

語彙知識の自己組織化に向けて
— EDR 電子化辞書の構造と開発手法 —

横井 俊夫
日本電子化辞書研究所

EDR 電子化辞書は語彙に関する知識を集めた知識ベースである。すなわち、単語の文法的属性、単語が表わす概念、概念を理解させるための基本的な知識などを大規模に体系的に蓄積したものである。このような知識ベースを開発するのが電子化辞書プロジェクトの目標であるが、もう一つの大きな目標として、開発のための方法論や支援システムがある。大規模な語彙を獲得していくための新しい技術の研究開発である。この技術は知識の獲得を極力自動化していくもので、電子化辞書の自己組織化機能の実現を目指すものである。この技術によって、電子化辞書の今後の改良・拡張を効率良く進めるためのシステムが実現される。

TOWARD THE SELF-ORGANIZATION OF A LEXICAL KNOWLEDGE BASE
— THE STRUCTURE AND DEVELOPMENT METHOD OF THE EDR ELECTRONIC DICTIONARIES —

Toshio Yokoi
Japan Electronic Dictionary Research Institute (EDR)
Mita-Kokusai Building Annex, 4 - 28, Mita 1- chome, Minato-ku, Tokyo 108, Japan

The EDR Electronic Dictionaries are a knowledge base containing accumulated knowledge on lexicon; a knowledge base that systematically stores, in full scale, grammatical attributes of words, concepts expressed by the words, and information that helps the concepts to be understood. While development of such a knowledge base is one of the EDR project's objectives, its other major objective is preparation of the methodology and support system for this development, that is, to develop new technologies for acquiring very large-scale lexical knowledge. These technologies are to automatize knowledge acquisition to the full extent, and to realize the self-organization of the electronic dictionaries. Through these technologies, a system will be established to efficiently improve and expand the dictionaries in the future.

1. はじめに

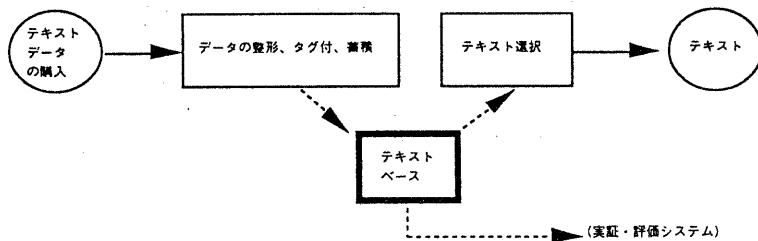
EDR電子化辞書は、語彙に関する知識を集めた知識ベースである。すなわち、単語の文法的属性、単語が表す概念、概念を理解させるための基本的な知識などを大規模に体系的に蓄積したものである。このような知識ベースを開発するのが電子化辞書プロジェクトの目標であるが、もう一つの大きな目標として、開発のための方法論や支援システムがある。大規模な語彙知識を獲得していくための新しい技術の研究開発である。この技術は知識の獲得を極力自動化していくもので、電子化辞書の自己組織化機能の実現を目指すものである。この技術によって、電子化辞書の今後の改良・拡張を効率良く進めるためのシステムが実現させる。

自己組織化機能の実現に向けて、電子化辞書の改良・拡張、維持・管理の統合的なラインを整理し、説明する。このラインは、コンピュータによる支援システムとそれを用いる辞書管理者、辞書作成者の作業が統合化されたものである。ただし、このラインは、現在は、各々がそれぞれに機能している段階であり、統合化は着手されたところである。また、説明は、理解を容易にするため若干細工をほどこしたり、細部を省略したりしてある。

2. 電子化辞書製造ライン

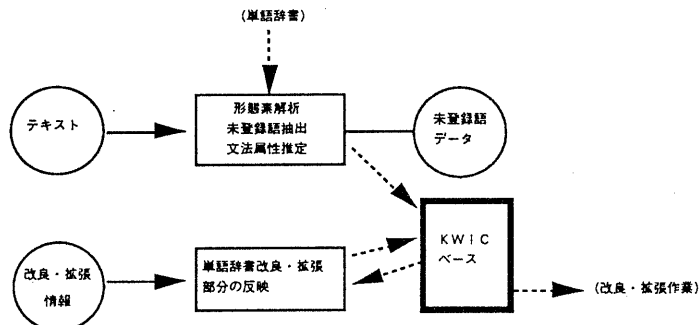
電子化辞書の改良・拡張、維持・管理の統合的なラインは、辞書を極力自動的に増殖、拡大していくためのラインであり、電子化辞書の製造ラインと呼ぶべきものである。以下に、この製造ラインの主要部となる部分を工程の流れ図によって説明する。ただし、紙面の都合上、説明は最少限にしてあるので、図の上で理解されたい。また、電子化辞書の内容を実感していただくために、付録Iに概念体系、付録IIに概念記述の具体的な内容を例示してある。

(1) テキストベース



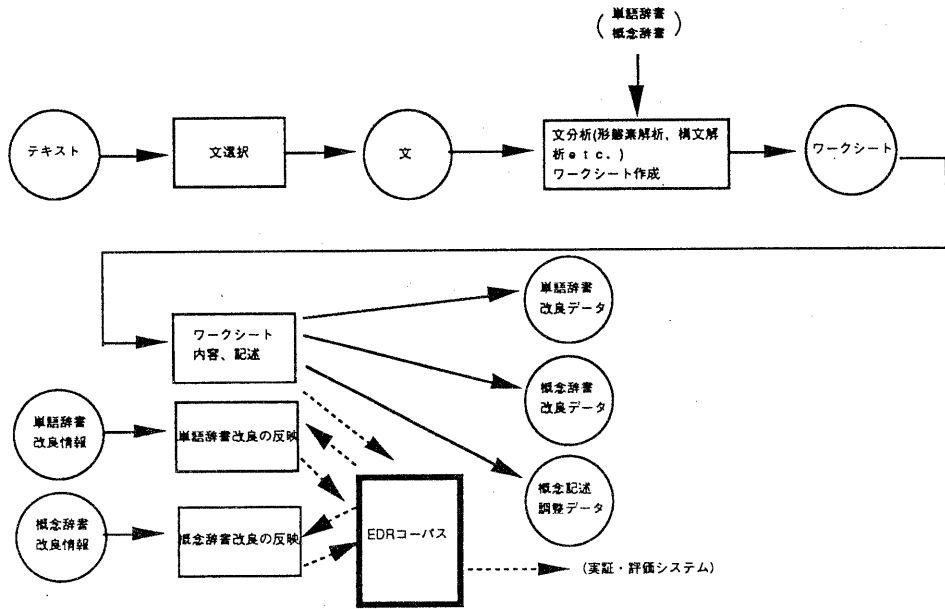
整形、タグ付等、選択等のための支援プログラムが用意されている。

(2) KWICベース



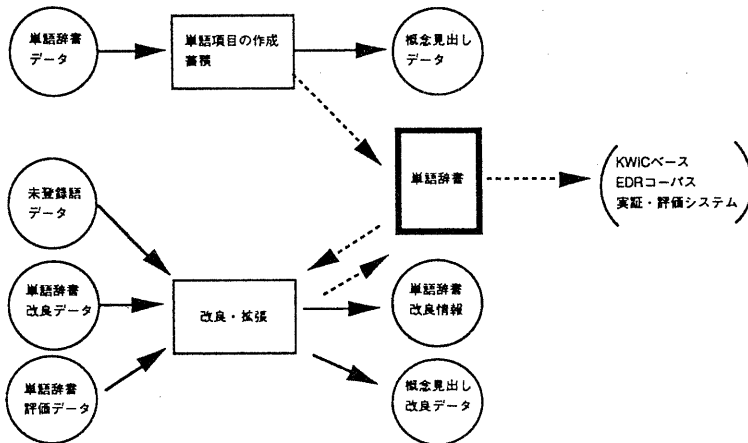
形態素解析、未登録語抽出、文法属性の推定とそれに基づくワークシートの作成は、すべてプログラムによって行われる。プログラムによりワークシートの記述内容がチェックされ、未登録語データにまとめられる。改良・拡張部分の反映もプログラムによって支援される。

(3) EDRコーパス



文選択の支援、文分析、ワークシート作成、各辞書の反映作業支援等に対するプログラムが用意される。

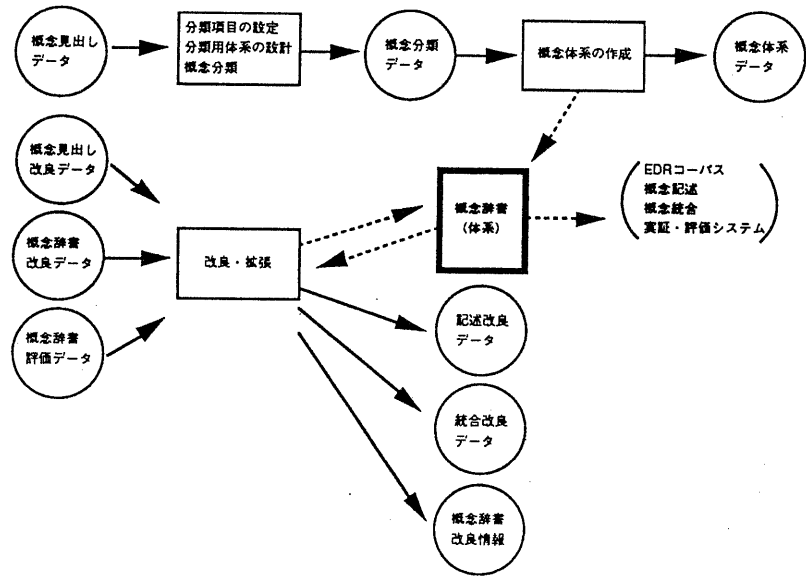
(4) 単語辞書



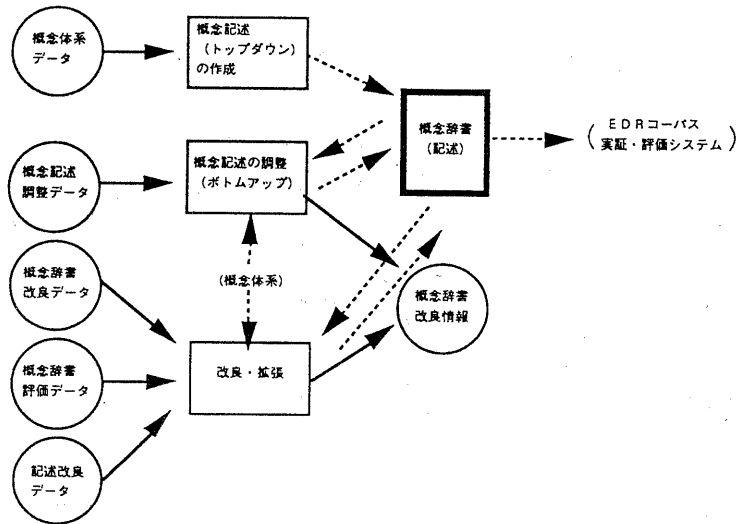
単語項目の作成、追加、変更、削除にともなう整合性のチェックを行うプログラムが用意される。

(5) 概念辞書

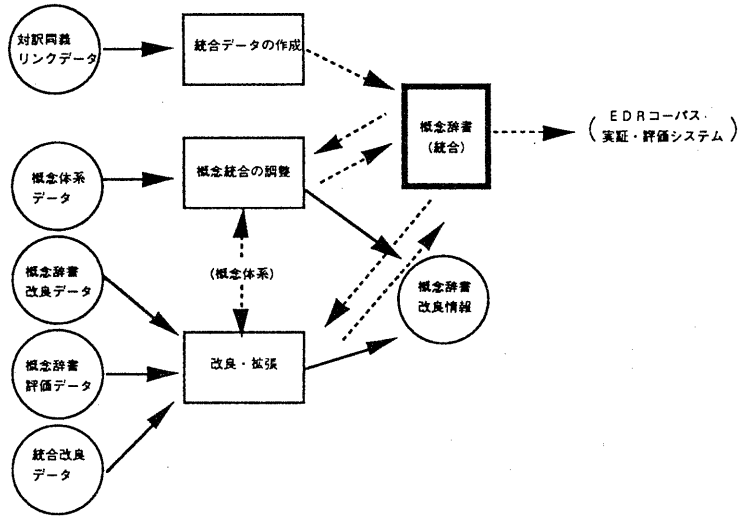
(i) 概念体系 (付録I 参照)



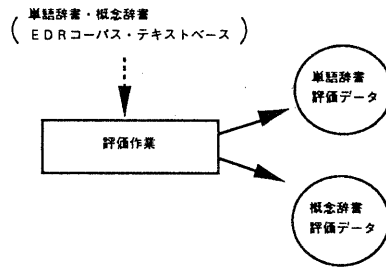
(ii) 概念記述 (付録II参照)



(iii) 概念統合



(6) 実証・評価システム



付録 I : 概念体系

(1) 分類項目の一部

この分類項目に応じて概念見出しが分類され、体系化されている。更に、現在この体系を元に体系の細分化を進めている。

形式

概念ID 分類番号/概念見出し/変更概念見出し(該当する概念見出しを持つ単語例)

- ▲ 二個所以上に配置されているもの、いわゆる多重リンクの部分
- , *, +, *, #, @などは、作業管理上、つけられている記号

3aa966	●O	000000	概念
30f6ae	●A	111-----	静物/c#thing{具体物, 物}
30f6af		111----	生命体(生物)
30f6bf		1112---	動物(動物)
30f6c1		111218	獲で捉えた動物
30f6c2		111211	鳥(鳥)
30f6c3		111212	魚(魚)
30f6c4		111213	貝(貝)
30f6c5		111214	哺乳類(哺乳類, 猿, 人間)
30f6c6		111215	爬虫類(爬虫類, 蛇)
30f6c7		111217	昆虫(昆虫, 虫)
30f6c9		111216	両性類(蛙)
3aa91a	@	111220	役割で捉えた動物
3aa91b	@	111221	家畜(家畜, 乳牛)
3aa91c	@	111222	ペット(ペット, 愛がん動物)
30f6ca		1113--	植物(植物)
30f6cb		111313	獲で捉えた植物
30f6cc		111311	樹木(木)
30f6cd		111312	草(草)
3f9638		114353	穀物(穀物, 米)
3aa91d	@	111320	役割で捉えた植物
30f6d2	#	11148	食用植物/食べられる部分をもつ植物(山菜, 果樹)
444576		114355	果物/果物がなる植物(柿の木, 野菜)
*30f6cf	▲	野菜	
444577		114353	穀物/実が穀物である植物(米)
30f6d0	#	11146	香料となる植物(ハーブ)
30f6e7	#	11162	香料となる植物の部分
30f6d1	#	11147	染料となる植物(藍, 紅花)
30f6d3	#	11149	飼料用植物(牧草)
30f6e9	#	11164	飼料となる植物の部分(まぐさ, 元草)
30f6d5	#	1114b	嗜好品となる植物(葉たばこ)
30f6eb	#	11166	嗜好品となる植物の部分(コーヒー, 煙草)
30f6d6	#	1114c	衣服の素材となる植物(綿, 麻)
30f6ec	#	11167	衣服の素材となる植物の部分(綿花, 繊維)
30f6ed	#	11168	木材となる植物の部分(桜木, 唐木)
3aa91e	@	111321	薬草(薬草)
3aa91f	@	111322	観葉植物(観葉植物)
30f6b0		1111--	人間(人, 人間)
30f6b2	*	1112-	社会との関係で捉えた人間(社会人)
3aa918	@	111310	役割で捉えた人間(責任者, 主役, 担当者, 売り手, 買手, 与え手)
30f6b6		11131	職業で捉えた人間(先生, 会社員, 医者)
30f6b7		11132	組織内の役割で捉えた人間(上司, 部下, 代表者, 社長, 部下)

(2) 具体的な上下関係例
(楽器の下位項目)

形式

下位概念のID 関係ind-of 上位概念ID

Fr:下位概念IDの概念見出し[その概念見出しを持っている単語

No Headwordと表示してある場合は、分類項目に該当する.]

To:上位概念IDの概念見出し

_000302	444601	<kind-of>	30f701	= 1/0/0
	Fr:	管楽器	[*** No Headword ***]	
	To:	11431f楽器	[*** No Headword ***]	
_000303	444602	<kind-of>	30f701	= 1/0/0
	Fr:	打楽器	[*** No Headword ***]	
	To:	11431f楽器	[*** No Headword ***]	
_000304	444603	<kind-of>	30f701	= 1/0/0
	Fr:	弦楽器	[*** No Headword ***]	
	To:	11431f楽器	[*** No Headword ***]	
_000305	44464a	<kind-of>	30f701	= 1/0/0
	Fr:	電子楽器	[*** No Headword ***]	
	To:	11431f楽器	[*** No Headword ***]	
_000306	44464f	<kind-of>	30f701	= 1/0/0
	Fr:	楽器を演奏する際に用いる器具	[*** No Headword ***]	
	To:	11431f楽器	[*** No Headword ***]	
_000307	1ee9b5	<kind-of>	30f701	= 1/0/0
	Fr:	楽器の共鳴箱[サウンドボックス]		
	To:	11431f楽器	[*** No Headword ***]	
_000308	3c5e5a	<kind-of>	30f701	= 1/0/0
	Fr:	楽器につける反響板[サウンディングボード]		
	To:	11431f楽器	[*** No Headword ***]	

(3) 楽器の下位項目の一つの鍵盤楽器に属する概念見出しの一部

_000001	0ee6f8	<kind-of>	1f585b	= 1/0/0
	Fr:	ハーブシコードという楽器(クラブサン)		
	To:	鍵盤楽器[キーボード]		
_000002	0efa6b	<kind-of>	1f585b	= 1/0/0
	Fr:	鍵盤で演奏する楽器(鍵盤楽器)		
	To:	鍵盤楽器[キーボード]		
_000003	0f9e42	<kind-of>	1f585b	= 1/0/0
	Fr:	セレスタという鍵盤をもつ打楽器(セレスタ)		
	To:	鍵盤楽器[キーボード]		
_000004	1042e6	<kind-of>	1f585b	= 1/0/0
	Fr:	Hammondオルガンという鍵盤楽器の一つで、パイプオルガンに似た音を出すもの[Hammondオルガン]		
	To:	鍵盤楽器[キーボード]		
_000005	104d6d	<kind-of>	1f585b	= 1/0/0
	Fr:	ボタン式の鍵盤をもつ、アコーディオンに似た、タンゴ演奏の主要楽器(バンドニオン)		
	To:	鍵盤楽器[キーボード]		
_000006	106696	<kind-of>	1f585b	= 1/0/0
	Fr:	オルガンという楽器(オルガン)		
	To:	鍵盤楽器[キーボード]		
_000007	1ee346	<kind-of>	1f585b	= 1/0/0
	Fr:	立式のピアノ[アップライトピアノ]		
	To:	鍵盤楽器[キーボード]		
_000008	1ef41c	<kind-of>	1f585b	= 1/0/0
	Fr:	コテージピアノ[コテージピアノ]		
	To:	鍵盤楽器[キーボード]		

付録II：概念記述

(1) 関係子一覧

関係子は、以下の通りである。なお、例は理解を容易にするためのものであり、概念辞書に、このままの概念関係が記述されているわけではない。

agent	有意志的動作を引き起こす主体。有意志者、自動物が主体になる。 例) " 動物は食べる" 〈食べる〉-agent→〈動物〉
object	1) 動作・変化の影響をうける対象 例) " 食物を食べる" 〈食べる〉-object→〈食物〉 2) 属性をもつ対象 例) " トマトは赤い" 〈赤い〉-object→〈トマト〉
beneficiary	事象や状態の受益者 例) " 人間に有用である" 〈有用〉-beneficiary→〈人間〉
purpose	目的 例) " ゴルフに行く" 〈行く〉-purpose→〈ゴルフ〉
manner	動作・変化のやり方 例) " ゆっくり話す" 〈話す〉-manner→〈ゆっくり〉
implement	有意志動作における道具・手段 例) " ナイフで切る" 〈切る〉-implement→〈ナイフ〉
material	材料または構成要素 例) " 牛乳から作る" 〈作る〉-material→〈牛乳〉
time	事象の起こる時間 例) " ある時刻に起きる" 〈起こる〉-time→〈時刻〉
time-from	事象の始まる時間 例) 定刻から働く" 〈働く〉-time-from→〈定時〉
time-to	事象の終わる時間 例) " 定時まで働く" 〈働く〉-time-to→〈定時〉
place	事象の成立する場所 例) " 部屋で遊ぶ" 〈遊ぶ〉-place→〈部屋〉
source	事象の主体または対象の最初の位置、最初の状態 例) " 京都から来る" 〈来る〉-source→〈京都〉
goal	事象の主体または対象の最後の位置、最後の状態 例) " 東京に行く" 〈行く〉-goal→〈東京〉
scene	事象の成立する場面 例) " ドラマで演じる" 〈演じる〉-scene→〈ドラマ〉
quantity	物・動作・変化の量 例) " りんごの重量" 〈りんご〉-quantity→〈重量〉
number	数 例) " 3 k g" 〈k g〉-number→〈3〉
condition	事象・事実の条件関係 例) " 泣くので帰宅する" 〈帰宅〉-condition→〈泣く〉

cooccurrence	事象・事実の同時関係 例) " 泣きながら帰宅する" 〈帰宅〉-cooccurrence→〈泣く〉
sequence	事象・事実の時間的前後関係 例) " 掃除してから帰宅する" 〈掃除〉-sequence→〈帰宅〉
part-of	全体一部分関係 例) 〈羽〉-part-of→〈鳥〉 〈鳥肉〉-part-of→〈鳥〉
kind-of	継承関係(継承すべき属性や性質を持つ概念と関係づける) 例) 〈鳥〉-kind-of→〈食物〉
similar	類義関係 例) 〈鼻〉-similar→〈nose〉

