

JOIS-Ⅲの自然語-統制語辞書

金子明夫

日本科学技術情報センター

日本科学技術情報センター(JICST)では、自然語から統制語(シソーラス用語)を参照できる「自然語-統制語辞書」(FT-CT辞書)を開発し、JICSTオンライン情報検索システムJOIS-Ⅲに検索支援辞書として搭載した。

開発の経緯及び辞書作成システムについて概観し、JOIS-Ⅲにおける利用方法と検索コマンドについて解説した。さらに同音語、限定句などの特殊な検索(参照)例についても検索画面を示して解説した。

この辞書は、システム解析とJICST情報員の評価により作成したFT-CT変換辞書と統計解析により作成したFT-CT共出現辞書をマージしてできる。現在の辞書規模は、FT-CT変換辞書11万語、FT-CT共出現辞書33万語、FT-CT辞書31万語となっている。

Free Term to Controlled Term Dictionary
in JOIS-Ⅲ

Akio kaneko

The Japan Information Center of Science and Technology

5-2, Nagatacho 2-chome, Chiyoda-ku, Tokyo 100, Japan

JICST(The Japan Information Center of Science and Technology) has developed a free term to controlled term authority file, by which a free term(FT) can be related to relevant controlled terms(CT:i.e.thesaurus terms). Bibliographic retrieval system,JOIS-Ⅲ(JICST Online Information System) has started servicing this file as a refererence dictionary.

This paper serveys a developing process and a system constructing this dictionary, and explains a way how to use this file and retrieval command through JOIS-Ⅲ. We also describe special retrieving examples, which involve terms with scope note, hompograph and homonym by showing a online output.

This dictionary has been constructed by integrating two dictionaries, one is a FT-CT transformig dictionary and another is a FT-CT co-word distionary. The former has been made through specialists of indexing and the latter through statistical cword analysis.

At present, FT-CT transformig dictionary has about 110 thousands entries, FT-CT cword dictionary 330 thousands, and online authority file 310 thousands.

1.はじめに

オンライン情報検索において、自然語による検索とシソーラス用語等の統制語による検索については、従来より種々の観点からその優劣について論じられている。

もし、自然語と統制語とを何等かの関係で関係づけ、両者の情報検索上の利点が同時に得られる検索方法が実現すれば検索上の適合性、網羅性等の向上が図れると期待される。

このため、JICSTでは自然語からそれと最も関連する統制語を案内する機能と、自然語と同じ文献に共出現した統制語を表示する機能をもつ自然語—統制語辞書（以下、FT-CT辞書）を開発し、JOIS-Ⅲに搭載して1980年1月よりサービスしている。

本文では、この辞書の開発経緯と作成システム（FT-CTシステム）とを紹介する。

2.FT-CTシステム

2.1.開発経緯

FT-CTシステムでは、自然語としてJICST情報員が索引した準ディスクリプタと日本語標題からシステムの的に切り出したタイトル語をJICST文献ファイルから抽出している。この他、JICSTシソーラスのディスクリプタ、非ディスクリプタも一種の自然語としてFT-CT辞書に取り込んでいる。また、準ディスクリプタとタイトル語が同形となる場合、それらを同形語として扱っている。これらをFT区分で区別している（表1）。

表1 FT区分

FT区分	自然語
1	準ディスクリプタ
2	タイトル語
3	同形語
4	ディスクリプタ
5	非ディスクリプタ

JICSTでは、自然語と統制語を関係づける辞書として以下の2つの辞書を検討・開発してきた：

- ①FT-CT共出現辞書（以下、共出現辞書）
- ②FT-CT変換辞書（以下、変換辞書）

及び両者をマージしたFT-CT辞書も含めて、3種類の辞書の作成システムとその関係を図1に示す。

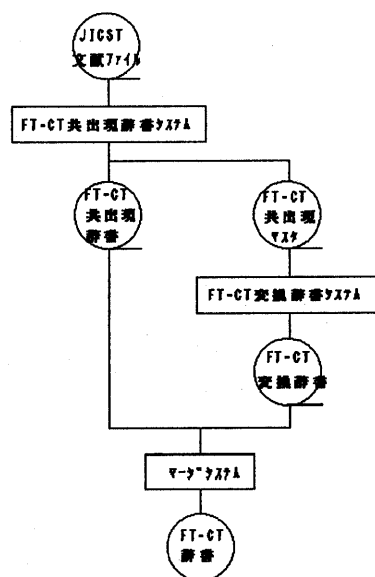


図1

(1)共出現辞書

共出現辞書は、JICST文献ファイルから抽出した自然語について、それと共出現した統制語を共出現頻度順に配列した情報を有する辞書である。

この辞書は、検索支援さらに準ディスクリプタ及びタイトル語と統制語索引との相関性の調査を目的として開発され、1987年にシステムを完成した。1988年に配列順に関する評価実験を実施し、その結果を基に1989年に辞書データを作成した。

(2)変換辞書

変換辞書は、自然語に対して、専門家（JICST情報員）が最適かつ必要十分と考える統制語を有する辞書である。なおこの辞書は、索引の統一すなわち一貫性を図るため及び索引作業を軽減するためという「索引支援」を主目的とした辞書である。

1984年にJICST内部で行った「人手作業による同一文献に対する自然語索引と統制語索引の比較実験」の結果をもとに、1985年から変換解析アルゴリズムを検討した。

この変換解析アルゴリズムを基に1987年にシステムを完成した。1988及び1989年の2年間で約11万語の変換データを作成した。

(3) FT-CT辞書

前記2辞書の開発と並行して、JOIS-Ⅲ用の検索支援辞書も検討した。

検討の結果、両辞書のデータを統合した辞書を検索辞書した方が望ましいと結論により、マージのためのシステムの開発に着手した。

1989年にシステムが完成し、31万語のFT-CT辞書を作成した。FT-CT辞書は1990年1月からサービスに供している。

2.2. システム概略

JICSTのFT-CTシステムは、別名「FT-CT自然語辞書作成システム」と称し、下記の3サブシステムで構成されている：

- ① FT-CT共出現辞書システム
- ② FT-CT変換辞書システム
- ③ マージシステム

例

TI	Corvette 対 944 Turbo 対 RX-7 Turbo (走行テスト)
KW	スポーツカー；走行試験；走行性能；加速性能；燃焼消費率；比較；アメリカ；西ドイツ；日本

以下では、本サブシステムの概要と共出現辞書の作成の実際について説明する。

(1) 自然語の抽出

まず、JICST文献ファイルから、記事毎にFT、CT、及び統一分類コードを抽出したキーワードファイルを作成する。

実作業では、2年分のJICST文献ファイルから自然語を抽出した。JICST文献ファイルは86年8月号から88年7月号までの2年分を選んだ。

このファイルから、各記事の各FT毎に共出現のCTと統一分類データを集積したFT-CT抽出ファイルを作成する。

実際には、計算機処理時間を考慮して、JICST文献ファイルの2号分、あるいは3号分にまとめてFT-CT抽出ファイルを作成した。

各サブシステムの関係と各々の辞書の作成過程を図1に示す。

2.2.1. FT-CT共出現辞書システム

本システムでは、JICST文献ファイル中の自然語（JICST情報員が文献に付与した準ディスクリプタと日本語標題から日本語自動化キーワード作成システムによって切り出されたタイトル語の2種類：FT）とディスクリプタ（CT）が同一の記事（文献）中に出現している回数（共出現頻度）を集計して共出現辞書を作成する。

例えば、下記の例に示すように、標題（TI）中の切り出し語「走行テスト」に対して、同一記事内のKW欄に共出現したディスクリプタ（スポーツカー、走行試験、走行性能など）が対応する統制語としてリストアップされる。このため、「アメリカ」、「日本」などの「走行テスト」に対応しないディスクリプタも累積の対象となり、若干のノイズが含まれる。

FT-CT抽出ファイルは、1個のFTに最大40個のCT情報と最大9個の統一分類情報を持つ。

(2) 集計処理と共出現頻度の算出

次に、抽出の対象となった2あるいは3号分のJICST文献ファイルの全記事の中で各FT毎に共出現データを集計してFT-CT共出現ファイルを作成する。この集計は、1FT当り最大15,000個の共出現CT情報と500個の統一分類コード情報をスタック出来るテーブルを使って計算機処理する。

そして、この2つのテーブルを共出現頻度の降順にソートし、上位から最大120個のCT情報と最大99個の統一分類情報を選んでFT-CT共出現ファイルを作成する。

(3) 累積処理

ここでは、FT-CT共出現ファイルを累積してFT-

CT共出現マスタを作成する。累積処理では、CT情報に関しては240個、統一分類情報に関しては200個のテーブルを用いて、累積処理毎に各テーブルを共出現頻度順にソートし、上位から最大120個のCT情報と最大99個の統一分類情報を選んでFT-CT共出現マスタを作成する。累積処理では、共出現頻度順で下位のCTのデータは累積されないで捨てられる。

2年分の累積処理を行って作成したFT-CT共出現マスタの総語数は約100万語であった。このうち、FTの出現頻度が1のものが全体の73%を占め、出現頻度2が11%、出現頻度3以上が16%であった。

実際には、後で変換辞書とマージしてできるFT-CT辞書の総語数を調節するため、準ディスクリプタに関しては出現頻度1以上、タイトル語に関しては出現頻度3以上のFT、及び1FTに対して10個のCTを抽出して約33万語のFT-CT共出現マスタを作成した。

(4) 共出現辞書作成

FT-CT共出現マスタをフォーマット変換し、共出現辞書を作成する。FTの見出し部（漢字部）が同形の時、CT情報を統合する。

共出現辞書にはシソーラス用語も取り込む。すなわち、ディスクリプタはそれ自身を、また、非ディスクリプタは対応するディスクリプタをCT情報として共出現辞書は持つ。

[USE A AND B] の関係を持つ非ディスクリプタはAとBをCT情報として持つことになる。

2.2.2. FT-CT変換辞書システム

このシステムは、2.1. 開発経緯で述べた変換解析アルゴリズムによって、自然語から統制語を抽出し、抽出された統制語を評価して変換辞書を作成するシステムである。

(1) 自然語の抽出

共出現辞書システムで作成した共出現マスタから抽出した約17万語の準ディスクリプタを変換辞書の見出し語の候補語とした。

(2) 変換解析

上記自然語を変換解析プログラムによって解析

し、変換解析結果を入力原票として出力した。

入力原票には参考データとして共出現した統制語も上位15語まで出力した。

変換解析では、各種の解析用辞書を用いて、基本的には自然語の先頭と末尾からの最長一致法によって統制語を発生させている（図2）。

自然語 : 電気双極子モーメント

統制語1 : 電気双極子

統制語2 : 双極子モーメント

自然語 : 水資源開発

統制語1 : 水資源

統制語2 : 資源開発

図2 最長一致法

(3) 結果のチェックと修正データの作成

入力原票によって、変換解析結果の正否を専門家（JICST情報員）が評価チェックし、「否」の場合は正しい統制語に修正した。

(4) 変換辞書の更新（作成）

1988年度及び1989年度の2ケ年にわたって作成した修正データ及び「正」データをもとに、約11万語の変換辞書を作成した。

2.2.3. マージシステム

開発経緯に記述したように、JOIS-Ⅲ用の検索支援辞書として共出現辞書と変換辞書の辞書データをマージしてFT-CT辞書を作成する。

(1) 同形・同音語のマージ

JOIS-Ⅲでは、FT入力にはFTのフリガナ（読み）で実行する。このため、FT-CT辞書のFTはそのフリガナ部ですべて異なりとなっていなければならない。

共出現辞書と変換辞書のFTが同形、あるいは同音の場合にはマージの仕方が問題となる。

マージ処理では、まず共出現辞書と変換辞書で同形となる（漢字部が同一）FTのデータをマージし、次に同音（フリガナが同一）となるFTの辞書データをマージを行う。

同音処理では、1 FTに対応するCTは上位20語まで採用する。

(2)CT情報の配列

同形語あるいは同音語のマージにおいて、両辞書のうちどちらのCT情報を優先するか（すなわち、CTの配列順序）が問題となる。いずれの辞書の順位を優先するかについては、変換辞書が「あるべき姿」であること、共出現データは（特に低出現頻度の自然語については）相関強度順に並ぶ保証がないため、これらのマージ処理では、変換辞書のCT情報を先に配列し、次に共出現辞書のCT情報を配列することにした。

3. JOIS-ⅢにおけるFT-CT辞書の検索

FT-CTシステムで作成したFT-CT辞書はJOIS-Ⅲへ渡され、オンライン用FT-CT辞書が作成される。自

然語を検索した場合、対応するオンライン用辞書のCTがタグファイルを見にいき、該当するCTの件数が表示されることになる。

自然語の入力は、¥LUコマンドにより行う。

入力形式は、

¥LU NW: 検索語

である。

¥LUコマンドが利用できるファイルは、JICST科学技術文献ファイル（81年～）、JICST科学技術文献ファイル（75年～80年）、JICST・医中誌国内医学文献ファイル（81年～）、JICST科学技術研究情報ファイルの4つである。

¥FILEコマンドにより、上記4つのファイルのいずれかを選択して、¥LUコマンドを入力する（検索例1）。

検索例1

U: ¥FILE 010 ←ファイルの選択
JICST (1981.01 - 1990.02) 4,571,196 (1990.04.14 UPDATE)
JICST COPYRIGHT

S: 質問を開始します 1990.04.26 10:47:32 質問番号 D2L47A06

U: ¥LU NW:キノセイヨク ←自然語の入力

NO	件数	検索語
@L01	333	KW:ケンコウショク 健康食品
@L02	135,641	KW:ショク 食品
@L03	72,609	KW:ヤクザヨ 薬理作用
@L04	209,264	KW:ヒト ヒト
@L05	3,095	KW:ホメオスタシス ホメオスタシス
@L06	108	KW:ショクビンカク 食品科学
@L07	22,240	KW:ショクビンカク 食品加工
@L08	1,308	KW:セイタイホウキョ 生体防御
@L09	27,803	KW:セイリカクセイイン 生理活性因子
@L10	3,834	KW:ロウカ(セイチャヨ) 老化

←FT-CT辞書のCT情報（検索タグ）

S: 出力終了しました

(注)

@L01: LOOK UP番号と呼び、後の検索等でこの番号により検索集合を指定できる。

件数: 検索語（FT-CT辞書のCT情報）の出現頻度であって、共出現頻度ではない。

なお、FT-CT辞書は、あくまでも思いついた概念、すなわち自然語に対応するであろう統制語の集合を提示する参照辞書である。

そのため、検索者が表示されたCT情報から、何を選択するかが重要となる。

4. FT-CT辞書におけるCT情報

以下では、いくつかの自然語についてのFT-CT辞

書のCT情報を説明する。

(1)タイトル語 (FT区分=2) で他に同音となるFTがない場合、共出現辞書のCT情報が10個そのまま表示される (検索例1)。準ディスクリプタの時も同様である。

(2)ディスクリプタ (FT区分=2) で、他に同音となるFTが存在しない場合、それ自身がCT情報として表示される (検索例2)。

検索例2

U: ¥LU NW:NAVIER-STOKES^{ナヴィエ}

NO 件数 検索語

@L01 5,167 KW:NAVIER-STOKES^{ナヴィエ} Navier-Stokes方程式

S: 出力終了しました

(3)ディスクリプタ (FT区分=2) で、他に同音となるFTが存在する場合。

例:「チョウデンドウ」

この例では、ディスクリプタの「超伝導」、タイトル語の「超伝導」、「超電導」が同音となる。

このため、各々のCT情報が統合されて出力される。

(4)同形語 (FT区分=3) で、共出現辞書と変換辞書のCT情報がマージされた場合。

検索例3

U: ¥LU NW:シ^シコ^コト^トウ^ウチ^チヤ^ヤク^クハ^ハ^ハ

NO 件数 検索語

@L01 13,290 KW:SCF^{シコトウチヤクハ} S C F 法

S: 出力終了しました

2.2.3.で述べたようにCT情報の配列順は、変換辞書のデータが先に表示され、共出現辞書データが続く。ただし、先頭の何番目までが変換辞書データかは一律ではなく、自然語によって異なり、かつ判別フラグも表示されていない。

(5)非ディスクリプタ (FT区分=2) で、他に同音となるFTが存在しない場合、USE先のディスクリプタがCT情報として表示される (検索例3)。

(6)非ディスクリプタ (FT区分=2) で、USE先のディスクリプタが複数の場合、USE先のディスクリ

プタが複数表示される (検索例4)。

検索例4

U: ¥LU NW:ネ^ネッ^ッホ^ホウ^ウチ^チヨ^ヨウ^ウケ^ケイ^イ

NO 件数 検索語

@L01 88,377 KW:ケイ^{ケイ}ソ^ソク^{ソク}キ^キ 計測器

@L02 5,125 KW:ネ^ネッ^ッホ^ホウ^ウチ^チヨ^ヨウ^ウ 熱膨張

S: 出力終了しました

(7)限定句がある場合

FTが同形、同音の場合の処理については上で述

べた通りである。

シソーラス用語には限定句をもつものがある。

FT-CTシステムではシソーラス用語もFT区分＝4あるいは5の自然語として扱っていることは既に述べたが、その他に限定句を取って新しく発生させた語もFT区分＝2のタイトル語として処理している。このタイトル語と他の準ディスクリプタ、タイトル語がと同形、あるいは同音の場合が生ずる。一方、JOIS-Ⅲ用FT-CT辞書はフリガナ部（幼音、促音も全て大文字として）で全て異なりとなっている必要がある。

このような状況から、シソーラス用語で限定句付きの場合、FT-CT辞書では限定句付きのFTと限定句を取ったFTの両方の見出し語の下にCT情報を作成している。また、限定句を取って同音となった場合は、上で述べたようなマージ処理を行ってCT情報を統合している。

例えば、フリガナ部から限定句を取って、「シヨウカ」となるシソーラス用語には

昇華【相転移】	シヨウカ【ソウテンイ】
硝化【反応】	シヨウカ【ハンノウ】
消化【代謝】	シヨウカ【タイシャ】
消火【火】	シヨウカ【ヒ】
*しょう果	シヨウカ

USE ベリー類 ベリールイ

の5種類ある。

限定句付きのフリガナを入力すればそれ自身のCT情報が表示される。

限定句なしで「シヨウカ」と入力してFT-CT辞書を検索した場合、上記5つのFTに対するCT情報と他に同音となるFTがあればそのCT情報がマージされて出力される（検索例5）。

検索例5

U: ¥LU NW:シヨウカ

NO	件数	検索語
@L01	1,349	KW:ベリールイ ベリー類
@L02	1,376	KW:シヨウカ(ソウテンイ) 昇華
@L03	1,201	KW:シヨウカ(タイシャ) 消化
@L04	1,106	KW:シヨウカ(ヒ) 消火
@L05	1,904	KW:シヨウカ(ハンノウ) 硝化
@L06	1,635	KW:シヨウキショウ 仕様記述
@L07	2,040	KW:ヨウキヨウシヨウ 要求仕様
@L08	14,079	KW:データ構造 データ構造
@L09	3,305	KW:ソフトウェア仕様 ソフトウェア仕様
@L10	10,969	KW:ソフトウェア工学 ソフトウェア工学

S: 出力終了しました 継続指示は¥LUです

U: ¥LU

NO	件数	検索語
@L11	2,412	KW:システム記述言語 システム記述言語
@L12	15,064	KW:ソフトウェア設計 ソフトウェア設計
@L13	21,781	KW:プログラミング 計算機プログラミング
@L14	43,147	KW:モデリング モデリング
@L15	2,223	KW:データモデル データモデル

@L06 以下のCTはタイトル語「仕様化（シヨウカ）」に対応する共出現辞書のCT情報である。

5. おわりに

現在JOIS-Ⅲに搭載しているFT-CT辞書の規模は31万語である。この程度の規模の辞書を思い付いてた自然語で検索した場合、どの程度ヒットするかが問題となる。JICSTで収集している用語統計データから、文献の日本語標題の切り出し語（タイトル語）を意識しながら（つまり、この程度の語ならタイトル語として抽出されているだろうと予想しながら）自然語を入力すれば、ほぼ1/3程度はヒットするのではないかと予想している。2.2.1.で述べたように出現頻度のFTが70%以上を占めているため、ただ単に語数を増やしただけではヒット率は高くない。FTの収集範囲を拡大し、出現頻度の高いFTを累積していくことが今後の課題である。

文献検索を、より「自然」に近づけるために「FT-CT辞書」を開発した。この辞書を「`¥LUNW:`」コマンドで参照し、検索者自身が必要な統制語を選択し、より効率の良い、洩れのない検索ができることを目指しているが、今回の辞書は第一版であり、辞書内容、及び支援機能を徐々に改良していく予定である。