1995年度日本オペレーションズ・リサーチ学会
春季研究発表会

E－Rモデルにおける実体定義文の構成方法（第1報）

－国語辞書における意味説明文の特性把握－

01011750　日本電信電話（株）　出原良夫　IDEHARA　Yoshio

1．はじめに

景気の長期低迷傾向にバブルの崩壊が加わり，昨今の企業は厳しい経営環境にさらされている．このような中で情報技術を活用することにより顧客志向の業務展開を行うことを目指し，旧来の分業体制を抜本的に見直して仕事の進め方を根本的に革新するリエンジニアリングが注目されている．このリエンジニアリングの実現においては，部門間でのデータベースの共有[1]，あるいは統合データベースの活用[2]が重要とされる．

一方，全社的，あるいは複数部門にまたがる情報システムの開発を，企業個々の情報戦略のもとに工学的な各種の技法を体系的に組み合せて実現を図るＩＥ(information engineering ：インフォメーション・エンジニアリング)が提唱されている[3]．ＩＥにおいては，情報中心の立場から企業全体，あるいは複数部門をまたがる業務領域で使用されるデータを冗長なくデータベース化するために，その論理構造を一定の形式で表現するデータモデルの作成が重要となる．

ＩＥ等データ中心のアプローチにおいてはデータベースの概念設計が重要となるが，この場合においては，現実世界の対象の意味をデータ構造に表現しやすいＥ－Ｒモデル(entity-relationship model)[4]が使用される．Ｅ－Ｒモデルは，業務で使用されるデータの有意な集合により表現した実体(entity)と，それらの間の関連（relationship）を表現する構造をもつものである．このモデルは，一般的に，①実体を四角形，また，それらの間の関連を矢線等により視覚的に表現したＥ－Ｒ図，②実体名とその業務上の意味を日本語により表現し一覧にした実体定義表，③実体名とこれに対応するデータ項目名を一覧にした属性定義表，④実体間の関連についての情報を一覧にした関連定義表の併用によって表現を行う．（[4]他）

これらはいずれも，個々の実体の意味や関連について開発，及びユーザ部門等で共通の認識を持ち，データベースの開発に反映するために有効な表現形式であるが，その作成作業は工学的アプローチによる生産性向上の図りにくいものであるため，作業者による個人差が生じやすく，また能率も上がりにくい現状にある．特に，②の実体定義表は，実体の意味をまず簡潔に把握するために有益な形式であるため，その作成方法について一定のルール化された手法を作成することは有益と考える．

本研究では，実体定義表の作成における実体の意味定義のための，一定の標準化された手法（定義文の説明法の形式，構文パターン，及び図示による方法等）の提案を目的とし，次のステップによって検討を進めることとする．

Step 1　実体定義文作成に個人差が生じる要因を分析するため，語の意味説明方法（説明文，及び挿図等）のサンプルとして国語辞書を用い，語の種類，説明パターン等による意味説明文の辞書間差等の特性を把握．

Step 2　語の意味の一般的な説明方法（説明法，及び構文パターンの類型化，図示法等）について調査・研究．

Step 3　実体の意味定義に最適な表現手法（実体定義文に最適な説明法，構文パターン，及び図示方法等）について研究のうえ提案，検証．

本報では，主としてStep 1に関する現在までの研究結果について述べる．

2．国語辞書収録語の意味説明文特性調査の方法，及びその結果

一般に，国語辞書においては，同義語・類義語との置き換えによる対訳型，その語の意味を直接説明する説明型，あるいはこれらの併用により意味記述が行われている．このうち，説明型の意味記述には，言語的解説と，百科的解説とがある．さらに，語はその意味が多数のもの（多義語）と単一なもの（ここでは仮に非多義語と呼ぶ）とに分かれる[6]．

意味説明文の詳しさの特性についてみると，対訳型においては辞書間で大同小異であるが，説明型においては編集方針や編者の個性が反映され[6]，また，多義語の場合に，その全ての語意を掲載するか，特定のものに限定するか（ここでは以降，網羅度と呼ぶ）によって差がみられる．このような理由によって同一の語の説明文の詳しさに辞書間で差が生じると考えられる．

ここでは，この詳しさの違いを個人差の生じ得る要因に由縁した結果ととらえ，それが，①多義語・非多義語別，②説明型（言語的解説・百科的解説）別，③対訳型・説明型・併用型別，④品詞別，にどのような特性を示すかを把握するため，説明文文字数のバラツキを評価尺度とし，以下によりその特性を定量的に調査・分析する．

2.1　対象辞書の選定

最近発行の国語辞書を中心に表1のとおり選定した．

2.2　サンプル語の選定

一つの辞書の各収録語の説明文平均文字数を，サンプリングによる許容誤差率10%，信頼度95%以下として推定するために必要な語数180をここでの必要サンプル量と決定し，全対象辞書に収録されている語の中からランダムに選定した．その品詞内訳を表2に示す．

2.3　サンプル辞書の選定

サンプル語の説明文文字数データをもとに対象辞書間で平均値の差の有意差検定を行い，平均値に有意な差のない辞書群として1～4を選定しサンプル辞書とした．

2.4　意味説明文の文字数データの主成分分析

語の説明文文字数の辞書間のバラツキの特性を抽出・把握するため，サンプル辞書別サンプル語別の説明文文字数データを表3に示すデータ行列として扱い，これに

対し分散・共分散行列による主成分分析を実施した．

この結果は表４のとおりであり，第１，２主成分による累積寄与率が90％を示している．したがって，これを第１，２主成分の散布図で示せば図１のとおりとなる．これから第１，２主成分の寄与率比に基づく重み付き平均以上のバラツキを示す51語（ここでは以降，バラツキの大きい語と呼ぶ）の抽出を行った．（表５）

２．５　説明文文字数のバラツキの特性の分析結果

２．４で抽出された説明文文字数のバラツキの大きい語について，各種の分析を行った結果を表６に示す．

表１　対象辞書

辞書番号	名称（版数）	発行年月（初版発行年）	出版社	語数（万語）	ページ数	価格（発売年）
辞書１	GR（２）	1993.1（1985）	SG	６．０	約 1510	2500（1993）
辞書２	IK（２）	1971.2（1963）	IN	５．７	約 1160	680（1971）
辞書３	IK（５）	1994.11（1963）	IN	６．２	約 1800	2300（1994）
辞書４	SM（３）	1981.2（1972）	SS	７．２	約 1290	1700（1982）
辞書５	SK（４）	1992.2（1960）	SS	７．３	約 1360	2200（1994）
辞書６	KZ（３）	1983.12（1955）	IN	約２０	約 2670	5800（1983）
辞書７	KZ（４）	1991.11（1955）	IN	約２２	約 2860	6500（1992）

表２　ｻﾝﾌﾟﾙ語の品詞内訳

品詞名	ｻﾝﾌﾟﾙ数	構成比（％）
名詞	１２３	６８．３
代名詞	２	１．１
名詞＆形容動詞	３	１．７
名詞＆動詞	２６	１４．４
動詞	１９	１０．６
形容詞	２	１．１
形容動詞	３	１．７
副詞，接続詞	２	１．１
計	１８０	１００

表３　文字数のデータ行列

ｻﾝﾌﾟﾙ辞書	サンプル語				
	1	2	----	j	------
1	X_{11}	X_{12}	----	X_{1j}	-----
2	X_{21}	X_{22}	----	X_{2j}	-----
⋮	⋮	⋮		⋮	
i	X_{i1}	X_{i2}	----	X_{ij}	-----
⋮	⋮	⋮		⋮	

表４　主成分分析の結果

主成分	Z_1		Z_2	
ｻﾝﾌﾟﾙ語	固有ﾍﾞｸﾄﾙ	因子負荷量	固有ﾍﾞｸﾄﾙ	因子負荷量
1	-0.004	-0.035	0.134	0.997
2	0.050	0.794	0.004	0.051
⋮	⋮	⋮	⋮	⋮
180	-0.063	-0.961	0.019	0.223
固有値	9247.9		5472.5	
寄与率	56.4	%	33.4	%
累積〃	56.4	%	89.8	%

主成分	主成分得点			
	辞書１	辞書２	辞書３	辞書４
Z_1	45.19	-90.49	-68.14	113.43
Z_2	-105.57	20.02	18.20	67.36

表５　バラツキの大きい語の品詞内訳

品詞名	抽出数	構成比（％）
名詞	３３	６４．７
代名詞	１	２．０
名詞＆動詞	７	１３．７
動詞	８	１５．７
副詞，接続詞	２	３．９
計	５１	１００

表６　分析結果

（１）全ｻﾝﾌﾟﾙ語（A）とﾊﾞﾗﾂｷの大きい語（B）との分布の比較

①	(A)	(B)	適合度検定
多義語	55 (31)	30 (59)	→分布が相違
非多義語	125 (69)	21 (41)	
計	180 語	51 語	

②		(A)	(B)	適合度検定
説明型	百科的	12 (7)	9 (18)	→分布が相違
	言語的	165 (93)	42 (82)	
	計	177 語	51 語	

③	(A)	(B)	適合度検定
対訳型	3 (2)	0 (0)	→分布相違と言えず
説明型	158 (88)	46 (90)	
併用	19 (10)	5 (10)	
計	180 語	51 語	

④	(A)	(B)	適合度検定
名，代名詞	125	34	→分布相違と言えず
名＆動，名＆形動	29	7	
動，形，形動，副，接	26	10	
計	180 語	51 語	

注：（ ）内の数値は構成比（％）

（２）説明文文字数のバラツキの大きい語のうち，意味説明の詳細度の差によるものの比率

多義語---- 20／30
非多義語- 21／21
計　41／51＝80.4％

（３）説明文文字数のバラツキの大きい多義語のうち，網羅度の差によるものの比率

22／30＝73.3％

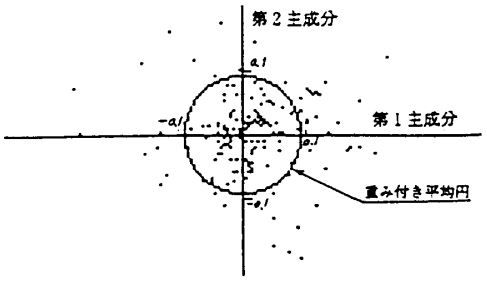

図１　第１，２主成分の散布図

３．まとめと今後の予定

分析結果より，説明文文字数の辞書間におけるバラツキに関する次の特性が把握できた．

① 多義語は非多義語に比べバラツキの大きいものが多く，その約７割は網羅度の差に起因するものである．

② 説明型の語のうち，バラツキの大きいものの約８割は説明の詳細度の差に起因している．

③ 説明型の語のうち，百科的解説の語は，言語的解説の語に比べバラツキの大きいものが多い．

④ バラツキの大きい語は特定の品詞に偏らない．

一方，[7] においては，R. Robinsonが提案した同義語提示，分析的，総合的，暗示的，表示的，例示的，規則提示，の７種類の語の説明法が示されている．

次は，今回得た分析結果に加え言語学，あるいは記号論的考察等により，説明法，構文パターン等の類型化等を行って語の意味の一般的説明方法を整理・体系化し，合わせて意味を一層明確にするための図示による方法等についても検討する．

さらに最終的には，Ｅ－Ｒモデルにおける実体定義の場合に特化した最適な説明法，構文パターン，あるいは図示方法等について提案・検証を行うこととしたい．

検討に当たって数々の有益な助言をいただいた林担当部長に感謝の意を表したい．

参　考　文　献

[1]Hammer,M. and Champy,J.：*Reengineering the Corporation*. Harper Business，1993.

[2]細田正勝：情報技術が産みだすリエンジニアリング．ｵﾍﾟﾚｰｼｮﾝｽﾞ・ﾘｻｰﾁ，Vol.39，No.8（1994），402－409.

[3]Martin,J.：*Information Engineering*.Prentice-Hall,Inc.，1989．（三菱ＣＣ研究会ＩＥﾀｽｸﾌｫｰｽ訳：インフォメーション・エンジニアリング．トッパン，1991．）

[4]Chen,P.P.:*The Entity-Relationship Model*. ACM Transactions on Database Systems,Vol.1，No.1(1976),9-36.

[5]町原宏毅，出原良夫：情報基盤整備のためのアプローチ．電子情報通信学会技術研究報告，DE94-54 (1994-09)，81-87.

[6]金田一春彦，林　大，柴田　武　他編：日本語百科大事典．太修館書店，1988.

[7]Waldron,R.A.：*Sense and Sense Development*. Andre Deutsch Ltd.，1979．（築島兼三訳：意味と意味の発展．法政大学出版局，1990．）