂	85
医学大辞典	○	○	○	27,000		南山堂	89
医学大辞典	○	○	○	45,000		医歯薬出版	87
和英医学用語大辞典	○			320,000		日外アソシエーツ	90
英和医学用語大辞典		○		230,000		日外アソシエーツ	89
薬科学大辞典	○	○	○	21,000	かな	広川書店	83
数学英和・和英辞典	○	○		6,000		共立出版	79
経営英和辞典		○		19,000		ダイヤモンド社	80
情報科学辞典	○	○		13,000		岩波書店	90
エネルギー用語辞典	○	○		2,600		日刊工業新聞社	76
原子力科学用語辞典	○	○		5,000		日刊工業新聞社	83
電気術語大辞典	○	○	○	25,000		オーム社	64
電気用語辞典	○	○		14,000	かな	コロナ社	82
建築大辞典	○	○		34,500	かな	彰国社	93
土木用語辞典	○	○		17,000	ローマ	技国堂出版	71
環境科学辞典	○	○		6,000		東京化学同人	85
金属工学辞典	○	○		5,000		地人書館	62
機械工学辞典	○	○		11,000		朝倉書店	88
地学辞典	○	○	○	17,000		平凡社	70

見出し語に対応する英語で、最大9個まで収録できる。同義語、異形語、日本語見出し語が同形で異なる意味の語もすべて収録する。

④情報源分野コード（複数）

JICST 科学技術文献ファイル1981～87年版のタイトルから抽出した語は、その見出し語が含まれていた文献の JICST 統一分類コードを付与した。したがって、この分野コードは、見出し語が出現した分野を示すもので、必ずしも見出し語の属する分野を示すものではない。

対訳語が複数の場合は、それぞれの語に対して分野コードを付与する。

⑤頻度

JICST 科学技術文献ファイル1981～87年版のタイトル中の出現頻度を調べた。

3. 翻訳辞書の作成

3.1 名詞翻訳辞書

対訳辞書を基に、翻訳辞書に必要な見出し語・対訳を選定した。対訳辞書には、複数の訳語を持っている語があるが、科学技術用語として一番多く使用されられると思われる語を代表訳語として同義語をまとめた。

次に使用分野や意味的に違う語を分類した。さらに翻訳辞書として必要な品詞細分類、名詞意味マーカ等の付与を行った（図3）。

翻訳辞書の構造は、形態素情報、構文意味情報、変換情報等の階層構造をとっており、データ形式、データ長、必須項目、検索項目、属性値等を規定している。

〔作成方法〕

- ① 科学技術振興調整費で作成した翻訳辞書から、約7万語を再編集した。
 - ② ①にない語を、JICST 対訳辞書から約25万語作成した。その他 JICST 科学技術文献ファイルの中で、オリジナル論文が日本語で記述されている文献のタイトルと抄録について、形態素解析を実行して未登録語の追加を行った。
 - ③ JICST シソーラスの日英対訳リストを取り込んだ。
 - ④ 機関名、地名等の固有名詞の追加を行った。
 - ⑤ 他機関で作成した医学用語についてフォーマット変換を行い、医学専門用語辞書を作成した。
- このような方法によって、対訳辞書を基にして名詞翻訳辞書を作成した。また翻訳例文中の未知語については、常時、翻訳辞書を作成している。1994年1月末で、科学技術辞書は約35万語で、医

TERMJET

和英タイトル リスト

(87.01.09) PAGE.000018

AA

B00-100017 0986295

科学技術教育

カガクギジュツキョウイク

#01 科学技術教育の衰退 だれが本当に心配しているのか

Decay in science and engineering education. Who really cares?

AA

#02 非科学技術者に対する科学技術教育 次の世代への新しい挑戦

Education of technology to non-technologists - A challenge for the next decade.

AA

図2 対訳辞書作成候補語リスト

%=====	
%	日本語品詞表
%=====	
\$\$JPOS	
NOUN:NOUN:	名詞
PRON:PRON:	代名詞
VERB:VERB:	動詞
ADJA:ADJA:	形容詞
ADJV:ADJV:	形容動詞
ADVERB:ADVERB:	副詞
AUXI:AUXI:	助動詞
POSTP:POSTP:	助詞
AFFIX:AFFIX:	接辞・接語
DETERM:DETERM:	連体詞
CONJUNC:CONJUNC:	接続詞
\$\$	
%=====	
%	日本語品詞細分類表
%=====	
\$\$JSPS	
KOYUU:KOYUU:	固有名詞
FUTUU:FUTUU:	普通名詞
SAHEN:SAHEN:	サ変名詞
DOUSA:DOUSA:	動作名詞
KAKUTEKI:KAKUTEKI:	格助詞の名詞
SETUTEKI:SETUTEKI:	接続助詞の名詞
HOBUN:HOBUN:	補文標識
FUKUTEKI:FUKUTEKI:	副詞の名詞
KEIDO_N:KEIDO_N:	形動語幹
KA_SAHEN:KA_SAHEN:	マル化サ変
GIMOND:GIMOND:	疑問代名詞
NINSYOD:NINSYOD:	人称代名詞
SIJID:SIJID:	指示代名詞
WAGOV:WAGOV:	和語
SAHENV:SAHENV:	サ変
KASAHENV:KASAHENV:	化サ変
TOKUSYUV:TOKUSYUV:	特殊
TEKI_KEI:TEKI_KEI:	～的
SIKI_KEI:SIKI_KEI:	～式
KATA_KEI:KATA_KEI:	～型
HOKA_KEI:HOKA_KEI:	その他の形容
KEIYO:KEIYO:	形容詞
SUU:SUU:	数
SUURYOU:SUURYOU:	量詞
KANSUU:KANSUU:	数詞
JOSUUSI:JOSUUSI:	助数詞
SETTO_GO:SETTO_GO:	接頭語
SETTO_JI:SETTO_JI:	接頭辞
SETUB_GO:SETUB_GO:	接尾語
SETUB_JI:SETUB_JI:	接尾辞

図3 属性値テーブル

学辞書は約18万語で、名詞翻訳辞書語数は、合計約53万語である(図4, 5)。

名詞翻訳辞書の持つ項目としては、以下の通りである。

①日本語見出し語

対訳辞書の見出し語に相当する。

②読み

対訳辞書のフリガナに相当する。

③日本語品詞細分類

普通名詞, 固有名詞, サ変名詞, 化サ変名詞, 形容動詞語幹, 動作名詞, 副詞的名詞, 格助詞的名詞, 接続助詞的名詞, 補文標識の10種類に分類する。

④名詞意味マーカ

名詞の意味素性を分類するために、13の大分類と49の小分類からなる意味の体系表を作成した。

大分類としては、生物, 無生物, 知的抽象物, 属性等である。

小分類としては、生物の下で人, 動物, 植物およびその他に分ける。

⑤英語見出し語

科学技術用語として、よく使われると思われる訳語を代表訳とする。ただし意味的に違う場合や使用分野によって訳語が異なる場合は、分類分けを行う。

⑥分野コード

対訳辞書の情報源分野コードに対応する。

3.2 動詞翻訳辞書

動詞翻訳辞書の作成においても、名詞翻訳辞書と同様に解析辞書と変換辞書の部分を組み合わせて、日英解析変換辞書として作成した。

辞書の構成としては、見出し語情報, 形態素情報, 構文・意味情報, 格支配情報および変換情報の部分から構成されている。

動詞辞書の作成候補語としては、JICST 科学技術文献ファイルの1か月分の日本語文献約3万文から、動詞の用法がある語について抽出し、さらに和英辞典, サ変名詞, カタカナ語等より見出し

%=====									
% 日英解析変換(名詞)辞書									
%=====									
\$\$\$									
01	[見出しNo]	] LEXN-1000 C	1- (1 1 1)	A/12	NUM	-	-	(14)	(GOBANGO)
01	[更新年月日]	] UPDA-1100 C	1- (0 1 0)	A/126	NUM	-	-	(24)	(UP_DATE)
01	[作成者]	] NAME-1200 C	1- (0 1 0)	K/126	-	-	-	(22)	(USER_NAME)
01	[見出し情報]	] JLIF-2000 C G	1- ()	/	-	-	-	(64)	(J_LEX_INF)
02	[日本語見出し語]	] JLEX-2100 C	1- (1 1 0)	K/126	-	-	-	(12)	(J_LEX)
02	[語基]	] LEXB-2400 C	1- (0 0 0)	A/126	NUM	-	-	(24)	(J_BASE)
02	[読み]	] LEXL-2500 C	1- (1 1 0)	K/126	-	-	-	(12)	(J_YOMI)
02	[異形語]	] LEXD-2700 C	M- (0 1 0)	K/126	-	-	-	(22)	(J_IKEI)
01	[形態素情報]	] JHIF-4000 A G	1- ()	/	-	-	-	(64)	(J_MORPH_INF)
02	[前接情報]	] PREN-4310 A	1- (1 0 1)	A/2	NUM	PREN	PREN	(14)	(J_FK_INF)
02	[後接情報]	] PSTN-4320 A	1- (1 0 1)	A/2	NUM	PSTN	PSTN	(14)	(J_BC_INF)
01	[構文意味情報]	] JSIF-6000 C G	1- ()	/	-	-	-	(64)	(J_MSENT_INF)
02	[日本語構文品詞]	] JPOS-6110 C	1- (1 1 1)	A/8	-	JPOS	XPOS	(11)	(J_CAT)
02	[日本語品詞細分類]	] JSPS-6120 C	1- (1 1 1)	A/8	-	JSPS	JSPS	(11)	(J_SUBCAT)
02	[意味マーカ]	] JSEM-6200 A	M- (0 1 1)	A/2	-	SEMK	SEMK	(21)	(J_SEM)
02	[変換情報]	] TINF-B000 T G	M- ()	/	-	-	-	(64)	(E_TRAN_INF)
03	[J-U-I-D]	] JUID-B150 T	1- (1 0 1)	A/2	NUM	-	JUID	(14)	(USAGE)
03	[英語見出し語]	] ELEX-B200 T	1- (1 1 0)	K/126	-	-	-	(12)	(E_LEX)
03	[英語構文品詞]	] EPOS-6130 T	1- (0 1 1)	A/8	-	EPOS	EPOS	(11)	(E_CAT)
03	[英語品詞細分類]	] ESPS-6140 T	1- (0 1 1)	A/8	-	ESPS	EPSS	(11)	(E_SUBCAT)
03	[分野コード]	] FIELD-B300 T	1- (0 0 1)	A/4	-	FIELD	FIELD	(21)	(FIELD_CODE)
03	[意味マーカ]	] SSEM-B400 T	1- (0 0 0)	A/2	-	SEMK	-	(21)	(J_SEM)
03	[意味細分類]	] KSSM-B500 T	1- (0 0 0)	A/2	-	-	-	(11)	(J_SEM_CAT)
03	[E-U-I-D]	] EUID-B120 T	1- (0 0 0)	A/2	NUM	-	-	(14)	(E_UID)
03	[他の英語見出し語]	] ETTF-5450 T	M- (0 0 0)	K/126	-	-	-	(22)	(OTHER_CHICE)
03	[後接修飾情報]	] PSTI-B700 T G	M- ()	/	-	-	-	(32)	(MODIFYING)
04	[表層形]	] PFAC-B710 T	1- (0 1 0)	K/126	-	-	-	(12)	(J_KAKU_LEX)
04	[深層格]	] PDEF-B720 T	1- (0 0 0)	K/20	-	JDEP	-	(11)	(J_DEEP_CASE)
04	[表層付属語]	] PPRE-B730 T	1- (0 0 0)	K/6	-	-	-	(12)	(E_PREP_LEX)
04	[英語見出し語]	] PELX-B740 T	1- (0 1 0)	K/6	-	-	-	(12)	(E_LEX)
04	[英語構文品詞]	] PPOS-B760 T	1- (0 1 0)	A/8	-	EPOS	-	(11)	(E_CAT)
04	[E-U-I-D]	] PUID-B750 T	1- (0 0 0)	A/2	NUM	-	-	(14)	(E_UID)
03	[前接修飾情報]	] PREI-B800 T G	M- ()	/	-	-	-	(32)	(MODIFIED)
04	[表層形]	] SFAC-B810 T	1- (0 1 0)	K/126	-	-	-	(12)	(J_KAKU_LEX)
04	[深層格]	] SDEF-B820 T	1- (0 0 0)	K/20	-	JDEP	-	(11)	(J_DEEP_CASE)
04	[表層付属語]	] SPRE-B830 T	1- (0 0 0)	K/6	-	-	-	(12)	(E_PREP_LEX)
04	[英語見出し語]	] SELX-B840 T	1- (0 1 0)	K/6	-	-	-	(12)	(E_LEX)
04	[英語構文品詞]	] SPOS-B860 T	1- (0 1 0)	A/8	-	EPOS	-	(11)	(E_CAT)
04	[E-U-I-D]	] SUID-B850 T	1- (0 0 0)	A/2	NUM	-	-	(14)	(E_UID)
03	[共起修飾語]	] CMDF-B900 T G	M- ()	/	-	-	-	(32)	(MODIFIER)
04	[日本語見出し語]	] CJLX-B910 T	1- (0 0 0)	K/126	-	-	-	(12)	(J_LEX)
04	[英語見出し語]	] CELX-B920 T	1- (0 0 0)	K/126	-	-	-	(12)	(E_LEX)
04	[E-U-I-D]	] CUID-B930 T	1- (0 0 0)	A/2	NUM	-	-	(14)	(E_UID)
04	[品詞]	] CPOS-B940 T	1- (0 0 0)	A/8	-	EPOS	-	(11)	(E_CAT)
03	[動詞共起情報]	] VCIF-BA00 T G	M- ()	/	-	-	-	(32)	(J_N_KYOUKI)
04	[用例]	] CESM-7710 T	1- (0 0 0)	K/126	-	-	-	(12)	(J_YOUREI)
04	[日本語見出し語]	] VCJL-BA20 T	1- (0 0 0)	K/126	-	-	-	(12)	(J_LEX)
04	[英語見出し語]	] VCFL-BA30 T	1- (0 1 0)	A/126	-	-	-	(12)	(E_LEX)
04	[E-U-I-D]	] VCID-BA40 T	1- (0 0 0)	A/2	NUM	-	-	(14)	(E_UID)
04	[英語構文品詞]	] VCEP-BA50 T	1- (0 0 0)	A/8	-	-	-	(11)	(E_CAT)
04	[格フレーム]	] CSFR-BA60 T G	1- ()	/	-	-	-	(32)	(CASE_FRAME)
05	[日本語表層格]	] JFAC-BA61 T	M- (0 0 0)	K/126	-	-	-	(12)	(J_SURFACE_CASE)
05	[日本語深層格]	] JDEP-BA62 T	1- (0 0 0)	K/20	-	JDEP	-	(11)	(J_DEEP_CASE)
05	[日本語位置]	] JPSI-BA63 T	1- (0 0 0)	A/3	-	JPFL	-	(11)	(J_POS_FLAG)
05	[英語表層格]	] EFAC-BA64 T	1- (0 0 0)	K/126	-	EFAC	-	(11)	(E_SURFACE_CASE)
05	[英語深層格]	] EDEP-BA65 T	1- (0 0 0)	A/126	-	EDEP	-	(11)	(E_DEEP_CASE)
05	[変換対応]	] ECRP-BA66 T	1- (0 0 0)	K/126	-	JDEP	-	(11)	(CORRESPONDENCE)
01	[備考]	] NOTE-FFFF C	M- (0 0 0)	K/126	-	-	-	(12)	(NOTE)
\$\$\$									

図4 名詞翻訳辞書構造

語の追加を行った。

なお、現在は、英文データベースの運用による未登録語の動詞を追加して、1994年1月末で約1.5万語を作成した。

辞書の構造としては、基本的には名詞辞書と同様な階層構造をとっているが、辞書属性の項目が多く、階層構造が深くなっている。形態素情報における動詞の活用情報、構文・意味情報におけるその動詞の意味的な情報や文型変化、格支配情報における日本語表層格、日本語深層格や日本語意

味マーカ、さらに変換情報における Condition の配下の日本語表層格、日本語深層格や日本語意味マーカおよび英語表層格や英語深層格等である(図6)。

動詞辞書の持つ項目としては、以下の通りである。

(1) 見出し語情報

①日本語見出し語

終止形で表す。

②語尾字数

終止形における語尾の字数とする。形態素解析において、活用形を推定するために語幹と語尾を分ける。五段活用は1で、一段活用とサ変活用はおおむね2である。

③読み

複数の読みがある場合は一つだけとする。意味的に違う場合には、格パターンで区別する。

(例) 開く：ひらく、あく

④異形語

意味と読みが同じで、表記の異なる語を記述する。送り仮名の異なる語も記述する。

⑤原形語

見出し語が可能動詞の場合に、派生するものと語。

(例) 歩ける：歩く(原形語)

(2) 形態素情報

①動詞活用型

五段、上一段、下一段、サ変(一般)、サ変(〇〇化)、ザ変、特殊の7種類に分類する。

②動詞活用行

ア行、カ行、ガ行、サ行、ザ行、タ行、ナ行、ハ行、バ行、マ行、ヤ行、ラ行、ワ行に分類する。

語尾字数、動詞活用型、動詞活用行より接続関係のテーブルを発生させる。

③後接情報

後接語との接続関係を示すもので、活用語については、語幹の情報を記入する。特殊活用の場合には、活用形ごとに記入する。

0001	[見出しNo]	]	M01	LEXN	0279744
0002	[更新年月日]	]	M01	UPDA	920520
0003	[作成者]	]	M01	NAME	芦崎
0004	[見出し情報]	]	M01	JLIF	
0005	[日本語見出し語]	]	M02	JLEX	頭
0006	[読み]	]	M02	LEXL	アタマ
0007	[形態素情報]	]	M01	JMIF	
0008	[前接情報]	]	M02	PREN	1
0009	[後接情報]	]	M02	PSTN	1
0010	[構文意味情報]	]	M01	JSIF	
0011	[日本語構文品詞]	]	M02	JPOS	NOUN
0012	[日本語品詞細分類]	]	M02	JSPS	FUTUU
0013	[意味マーカ]	]	M02	JSEM	EP,EL
0014	[変換情報]	]	M02	TINF	
0015	[J-U-I-D]	]	M03	JUID	1
0016	[英語見出し語]	]	M03	ELEX	head
0017	[英語構文品詞]	]	M03	EPOS	N
0018	[英語品詞細分類]	]	M03	ESPS	COM
0019	[分野コード]	]	M03	FILD	AA00
0020	[意味マーカ]	]	M03	SSEM	EP,EL
0021	[E-U-I-D]	]	M03	EUID	1
0022	[他の英語見出し語]	]	M03	ETTF	caput
0023	[動詞共起情報]	]	M03	VCIF	
0024	[用例]	]	M04	CESM	人が難間で頭を抱え込む
0025	[日本語見出し語]	]	M04	VCJL	抱え込む
0026	[英語見出し語]	]	M04	VCEL	be
0027	[E-U-I-D]	]	M04	VCID	2
0028	[英語構文品詞]	]	M04	VCEP	V
0029	[格フレーム]	]	M04	CSFR	
0030	[日本語表層格]	]	M05	JFAC	が
0031	[日本語深層格]	]	M05	JDEP	主体
0032	[英語表層格]	]	M05	EFAC	SUBJ
0033	[英語深層格]	]	M05	EDEP	AGT
0034	[変換対応]	]	M05	ECRP	主体
0035	[格フレーム]	]	M04	CSFR	
0036	[英語表層格]	]	M05	EFAC	ATRSUB
0037	[英語深層格]	]	M05	EDEP	MAN
0038	[変換対応]	]	M05	ECRP	\$ puzzled
0039	[格フレーム]	]	M04	CSFR	
0040	[日本語表層格]	]	M05	JFAC	を
0041	[日本語深層格]	]	M05	JDEP	対象
0042	[共起名詞]	]	M05	JMEI	頭
0043	[日本語位置]	]	M05	JPSI	DEL
0044	[動詞共起情報]	]	M03	VCIF	
0045	[用例]	]	M04	CESM	メンテナンスを頭に置く
0046	[日本語見出し語]	]	M04	VCJL	置く
0047	[英語見出し語]	]	M04	VCEL	take
0048	[E-U-I-D]	]	M04	VCID	4
0049	[英語構文品詞]	]	M04	VCEP	V
0050	[格フレーム]	]	M04	CSFR	
0051	[日本語表層格]	]	M05	JFAC	が
0052	[日本語深層格]	]	M05	JDEP	主体
0053	[英語表層格]	]	M05	EFAC	SUBJ

図5 名詞翻訳辞書例「頭」

0070:0002845:00 (VERB SAHENV) 90-07-02	
0001 [見出し語No]	] D01 LEXN 0002845
0002 [見出し情報]	] D01 JLIF
0003 [日本語見出し語]	] D02 JLEX 翻訳する
0004 [語尾字数]	] D02 GOBL 2
0005 [語基]	] D02 LEXB 1-2,3-4
0006 [読み]	] D02 LEXL ホンヤクスル
0007 [形態素情報]	] D01 JMIF
0008 [動詞活用型]	] D02 VTYP サ変
0009 [動詞活用行]	] D02 VLIN サ行
0010 [前接情報]	] D02 PREM 2
0011 [後接情報]	] D02 PSTN 13
0012 [活用情報]	] D02 INFX 12
0013 [構文意味情報]	] D01 JSIF
0014 [日本語構文品詞]	] D02 JPOS VERB
0015 [日本語品詞細分類]	] D02 JSPS SAHENV
0016 [格パターン]	] D02 CASE V1
0017 [自動詞/他動詞区分]	] D03 VITF T
0018 [アスペクト]	] D03 DASP 結果
0019 [態]	] D03 DVOC 受身, 使役, 可能, てある
0020 [意志]	] D03 JWIL 有
0021 [文型変化]	] D03 SYNT
0022 [受身]	] D04 PASS 1
0023 [使役]	] D04 CAUS 2
0024 [可能]	] D04 CAPA 1
0025 [格支配情報]	] D03 JCPR
0026 [用例]	] D04 CESH 人
0027 [日本語表層格]	] D04 CJFC が
0028 [日本語深層格]	] D04 CJDP 主体
0029 [意味マーカ]	] D04 CCSM ON,OF
0030 [構文型]	] D04 CSYF N
0031 [必須格/任意格]	] D04 JOBL 1
0032 [格支配情報]	] D03 JCPR
0033 [用例]	] D04 CESH 英語
0034 [日本語表層格]	] D04 CJFC を
0035 [日本語深層格]	] D04 CJDP 対象
0036 [意味マーカ]	] D04 CCSM XX
0037 [構文型]	] D04 CSYF N
0038 [必須格/任意格]	] D04 JOBL 1
0039 [格支配情報]	] D03 JCPR
0040 [用例]	] D04 CESH 日本語
0041 [日本語表層格]	] D04 CJFC に
0042 [日本語深層格]	] D04 CJDP 終状態
0043 [意味マーカ]	] D04 CCSM IO
0044 [構文型]	] D04 CSYF N
0045 [必須格/任意格]	] D04 JOBL 1
0046 [変換情報]	] D03 TINF
0047 [Condition]	] D04 COND 1
0048 [用例]	] D04 EXMS 人がスワヒリ語を日本語に翻訳する
0049 [英語見出し語]	] D04 ELEX translate
0050 [英語構文品詞]	] D04 EPOS V
0051 [EUID]	] D04 EUID 2
0052 [格フレーム]	] D04 CSFR
0053 [日本語表層格]	] D05 JFAC が
0054 [日本語深層格]	] D05 JDEP 主体
0055 [英語表層格]	] D05 EFAC SUBJ
0056 [英語深層格]	] D05 EDEP AGT
0057 [変換対応]	] D05 ECRP 主体
0058 [格フレーム]	] D04 CSFR
0059 [日本語表層格]	] D05 JFAC を
0060 [日本語深層格]	] D05 JDEP 対象
0061 [英語表層格]	] D05 EFAC OBJ
0062 [英語深層格]	] D05 EDEP OBJ
0063 [変換対応]	] D05 ECRP 対象
0064 [格フレーム]	] D04 CSFR
0065 [日本語表層格]	] D05 JFAC に
0066 [日本語深層格]	] D05 JDEP 終状態
0067 [英語表層格]	] D05 EFAC = into
0068 [英語深層格]	] D05 EDEP GOA
0069 [変換対応]	] D05 ECRP 終状態

図6 動詞翻訳辞書例「翻訳する」

④特殊活用形

特殊活用になる語は、平仮名等の異形語も含めて、未然、連用…の各活用形ごとに語形と後接情報を記入する。

(3) 構文・意味情報

①格パターン

日本語の格支配関係を分類し、必須格の多いものから一連の番号を付ける。

②アスペクト

動詞には、その動作や現象の起こり方による分類が考えられる。例えば、「発射する」は、それ自体は一瞬で生起することがらであり、「発射し続ける」とは「発射」を持続させるのではなく、繰り返すことを表している。逆に、「放射する」は、ある時間だけ持続することであり、「放射し続ける」とは「放射」を繰り返すことではない。また、「直交する」のように関係を表し、そのことがらの始まりや終わりを問題にしない動詞もある。

こうした、動詞の表すことがらの起こり方の違いを「アスペクト素性」という。このアスペクト素性により動詞を、瞬時、継続、結果、推移、状態、準状態の6種類に分類する。

英語側で動詞の時制を決定する手がかりとする。

③態

受身、使役、可能、てある、に対してその動詞に生じうるもので、しかも格パターンに変化をもたらす語。

受身では、「れる」、「られる」が付けられる語。使役では、「せる」、「させる」が付けられる語。可能では、一段活用に「られる」、サ変語幹に「できる」が付けられる語。

能動 秘書が郵便物を机の上に置く。

受身 郵便物が(秘書によって)机の上に置かれる。

てある 郵便物が机の上に置いてある。

能動 (円高により) A社の業績が悪化する。

る。
使役 円高がA社の業績を悪化させた。

能動 施設を利用する。
可能 施設が利用できる。
可能 施設を利用できる。

④意志

動詞には、「やろうとしてできる動作」と「そうになっている現象」を表すものという分類が考えられる。「記述する」、「処理する」等は、「そうしよう」としてできるが、「発光する」、「爆発する」等は普通の状況では「そうしよう」と思ってもできることでない。また、両者の間には、「達成する」、「成功する」等の、ある程度意志で左右できても完全には制御できるわけではないことを表す動詞もある。このような意志性を分類する。

⑤意味コード

記述動詞は、論文作成者が論文の筋立てを説明するために用いる動詞のうち、有意性の語。

(例) 解説する、提案する

準記述動詞は、論文作成者が論文の筋立てを説明するために用いる動詞のうち、無意志または中間の語。

(例) 分かる、思う

対称動詞は、「が」、「を」、「に」格の名詞のいずれかが対称的な語を持つ語。

(例) 対応する、比較する、またがる

⑥文型変化パターン

格支配情報が、受身、使役、可能の場合にどのように変化するかを記述する。

態で取りうる動詞の変化の文型を示す。

(4) 格支配情報

①日本語表層格

格助詞および格助詞相当語を記入する。必須格としては自動詞では「が」格、他動詞では「が」、「を」、「に」格を要求し、その他英語で表現するときに特別な前置詞を持つ格および文型変化パターンに含まれる格について

記入する。

②日本語深層格 (表5)

文の表層上の主語、目的語…に対して、主体格、対象格、道具格、場所格、時間格…を記述する。

③名詞・名詞意味マーカ (表6)

上記格に生ずる意味マーカを記入する。

④必須性

動詞が直接支配する格を必須格とする。

(5) 変換情報

①英語見出し語

日本語見出し語に対して、和英辞典を用いてその語について意味・用法をカバーする英語の候補語を抽出する。KWICを参照して意味・用法が科学技術文献に適切な訳語を選び出す。

②Condition

同一格パターン内で意味マーカによって訳し分けができる場合に、一連番号を付与する。

③E-UID

英語見出し語の取りうる英語動詞の格パターンの番号。

④英語表層格

英語には、日本語の助詞に相当する語がないので、主語、第一目的語、第二目的語、主格補語、目的格補語、その他副詞類を記述する。

⑤英語深層格

日本語深層格に準ずる。ただし英語側では、日本語の主語に該当する深層格は、意志性の有無、自動詞と他動詞によって区別する。

3. 3 翻訳辞書の評価および改良

1993年度の翻訳システム運用環境では、名詞53万語、動詞1.5万語の翻訳辞書を使用している。辞書データの評価には、翻訳対象文の入力時の形態素解析の未知語と英語に訳出された時の未登録語を収集して、辞書データに反映している。

名詞翻訳辞書は、主として対訳辞書と他機関で作成した専門用語辞書の導入等により作成した。

表5 日本語深層格

日本語名	英語名	用 例
(1) 主 体	SUBject	～が
(2) 対 象	OBject	～を
(3) 受 け 手	RECIpient	～に与える
(4) 与 え 手	ORigin	～から受ける、奪う
(5) 相 手 1	PARtner	～と協議する、異なる、～に関連する
(6) 相 手 2	OPPonent	～から保護する、独立する
(7) 時	TIME	1980年に
(8) 時・始点	Time-FROM	5月から
(9) 時・終点	Time-TO	来年まで
(10) 時 間	DURation	5分間加熱する
(11) 場 所	SPAcE	～に位置する、～で発生する
(12) 場所・始点	Space-FROM	～から帰る
(13) 場所・終点	Space-TO	～へ送る、～に到達する
(14) 場所・経過	Space-THROUGH	～を通る、上空を飛ぶ
(15) 始 状 態	SOURce	5.5%から6%へ引き上げる
(16) 終 状 態	GOAl	英語から日本語に翻訳する
(17) 属 性	ATTRiute	適応性に富む、欠ける、乏しい
(18) 原因・理由	CAUSE	事故で死ぬ、～から分かる
(19) 手段・道具	TOOl	イオン法で、ドリルで
(20) 材 料	MATERial	ペーストで作る
(21) 構成要素	COMPonent	～から成る、～で構成する
(22) 方 式	MANner	並列に、10m/secで
(23) 条 件	CONDition	焦点深度で決まる
(24) 目 的	PURpose	～に適する、備える、必要な
(25) 役 割	ROLe	議長に選ぶ、～として用いる
(26) 内容規定	CONtent	～と呼ぶ、述べる、みなす
(27) 範囲規定	RANge	～について、～に関して
(28) 提 題	TOPic	～は、～とは
(29) 観 点	VIEWpoint	立場から、～の点で
(30) 比 較	COMpaRison	～より大きい、～に劣る、～を上回る
(31) 陪 伴	ACCOMPany	とともに、～に伴って
(32) 度 合	DEGREE	5%増加する、3キロヤせる
(33) 陳 述	PREdicative	～である
(34) そ の 他	ETC	

注: 英語名中、大文字の部分(3字)を略称とする。

表6 名詞意味マーカ

今後とも未登録語の追加と専門用語辞書の導入等で、辞書の拡充を図っていく予定である。

動詞辞書は、動詞を抽出した翻訳文の結果からの改良と、さらにサ変名詞等の対応から拡充を図っていく。

4. 翻訳実行時辞書

4.1 オリジナル辞書

機械翻訳システムのマスター辞書は、ADABASの翻訳辞書とした。このADABAS辞書には、辞書データが数百メガバイトの大きさに記録されている。翻訳実行システムでは、この

ADABAS 翻訳辞書を直接アクセスすることを避け、翻訳時の各処理に対応した個別の辞書を作成し、使用する。

これは、使用するデータが処理ごとで大きく異なるためであり、これらの処理に対応した辞書をまとめて、翻訳実行時辞書と称する。

ADABAS 辞書と翻訳実行時辞書との間に設定されるのが、オリジナル辞書である。実際の機械翻訳システムにおいては、本体部の翻訳各処理からアクセスを受けるのは翻訳実行時辞書であって、オリジナル辞書ではない。

それゆえ、翻訳実行時辞書側から見ると、オリ

ジナル辞書をマスター辞書と見ることもできるが、機械翻訳システムにおいては、あくまで ADABAS の翻訳辞書をマスター辞書とし、オリジナル辞書は翻訳実行時辞書のデータの緩衝帯であり、基データの集合体と位置付けた。

オリジナル辞書とは、オリジナル解析辞書、オリジナル変換辞書、オリジナル生成辞書の三つをさす。ADABAS 辞書データを基にフォーマットを変換して作成する。オリジナル辞書のデータは、基本的に、ADABAS 辞書と、同質のものである。ADABAS 辞書と同じデータの階層構造を持ち、個々のデータは、可能な限りコード化されている。

実際にオリジナル辞書を作成するには、ADABAS 辞書のデータそのものではなく、アンロード形式に一度直された、アンロードデータと呼ばれるデータから作成する。この時、アンロードデータは、辞書情報が、すべて KEIS 漢字コードで記述されているものを、2 バイトの16進コードに変換する。これをコード化するという。コード化された、オリジナル辞書より、翻訳実行時辞書を作成するので、翻訳実行時辞書自身も、コード化された情報を持っている。

オリジナル辞書は、翻訳実行時辞書の基になるデータである。基になるデータとして、さまざまな要因を含んでいる。主な役割を以下に示す。

- ①翻訳実行時辞書のベース・データである。
 - ②辞書データ中のエラーの削除（オリジナル辞書作成の前にデータのチェックを行いエラーデータを取り除く）
 - ③辞書更新データの蓄積
- オリジナル辞書のデータが2 バイト16進コード

で記述されているので、ADABAS 辞書や、アンロードデータと比較すると、非常にコンパクトなものになっている。コンパクトで扱いやすく、アクセスレスポンスも向上させることができる。

これにより、翻訳実行時辞書に対して、ダイナミックかつ即時的にデータの更新を行うことができる。オリジナル辞書のデータは、作成時にエラーチェックをかけられており、信頼度は高い。

オリジナル辞書のデータは可変長である。辞書データの更新を考える時、可変長のデータの連続では、非常に不便である。そこで、辞書を内部でいくつかのブロックに分けることを考える。

オリジナル辞書では、ファイル内部を同じ大きさに区切っていく。これをブロック化という。1 ブロック内を、辞書データが格納される実データ域と更新時のデータ増加分として空領域がある。1 ブロックごとのデータの個数は一定ではない。また、一つの単語のデータが二つのブロックにまたがることはない。

辞書検索時のレスポンス向上を考えると、データ自身は品詞別、見出し語の KEIS コード系昇順ソートを行い、ブロック化を行うものとする。

4. 2 属性辞書

属性辞書は、解析、変換、生成の各処理において、文法規則を適用するために必要な情報として、単語の意味的な要素を木構造に付与する。属性辞書はオリジナル辞書を基にそれぞれに必要なデータを抽出し、作成する。データの抽出とは、属性辞書として必要な辞書属性（各辞書、各品詞によって異なる）のみを選択することである。

**** データ行の先頭 ****

```
愛する _VERB_HENKAN. SG(#RI(愛する _VERB_HENKAN);) 愛する _VERB_HENKAN. RR
(#VD(NAME=YR);#MC:<#P>;#P.#UD!=%;#SSO:#CDC(T,#P.#LX:#P.#UD:#HNK,<#P>,<YR>);) 愛する _V1_HENKAN. SG(#RI(愛する _V1_HENKAN1);) 愛する _V1_HENKAN1. R
R(#MC:<(#P.#K)>;#K=#DS(<NOM1 NOM2 >,<?>,LR);NOM1.#DP==SUB;NOM2.#DP==OBJ
NOM1.#EDC=EXP;NOM1.#ESC=SUBJ;NOM1.#EOB=YES;NOM2.#EDC=OBJ;NOM2.#ESC=OBJ1;
B=YES;#P.#EL=" l o v e ";#P.#EC=V;#P.#EU=2;) 案ずる _VERB_HENKAN. SG(#RI(
Fマレマ _VERB_HENKAN);) 案ずる _VERB_HENKAN. RR(#VD(NAME=YR);#MC:<#P>;#P.#UD!=
#SSO:#CDC(T,#P.#LX:#P.#UD:#HNK,<#P>,<YR>);#CR:<YR>);) 案ずる _V1_HENKAN. SG(
案ずる _V1_HENKAN1, 案ずる _V1_HENKAN2);) 案ずる _V1_HENKAN1. RR(#MC:<(#P
```

図7 ルール辞書

オリジナル辞書のデータは2バイト16進コードであるが、属性辞書はこれを2バイトの連続した数値(属性番号)に変換する。属性番号は、翻訳実行系で実際に値として使用されるデータ表現と同じである。

4. 3 ルール辞書

ルール辞書とは、オリジナル辞書を基として、文法のルールと同一の形をした翻訳実行部がアクセス可能な形式を持った辞書である。

ルール辞書の内部は、ルール名をキーとした単語ごとの実行形式ルールの集合である。オリジナル辞書や属性辞書は見出し語がキーとなっているが、ルール辞書は、単語固有のルールであり、一定の形式のルールの枠に、各語の辞書情報を埋め込んだものである(図7)。

4. 4 形態素解析辞書

形態素解析辞書は、翻訳実行時辞書の一つで、入力文の形態素解析を実行するための辞書であ

る。形態素解析辞書は、形態素解析処理のために、オリジナル辞書とは異なるデータの階層構造を持っている。またデータのソートも見出し語だけではなく、品詞の優先度も考慮した形態素解析専用の辞書である。

4. 5 形態素生成辞書

形態素生成辞書は、形態素生成に必要な辞書属性をデータとして持つ。形態素生成実行部が、直接アクセスする辞書として位置付ける。辞書としては、属性辞書と並ぶ、実行時辞書の一つである。

形態素生成辞書のデータ項目は、生成オリジナル辞書から必要な項目を抽出して、作成する。オリジナル辞書を入力とするために、個々の単語について記述される内容は、生成オリジナル辞書のデータに依存する。

□□□□□□□□□□□□□□	
東京都において機械翻訳システムを利用	
□□□□□□□□□□□■ ■ ■	
東京都において機械翻	1 0
□□□□□□□□□□	
東京都において機械	9
□□□□□□□□	
東京都において機	8
□□□□□□□	
東京都において	7
□□□□□□	
東京都におい	6
□□□□□	
東京都にお	5
□□□□	
東京都に	4
□□□	
東京都	3
□□	
東京	2

図8-1 切り出し語

■□□□□□□□□□□■ ■	
東京都において機械翻訳システムを利用	
京都において機械翻訳	1 0
京都において機械翻	9
京都	2
■ ■ □□□□□□□□□□ ■	
都において機械翻訳シ	1 0
都に	2
■ ■ ■ ■ ■ ■ □□□□□□□	
システムを利用	7
システム	4
シス	3
■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ □□	
利用	2

図8-2

表7 JICST 科学技術文献ファイル中の語の頻度分布
('90年度, '91年度及び, '91+92年度)

頻 度	' 90年度	' 91年度 ('90年度比)	'91+'92年度 ('90年度比)
10,000 ~	92	169 (1.84)	442 (4.80)
5,000 ~ 9,999	165	272 (1.65)	560 (3.39)
1,000 ~ 4,999	1,346	1,952 (1.45)	3,404 (2.53)
500 ~ 999	1,371	1,956 (1.43)	3,294 (2.40)
200 ~ 499	3,346	4,705 (1.41)	7,893 (2.36)
100 ~ 199	4,425	6,394 (1.44)	9,528 (2.15)
50 ~ 99	6,923	9,628 (1.39)	14,262 (2.06)
20 ~ 49	14,283	20,273 (1.42)	26,482 (1.85)
10 ~ 19	16,148	21,612 (1.34)	25,174 (1.56)
5 ~ 9	20,212	25,996 (1.29)	28,362 (1.40)
1 ~ 4	44,458	50,402 (1.13)	45,240 (1.02)
合 計	112,769	143,359 (1.27)	164,641 (1.46)
辞 書 語 数	345,692	364,660 (1.05)	364,660 (1.05)
抄録件数 (日本語)	213,737	318,258	640,684

4. 6 辞書データ構造

① ADABAS 辞書およびアンロードデータ

JICST で作成する翻訳辞書データで, KEIS 漢字コードにより全データが記述されている。データは, データの種類を表す名前 (属性名) と, データの値 (属性値) との組で表現されており, この組が幾重にも入れ子になった階層構造を持っている。

② オリジナル辞書データ

アンロードデータをフォーマット変換することにより, 作成する。データは可能な限り16進コードに置き換えられ, ADABAS 辞書やアンロードデータと同じ階層構造を持っている。

③ 翻訳実行時辞書データ

オリジナル辞書から必要なものを抽出, もしくは展開して作成する。データは16進コードで表現されるが, 階層構造を持っていない。属性名と属性値の組データが一次的に配置されている。

④ 対応テーブルデータ

データチェックおよびフォーマット変換時に使用するもので KEIS 漢字コードと, 16進コードが対応づけられて記述されている。データチェックという点から, 属性名と, 複数の属性値の組という構造を持っている。

5. 頻度付き辞書の作成

5. 1 科学技術文献ファイルより用語の切り出し

JICST 科学技術文献ファイル中のタイトルや抄録文に用いられている用語の頻度分析のために, 文字列切り出しシステムを開発した。今回の頻度情報は, 機械翻訳で使用している形態素解析や, 従来 JICST で用語分析に使用している TERM-JET は使用しなかった。

抄録文を機械的に文字列の先頭から2~10文字ずつの文字列集合を作成する。切り出し規則としては, タイトル, 抄録とも文の先頭文字列から2~10文字切り出す。そして次の2文字目からまた2~10文字ずつ切り出す。そしてこの切り出された単語集合をソートし, 重複語をカウントする。

科学技術文献ファイルの90年度分と91,92年度分について単語の切り出しを行った。用語は半月分で約1,000万語になり, 速度向上のためにその時点で翻訳辞書との照合を行った (図8)。

(例)

東京都において機械翻訳システムを利用……………
(切り出し語)

東京都, 京都, におい, おい, 機械, 機械翻訳, 翻訳, システム, ス, ステム, 利用

文字列と翻訳辞書との照合を行ったので, 実際の用法と違う場合がある。サ変名詞は, 実際には動詞の用法も含まれる。

5. 2 翻訳辞書との照合

翻訳辞書（科学技術用語35万語）から、2～10文字の用語を異形語を含めて展開する。科学技術文献ファイルからソートされた単語集合と、翻訳辞書とのマッチングを行い、マッチングした単語の頻度をカウントする。頻度情報付きのデータを次の半月分の入力源として、順次半月分ずつ累積していく。そしてJICST科学技術文献ファイルとのマッチングが全て終了した段階で、単語集合から頻度統計リストを出力する。

5. 3 頻度付き辞書

辞書とマッチング処理

約11万語に頻度情報が付与された。(90年度)

約14万語に頻度情報が付与された。(91年度)

約16万語に頻度情報が付与された。(91+92年度) (表7)

90年度と91+92年度を比較すると、頻度情報付きの語数が飛躍的に増加した。初期に作成していた翻訳辞書のソースは、主として対訳辞書から作成していた。この見出し語は、英語論文から切り出した日本語であった。89年から日本語論文について、形態素解析の未知語を収集し、さらに90年から英文DBの後編集結果の未知語中心に作成したので、頻度情報付きの語数が増加したのではないだろうか。

頻度数	日本語見出し	ヨミ	頻度数	日本語見出し	ヨミ
[00175672]	こと	コト	[00039455]	ラン	ラン
[00154334]	より	ヨリ	[00038998]	設計	セツケイ
[00130230]	検討	ケンコウ	[00038604]	使用	シヨウ
[00112596]	シス	シス	[00038040]	デー	デー
[00106622]	システム	システム	[00037850]	抄録	ショウロク
[00106323]	システム	システム	[00037122]	データ	データ
[00104084]	研究	ケンキュウ	[00036958]	同所	ドウショ
[00097123]	結果	ケツカ	[00036950]	情報	ジヨウホ
[00095815]	ため	タメ	[00036589]	ネル	ネル
[00092761]	開発	カイハツ	[00036038]	投与	トウヨ
[00078210]	技術	ギジュツ	[00035920]	クリ	クリ
[00077694]	測定	ソクテイ	[00035657]	なか	ナカ
[00074083]	リン	リン	[00035269]	管理	カンリ
[00073925]	特性	トクセイ	[00034897]	患者	カンジャ
[00068235]	構造	コウゾウ	[00034635]	分析	ブンセキ
[00068147]	影響	エキョウ	[00034329]	組織	ソウキ
[00067269]	また	マタ	[00034210]	作用	サヨウ
[00067164]	処理	ショリ	[00033822]	検査	ケンサ
[00065188]	解析	カセキ	[00033654]	それ	ソレ
[00064451]	示し	シシ	[00033009]	療法	リョウホ
[00063282]	効果	コウカ	[00032744]	材料	ザリョウ
[00062644]	細胞	サイボウ	[00032441]	温度	オンド
[00061939]	実験	ジツケン	[00032331]	条件	ジョウケン
[00060089]	試験	シケン	[00031730]	方式	ホウシキ
[00058538]	方法	ホウホ	[00031412]	臨床	リンショウ
[00057089]	調査	チョウサ	[00031392]	構成	コウセイ
[00056746]	制御	セイギョ	[00031247]	表面	ヒョウメン
[00056727]	プロ	プロ	[00030905]	改善	カイゼン
[00055932]	機能	キノウ	[00030425]	疾患	シツカン
[00055903]	変化	ヘンカ	[00030302]	濃度	ノウド
[00054681]	評価	ヒョウカ	[00030147]	化学	カガク
[00053016]	利用	リヨウ	[00030089]	応用	オウヨウ
[00052343]	問題	モンダイ	[00030087]	分布	ブンブ
[00052312]	特集	トクシュウ	[00030046]	オン	オン
[00050555]	これ	コレ	[00029927]	活性	カクセイ
[00049795]	ロー	ロー	[00029895]	えら	エラ
[00049341]	治療	チリョウ	[00029774]	有効	ユウコウ
[00049059]	もの	モノ	[00029738]	速度	ソクド
[00048757]	考え	カンガエ	[00029415]	以上	ジョウ

図9 頻度順の用語

なお、翻訳辞書語数と頻度付き辞書語数に約2.5倍の差があるのは、頻度情報に使用する翻訳辞書の見出し語の長さが2～10文字であるためであろう。翻訳辞書には、カタカナで始まる見出し語が約10万語あり、10文字以上の専門用語もかなり多い。カタカナ語は主として対訳辞書に由来するものが多く、辞書作成のコーパスの違いによるものである。

JICSTの科学技術論文の抄録を翻訳するためには、最新科学技術用語を取り込んだ最低20万語程度の辞書が必要と思われる(図9)。

5. 5 収録データ項目(図10)

①用語 No

3桁のロットIDと7桁の数字からなっている。

②更新年月日

8桁の数字からなっている。はじめの4桁は西暦年、次の2桁は月、その次の2桁は日を表す。

③頻度

JICST科学技術文献ファイル90年度版のタイトルと抄録に出現した度数。ただし、日本語見出し語が2～10文字までの語に限る。

数字、記号を含む日本語見出し語には、頻度情報は付かない。

最大値は9999である。

④日本語見出し語

JICSTシソーラス用語、JICST科学技術文献ファイルのタイトルと抄録から抽出した語。

⑤読み

日本語見出し語の読み。

⑥異形語

読みが同じで、表記が異なる語。送り仮名の異なる語も含む。

⑦分野コード

分野コードはJICST科学技術分類表の分類コードを付与した。下2桁は00固定、英語見出し語1語に付き最大4個まで付与している。

⑧英語見出し語

(3)分野コード	AA00
英語見出し語(1)	indigo blue
用語No	M010013162
更新年月日	19901128
頻度	8
日本語見出し語	アイアン
読み	アイアン
(1)分野コード	AA00
英語見出し語(1)	iron
用語No	M010013163
更新年月日	19901128
頻度	0
日本語見出し語	アイアンフォーメーション
読み	アイアンフォーメーション
(1)分野コード	DE00
英語見出し語(1)	iron formation
用語No	M010013164
更新年月日	19901128
頻度	0
日本語見出し語	アイアンロー
読み	アイアンロー
(1)分野コード	AA00
英語見出し語(1)	iron law
用語No	M010106449
更新年月日	19901128
頻度	5
日本語見出し語	愛育
読み	アイイク
(1)分野コード	AA00
英語見出し語(1)	tender nurture
用語No	M010013165
更新年月日	19901128
頻度	0
日本語見出し語	アイウェア
読み	アイウェア
(1)分野コード	GP00
英語見出し語(1)	glasses
用語No	M010013166
更新年月日	19910528
頻度	4
日本語見出し語	アイウエオ順
読み	アイウエオジュン
異形語	あいうえお順
(1)分野コード	AA00
英語見出し語(1)	alphabetical order
用語No	M010106450
更新年月日	19901128
頻度	2
日本語見出し語	愛煙家
読み	アイエンカ
(1)分野コード	AA00
英語見出し語(1)	habitual smoker
用語No	M010013167
更新年月日	19910618
頻度	50
日本語見出し語	アイオノマ
読み	アイオノマ
異形語	アイオノマー
(1)分野コード	AA00
英語見出し語(1)	ionomer

図10 頻度付対訳辞書

***** 翻訳抽出ファイル プリントアウト *****

データ

9108544980012

高精細画像ファイリング・システムの商品化。

Commercialization of a high-definition image filing system.

NTTは、印刷に使われるカラー写真の様に極めて高精細なフルカラー画像を、圧縮して記憶化しデータベース化した。

Image encoding is applied to extremely high-definition full color images like color photographs used in printing, and stored into a database by NTT.

INSネットを用いて遠隔地へ高速に伝送するファイリング・システムを商品化した。

Filing system allowing high speed remote transmission using Network was commercialized.

国際標準に準じた画像圧縮方式とオート・チェンジャ型光ディスク装置の採用により、数万枚のA4サイズのフルカラー画像をファイリングできる。

By the adoption of image encoding method in accordance with international standards and auto changer-type optical disk drive, tens of thousands of A4-size full color images can be filed.

9108544980016

大量の文書の効率的な管理運用の実現。

Realization of efficient management and operation of large-volume documents

図11 バイリンガルコーパス

代表的訳語と意味的に訳しわけられない同義語を付与した。最大5語まで採択した。

の出力結果を利用して、Example Based MTへの利用を検討している(図11)。

6. バイリンガル・コーパスの作成

6.1 翻訳結果の後処理

翻訳出力結果に対して、未登録語や係受け、句並列および構文の修正を手で行う。この作業は、複数の翻訳専門家が全抄録文についてチェックを行い、品質向上に努めている。翻訳文は、日本語対英語がほぼ1対1に修正されている。英語文のみ英文DB作成に使用される。

翻訳実行部における後編集結果の日本語と英語

7. 結 論

科学技術の抄録文といったかなり制限された分野で、大量文書の文字列を丹念に集めて、大規模な辞書との照合は行われていなかった。多少 unnecessary 用語の頻度がカウントされている。しかしコーパスと併用すれば、最新科学技術用語も含まれた大規模な辞書であり、言語処理の展開に新しい道を開くものであると確信している。

参 考 文 献

- 1) 高野文雄他; JICST ファイルにおける科学技術用語の用語統計分析第20回情報科学技術研究集会発表論文集1983, PP63-78
- 2) 日英科学技術用語辞書データベース開発に関する報告書日本科学技術情報センター(1983)
- 3) 日本電子工業振興協会: 日本語処理技術に関する調査研究(61-C-535)(1986)

(次回は No. 3 6月号に掲載予定)