複合語解析技術を用いたデータ項目名称の標準化手法

黒川 清† 中川 優† 関根 純†

近年，情報の高度な利用を促進するために，データを部品として体系化し各システムで利用する，データ中心のシステム構築法に期待が高まっている．この構築法では，データのさまざまな属性についての体系化が必須となり，特に，データの概念を表すデータ項目名称の標準化は，異種のデータベースを統合利用するシステムの構築では重要な技術となる．我々は，Durell の提案するデータ項目名称の命名規則を基本に，標準的なデータ項目名称の付与とチェックを支援するデータ標準化ツールを作成した．本ツールの名称チェック機能の核となる技術に，データ項目名称を構成する用語を辞書化する技術がある．しかし，用語辞書構築の基準にはヒューリスティックな要素があり，このような用語辞書の品質は不安定で，これを用いた名称チェック機能は信頼性に欠け広く実用に供しづらかった．本稿では，データ項目名称が複合語であることに着目し，複合語解析結果の品詞，意味カテゴリの概念を導入した用語辞書の構築方法を明らかにする．次に，用語の品詞情報を活用した新しい命名規則を提案する．これらの提案により，用語辞書の構築が高い精度で容易に行えるようになり，また，人の解釈では問題がないにもかかわらず名称チェック機能によりエラーとなるデータ項目名称をなくすことができた．これらの成果により，今後，整備されたデータ項目名称を核に，他のデータ属性に関する標準化作業が進展するものと思われる．

Standardization Method of Data Element Names Using Compound Word Analysis

Kiyoshi Kurokawa,† Masaru Nakagawa† and Jun Sekine†

This paper describes the method of constructing a semantic dictionary which assists data administrators in standardizing data element names of databases. The semantic dictionary consists of words constituting data element names and their attributes, and it is built in a data standardization tool which supports functions for checking data element names based on Durell's naming convention in order to keep data element names consistent, and functions for clustering data elements in existing databases by similarity among words. New methods of segmenting words and deciding word attributes necessary for checking data element names are proposed based on the part of speech and the semantic category of each word. Also, an improved version of Durell's naming convention allowing more natural naming in Japanese is proposed. These proposals enable data administrators to add new words to the semantic dictionary during their data standardization efforts and to improve the reliability of the semantic dictionary.

1. はじめに

従来のプロセス中心の設計法によるシステム開発では，短期間でのシステム更改が困難となり，また膨大なバックログが発生するなどの現象を招き，企業ニーズに十分応えきれないという反省から，近年，企業に蓄積されているデータを体系的にデータベース化し利活用しようという，データ中心のシステム設計法[1]に対する期待が高まっている．そこでは，情報の源はデータであるという立場から，情報の多様化を吸収するために部品としてのデータを体系化し，その組み合わせで各システムからの要求に対応することを考えている．この部品に相当するデータの名前，型，桁数など，さまざまな属性についてのデータの体系化[2]のうち，データ項目名称（日本語名）の標準化[3]は重要な技術となる．すなわち，システムアナリストやシステム設計者いずれに対しても，曖昧性が少なくわかりやすい名称を生成できれば，共通概念の下でデータ設計，および，業務設計が図れるからである．

我々は，このような標準化された名称の生成を可能とする Durell の提案する命名規則[4]を基本に，日本語のデータ項目名称の付与とチェックを支援するデータ標準化ツール[5]を作成した．本ツールは，命名規則に基づく名称チェック機能，および，類似する既存のデータ項目名称を検索する機能などを実現している．これらの機能の核となる技術に，データ項目を構成し

† NTT 情報通信網研究所
Network Information Systems Labs., NTT

ている用語の役割や用語間の類似度などの基本情報を管理する用語辞書の構築・利用技術がある．しかし，従来，データベース技術の分野で，このような本格的な用語辞書を構築するための技術は確立されておらず，専門家でさえ容易に構築することはできなかった．例えば，現在までに我々が構築した用語辞書は，用語登録の基準が不明確なまま構築したため品質が悪く，これを用いた名称チェック機能の信頼性には問題があり，広く実用に供しづらかった．

本稿では，用語辞書を強化するため自然言語処理における成果を活用することを考え，特に，データ項目名称が複合語であることに着目し，複合語解析結果の品詞，意味カテゴリに基づく新しい用語辞書の構築方法を提案する．また，多数のシステムにおける実データの評価により，従来から提案されている命名規則では多様な日本語のデータ項目名称に対応しきれない場合が多々あることがわかった．そこで，用語の品詞情報を活用した新しい命名規則を提案する．これら二つの提案により，自然言語の専門家でなくともデータ標準化の作業中でも用語辞書の構築が可能となり，また，より自然な品質のよいデータ項目名称の付与が可能となり実用性が高まることがわかった．以下，2章で従来用いてきた命名規則とその問題点を明らかにし，3章で用語辞書構築に関する用語区切り，用語の役割の判定法について述べ，4章で企業内のデータ項目名称に適した命名規則の改善案を提案し，5章でまとめと今後の課題を示す．

2. 既存の技術とその問題点

本章では，データ項目名称の付与に関して我々の開発したデータ標準化ツールで実現している機能を明らかにし，具体的な問題について分析を行った．ここで，データ項目とはデータベースにおけるデータ操作の最小単位であり，データ項目名称とはデータ項目に付与される日本語の名前であると定義する．例えば，管理すべき顧客の識別番号を表すデータ項目には「顧客コード」というデータ項目名称を付与する．

2.1 命名規則

データ項目名称は用語の組み合わせ，すなわち，複合語として構成されている．この用語の組み合わせに構文的制約を加えることで，誰にもわかりやすい名称にするのが命名規則の考え方である．日本語の特徴を考慮し Durell の命名規則に制約を加えた提案[6]とその例を**図1**に示す．

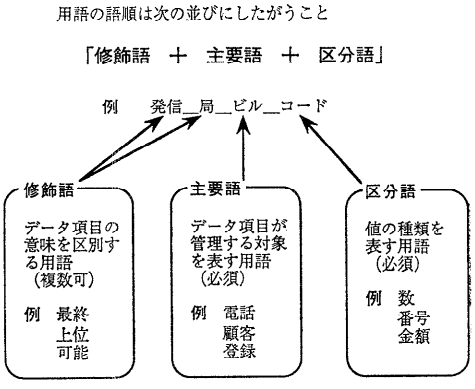


図1 命名規則
Fig. 1 Naming conventions of data element.

例えば，データ項目の名称として「ビル」を登録するものとする．しかし，この名称だけでは「ビル」の「名前」を表すものか「コード」を表すものかわからないという問題がある．そこで，命名規則ではデータ項目がどのような値かを表す用語（この例の場合「コード」），および，その用語の前方に，管理する対象が何かを表す用語（この例の場合「ビル」）が必須であることを規定する．このとき，管理する対象を表す用語を主要語，内容を表す用語を区分語と呼ぶ．さらに，この二つの用語から構成される名称だけでは，データ項目名称が一意にならない場合，意味を区別する用語である修飾語を任意個，主要語の前に接続させて識別性を高める．

2.2 支援ツールとその機能

新規システムにおけるデータの体系化を図るため，および，既存システムのデータ項目名称を分析してその名称の整備，利活用を促進するために，前節で示した命名規則を適用した，データ標準化支援ツール（DB prompt/NAME）を作成した．本ツールのデータ標準化支援機能[5]として，新規開発システムのデータ項目名称を統一するための名称チェック機能がある．この名称標準化のための核となる技術は，データ項目名称に使用される用語を納めている用語辞書の構築・利用技術であり，この用語辞書には，以下のような情報が管理されている．

（1） 用語

データ項目名称を構成する最小の単位であり，その一語で意味を持つものとする．

（2） 標準用語

企業内で使用する用語のうち，公式に通用する用語を標準用語とする．例えば，カタカナ英語における語

尾の長音記号の付与を統一するような，用語の表現を統一するために，類似した用語はそのうちの一つを選択して標準用語とする．例えば，「ユーザー」，「ユーザ」「USR」などの用語は，「ユーザ」を標準用語にすると決める．

（3） 用語種別

Durell の命名規則における，区分語，主要語，修飾語のような用語の役割の概念が，データ項目名称に出現している用語でどのように使われているかを分析した結果，これら三つの概念は一つに固定されるわけではなく，状況に応じてかわり得ることがわかった．これを分類すると，用語は「修主区」，「修主」，「修」の3種類に分類・整理できた(**表1**)．用語種別が「修主区」の用語は修飾語，主要語，区分語として，また，「修主」の用語は修飾語，主要語として用いることができ，「修」の用語は修飾語としてのみ使用可能であることがわかった．例を**図2**に示す．用語辞書では，用語ごとに上記三つのうちのどの分類に属するかを管理することにした．

これらの情報をもとに，標準化支援としての名称チェックの機能を**図3**のように実現している．

a．データ項目名称を投入する．

b．入力されたデータ項目名称を，用語辞書に登録されている用語に分解する．

c．その用語を標準用語に変換し，データ項目の標準名称を生成する．

d．標準用語に設定されている用語種別の並びが，命名規則に従っているかチェックする．具体的には，以下のようなチェックを行う．

①最右端の標準用語の用語種別が区分語の役割を果たす「修主区」であること．

②区分語の左にある標準用語の用語種別が主要語の役割を果たす「修主区」または「修主」であること．

e．従っていない場合，エラーレポートを出力する．

検証の結果，これらの名称チェック機能の一部に問題があることがわかった．次にその問題点と分析結果を示す．

2.3 問　　題

我が社の基幹サービス部門（19 システム）のデータ項目名称に基づき構築した用語辞書（**表2**），および，図1の命名規則を用いて，既存の 11 システム

表 1 用語の役割と用語種別の関係
Table 1 Relationship between word role and word type.

用語種別＼役割	修飾語	主要語	区分語
修主区	○	○	○
修　主	○	○	
修	○		

(a)用語種別が「修主区」の例（番号）

電話_番号　　登録_番号_区分　　番号_更新_日付
↑　↑　　↑　↑　↑　　↑　↑　↑
主要語　区分語　　修飾語　主要語　区分語　　修飾語　主要語　区分語

(b)用語種別が「修主」の例（支店）

支店_コード　　支店_社員_数
↑　↑　　↑　↑　↑
主要語　区分語　　修飾語　主要語　区分語

(c)用語種別が「修」の例（上位）

上位_局_番号
↑　↑　↑
修飾語　主要語　区分語

図 2 用語の役割の変化
Fig. 2 Variation of word role.

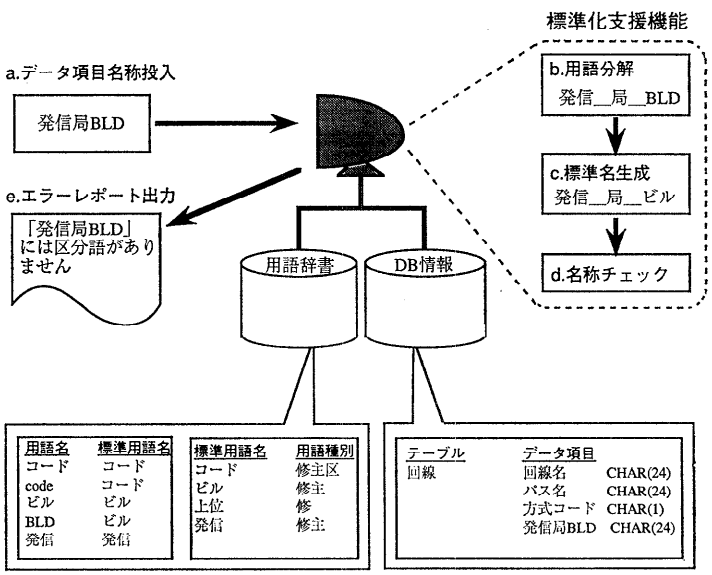


図 3 ツールの標準化支援機能
Fig. 3 Standardization facilities.

9,989 件のデータ項目名称の名称チェックを行った．これによりエラーとなった 3,862 件の名称について，その名称の可否を専門家が判断し，問題があってエラーとなるもの (3,387 件)，問題がないのにエラーとなるもの (475 件) に分類した．さらに，問題がないのにエラーとなった名称について，専門家の判断とツールでの処理結果がくい違った原因の分析と，その分類を行った．その結果，三つに分類できた (**図 4**)．

問題 1　用語辞書の用語区切りが誤っていたためにエラーとなった (2%)

問題 2　用語辞書の用語種別の設定に誤りがあったためにエラーとなった (2.5%)

問題 3　命名規則で許容できないためエラーとなった (7.8%)

表 2　用語辞書の諸元
Table 2　Statistics of thesaurus.

用語数	1,336
標準用語数	1,297
内　訳	
修主区	148
修　主	918
修	231

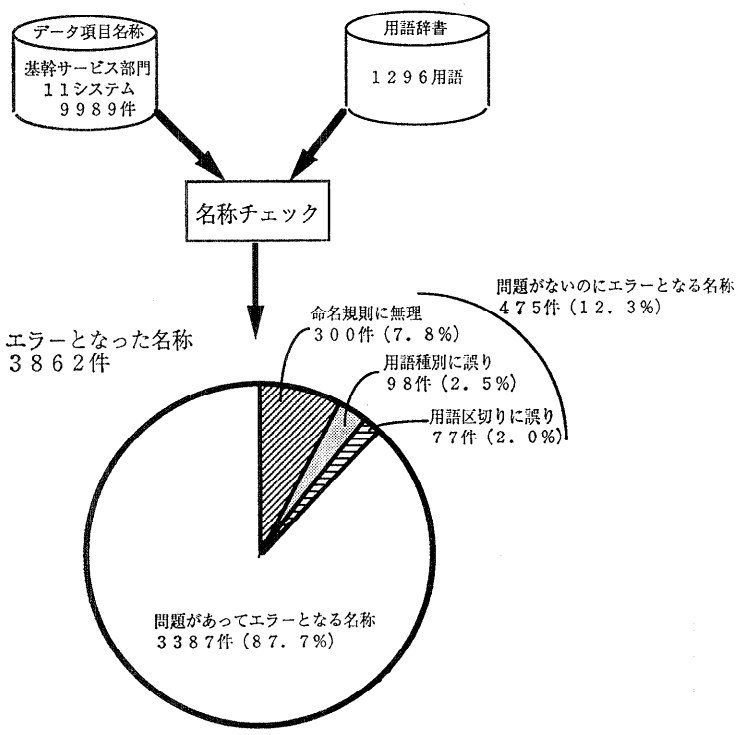


図 4　データ項目名称の分析結果
Fig. 4　Result of data element name analysis.

この分析から，従来の用語辞書，図 1 の命名規則を用いた名称チェックでは，信頼性に問題があることがわかった．その原因は次のように考えられる．

まず問題 1，2 については，用語辞書への登録基準 (用語区切り，用語種別の判定) が明確でないということがあげられる．どのような用語がどのような役割を果たすかということが明らかでないため個人差が生じ，また，同一人物であっても時と場合により考え方が異なる場合があるためである．例えば，問題 1 については，「取引先」という用語が「取引」と「先」に分けられて登録されていたが，「先」という用語だけでは意味が明確でないため，「取引」と接続させて一用語とした方が妥当と考えられる．また，問題 2 についても，「工数」という用語の用語種別が「修主」で登録されていたが，「予約__工数」というデータ項目名称もあるので，この用語の用語種別は「修主区」が妥当と考えられる．

次に問題 3 については，図 1 の命名規則をそのまま日本語の名称に適用するには無理があるということがあげられる．例えば，問題ないと思われる「収容__可能__条数」というデータ項目名称では，それぞれの用語の用語種別が「収容：修主」，「可能：修」，「条数：修主区」であるため，区分語の前に修飾語がありエラーとなる．ここで「可能：修主」とする考え方もあるが，「可能」は管理する対象を表さないので主要語とするのは適切ではない．命名規則に合致するよう「可能__収容__条数」と名称を変更する考え方もあるが，これでは人の解釈として不自然な日本語名称となり意味が曖昧になってしまう．

このような問題を起こさない用語区切り，用語種別の判定基準を作り，社内で使用されるデータに対応できる命名規則を提案することが必要となった．3 章で用語辞書構築の基準を示し，4 章で命名規則の改善案を示す．

3. 用語辞書構築方法

データ項目名称は複数の品詞から構成された複合語であり，その複合語解析の研究については，自然言語の研究開発から種々の成果がでてきており，我が社でもその成果の一つとして，キーワード自動抽出システム INDEXER[7] がある．そこで，2章で述べた，用語区切り，用語種別の判定のために，INDEXER の処理結果を利用することを考えた．

3.1 INDEXER の利用

INDEXER は，大量の文献情報データベースから必要な情報を早く引き出すために必要な，キーワードを自動抽出するシステムとして開発された．この INDEXER が提供するいくつかの機能のうち，複合名詞分割機能を用語区切りと用語種別の判定に用いることを考えた．INDEXER の入出力情報と処理の流れを**図5**に示す．INDEXER にデータ項目名称を入力すると，INDEXER 内部で形態素解析および意味解析が行われる．複合語であるデータ項目名称は，複合名詞解析ルーチンにおいて，語と語の前後関係を解析する係り受け解析法[8]によって，語の区切りおよび語と語の相互関係の同定処理が行われ，その結果，語，カナ読み，品詞種別，意味カテゴリが出力される．ここで，意味カテゴリとは世の中の事象や概念を抽象的概念からより具体的概念へ階層化した分類体系である．INDEXER では，分類語彙表[8]を基本に，それをさらに詳細化した 2,715 種類の意味カテゴリを管理しており，同定処理の結果として各単語ごとに第一候補から第三候補まで三つの意味カテゴリの候補を出力する．この結果を，用語区切り判定法（3.2 節）と用語種別の判定法（3.3 節）に用いることにした．

3.2 用語区切り判定方法

3.2.1 判定アルゴリズム

命名規則で用いる用語は，その一語で，事物，事象，値など，意味のあるものとしている．一方，INDEXER の出力である語には名詞，動詞，形容詞など様々な品詞種別があり，接辞である「最」，「的」のように，他の品詞と結合して意味あるものになる語もある．すなわち，一つ以上の語の組み合わせで用語は成り立っていると考えられる．そこで，このような用語と語の関係を調べるため，まず，2.3 節の分析に用いたシステムのデータ項目名称を，INDEXER で処理して語ごとに分解し，その結果から用語がどのような語から構成されるかを調べた．その結果，

1. データ項目に使用される語の品詞種別は限定される，
2. 一つ以上の語からなる用語の品詞種別組み合わせパターンは限定される

ことがわかった．その構成を**表3**に示す．表3の考え方をフローにまとめたものを**図7**に示す．このフローは以下の考

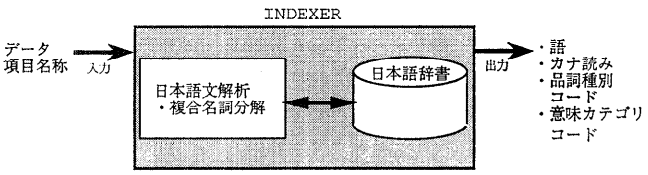


図5 INDEXER の処理概要
Fig. 5 General flow of INDEXER.

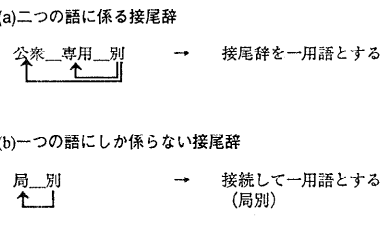
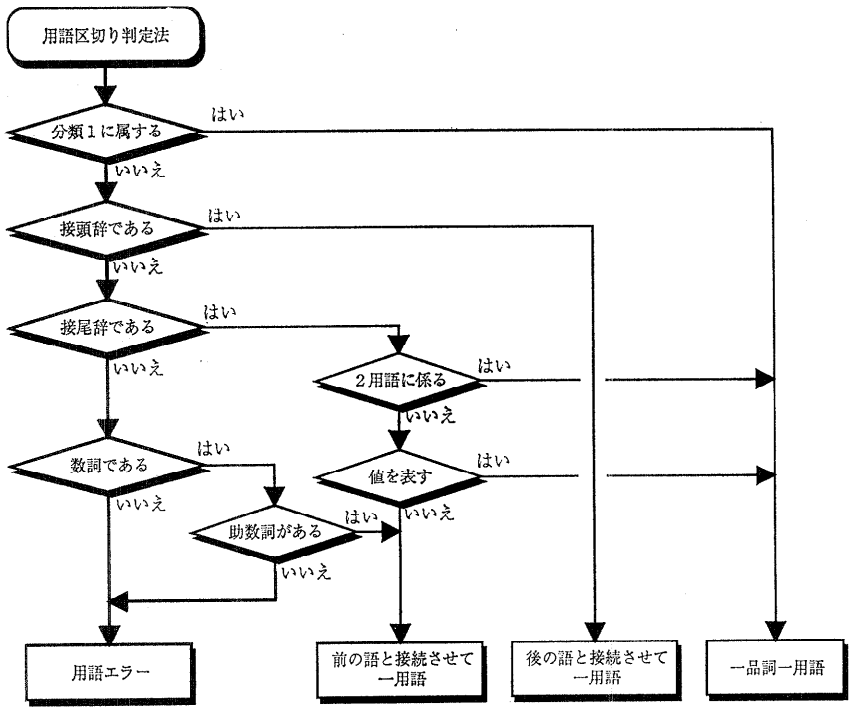


図6 接尾辞の係り受け
Fig. 6 Semantic dependent relationship between words.

表3 用語の品詞構成と用語種別
Table 3 Composition and classification of words.

用語種別 / 用語(品詞構成)	修主区	修主	修
一般名詞	番号，金額	回線，電話	上位，下位
時　詞	月，日	休日	午後
形容動詞型名詞	—	—	固有，可能
サ変動詞型名詞	—	受注，契約	—
動詞転成型名詞	—	受付，繰越	—
他動詞	—	引込，打込	—
副　詞	—	—	以上，随時
固有名詞	—	—	NTT
接尾辞	値，数	—	別
接頭辞＋一般名詞	—	上支線	最下位
接頭辞＋サ変動詞型名詞	—	無応答	
接頭辞＋動詞転成型名詞	—	再割付	—
接頭辞＋形容動詞型名詞	—	—	不完全
一般名詞＋接尾辞	—	事業部	自動的
サ変動詞型名詞＋接尾辞	—	契約者	修正用
他動詞＋接尾辞	—	振込済	引込用
動詞転成型名詞＋接尾辞	—	受付後	繰越用
数詞＋助数詞	—	—	1.5M

(注) — 印は生起しないケースを示す．



分類1：一般名詞，形容動詞型名詞，時詞，副詞，固有名詞
サ変動詞型名詞，他動詞，動詞転成型名詞

図 7 用語区切り判定法のフロー
Fig. 7 General flow of word segmentation.

え方に基づく．INDEXER により，一般名詞，形容動詞型名詞，時詞，副詞，サ変動詞型名詞，固有名詞，他動詞，動詞転生型名詞に分類される語は，用語として意味が明確であると考えられるので，一語で一用語とする．ここで，形容動詞型名詞は「固有，可能」などのように語尾に「だ」や「な」という付属語を接続しても不自然でない品詞種別であり，サ変動詞型名詞は「受注，契約」などのように語尾に「する」という付属語を接続して動詞となる品詞種別であり，動詞転生型名詞とは「受付，繰越」のように語尾に付属語を接続する（「受付る」，「繰越す」）ことにより動詞に変化する品詞種別である．接頭辞に分類される語は，単独では意味が明確でないため，直後の語に接続させて一用語とする．接尾辞に分類される語は，単独では意味が明確でないため，直前の語に接続させて一用語とする．ただし，二つ以上の語にかかる接尾辞（**図 6**）は，二つの語（公衆，専用）が並列に置かれており，一方に接続させると全体の意味が異なるため，独立させることとする．また，接尾辞の中には区分語となるものがあるので，これは独立な用語とする（3.3 節を参照）．数詞と助数詞に分類される語は，それぞれ単独では意味が明確でないため，数詞と助数詞を接続させて一用語とする．なお，判定不可能となる語は，もともとデータ項目名称として使用できない品詞（助動詞，副助詞など）であるので，エラーとしている．

3.2.2 評　　価

語の組み合わせによる用語区切り判定法の有効性を確かめるため，分析に用いたシステムと同分野の別システムのデータ項目名称に対して検証実験を行った．その結果，全 1,283 データ項目名称において，重複を除く 330 データ項目名称から生成された 888 用語のうち，97% は正しく区切ることができた．区切り誤りは 28 用語（約 3%）であった．これらは，ほとんどのものが「D 60」，「M 20」などといった特殊な略号であり，INDEXER で正しく処理できなかったことが原因である．他の誤り原因は，「電源電圧」などのよう

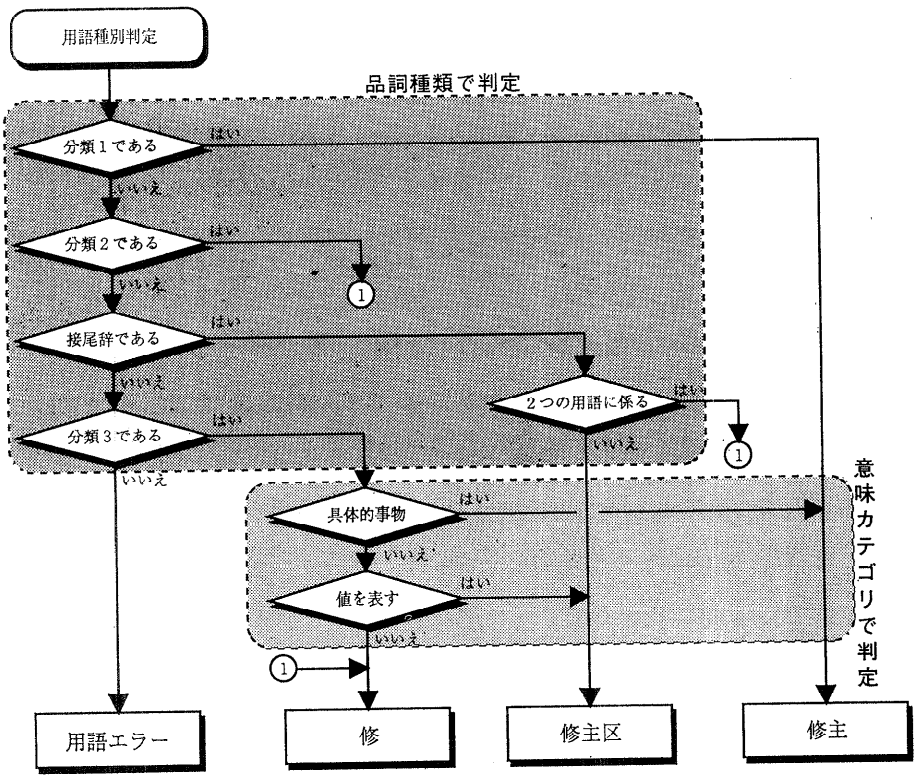
な，複合語が INDEXER の持つ日本語辞書に登録されていたためであった．これら二つの問題は，INDEXER の日本語辞書を改善することにより，容易に解決することができるので，この手法の有効性は確認できた．

3.3 用語種別判定方法

3.3.1 判定アルゴリズム

次に，文法上の役割によって分類したものが品詞種別であることから，用語種別の判定に品詞種別を用いることを検討した．2.3 節の分析に用いたシステムの用語について，品詞種別と用語種別の関係を調べた結果，品詞種別だけで用語種別が判定できるものと，そうでないものがあることがわかった．

一般名詞と時詞，接尾辞を除く他の用語は，品詞種別だけで用語種別が決まることがわかった．品詞種別と用語種別の分類・整理結果を表 3 に示す．この結果を用いた用語種別判定の考え方のフローを**図 8** に示す．まず，INDEXER で分類される語で，サ変動詞型名詞，動詞転生型名詞，他動詞の品詞種別を持つものは，イベントの事象，または，状態を表すと考えられ，これらはデータベースにおいてはデータ管理の対象になるので，用語種別は「修主」とする．また，固有名詞，形容動詞型名詞，副詞，数詞＋助数詞は，他の用語に接続してその意味を限定する働きがあると考えられるが，これらは単独でデータベースにおける管理の対象とならないので，用語種別は「修」とする．また，接尾辞は種類が少ないので総当たりで用語種別を決定した．なお，接頭辞，接尾辞が接続した用語の用語種別は，接続された用語の品詞種別，意味カテゴリを引き継いで判定する．また，例外処理として，接尾辞の「用」，「的」が接続した用語の用語種別は，その接尾辞が接続することにより形容動詞型名詞に変化



分類１：サ変動詞型名詞、動詞転成型名詞、他動詞
分類２：固有名詞、副詞、形容動詞型名詞、数詞＋助数詞
分類３：一般名詞、時詞

図 8 用語種別判定のフロー
Fig. 8 General flow of word type classification.

するため，「修」とする．その他の接頭辞（不，未），接尾辞（化）でも，形容動詞型名詞に変化するが，これらは状態を表す用語となるため，接続された用語の用語種別を引き継ぐこととする．

一方，一般名詞，時詞については，品詞種別だけでは用語種別の判定はできないので，INDEXER が出力する意味カテゴリを用いて，さらに判定を行うことにした．INDEXER が出力する意味カテゴリは 2,715 種類あり，すべての意味カテゴリを 3 種類の用語種別に分類・整理した．意味カテゴリと用語種別の関係の例を**図 9** に示す．データの実体と考えられるカテゴリ（業務，稼業など）は「修主」，値に置き換えることができるカテゴリ（番号，価格など）は「修主区」，抽象的な関係を表すカテゴリ（正，副など）は「修」とした．なお，INDEXER の処理の結果，意味カテゴリとして出力される三つの意味カテゴリのうち，第一候補を用いて用語種別を決定した．

3.3.2 評　価

用語の品詞種別，意味カテゴリを用いた用語種別判定法の有効性を確かめるために，3.2 節と同じシステムのデータ項目名称に対して検証実験を行った．生成された 888 用語のうち，88% は用語種別を正しく判定することができた．誤りは 104 用語（12%）であったが，これらの誤りの中に INDEXER で処理できない文字（「No.」，「I」など）を含むものがあり，これらを除く用語種別の誤りは 51 用語（6%）であった．この誤りは「コード」，「ID」など，すべて意味カテゴリで判定したものであった．その原因は，INDEXER が第一候補としてデータ項目で使用される意味と異なる意味カテゴリを出力したためである．しかし，第三候補までには適切な意味カテゴリが出力されているので，三つの意味カテゴリを考慮して適切なものを選択できれば，用語種別判定の精度は向上すると思われる．よって，意味カテゴリを用いることの有効性が確かめられた．

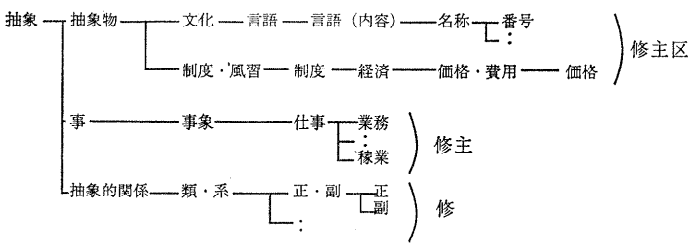


図 9 意味カテゴリと用語種別の関係（抜粋）
Fig. 9 Relationship between semantic categories and word types (extraction).

3.4 用語辞書の検証

3.2 節，3.3 節の用語区切り，用語種別の判定法をもとに現在構築されている用語辞書（表 2）の見直しを行った．その結果，用語区切り，用語種別の設定誤りは計 17% あった．

1. 用語区切りの誤り（1.2%）
2. 用語種別の誤り（15.8%）

2.3 節の問題分析において，9,989 データ項目の名称チェックを行ったが，用語辞書の誤りのためにエラーとなったものはエラーの 4.5% を占めた．これにより，用語辞書の誤りが名称チェックに及ぼす影響は潜在的なものであり，実際の誤りは約 4 倍あることがわかった．これらを修正することにより，2.3 節の問題分析で用語辞書の誤りのため生じた 4.5% のエラーはすべて解消され，用語辞書の品質も向上した．

4. 命名規則の改善

データ項目名称の命名規則は，Durell の命名規則に始まり，それをもとに各社工夫が行われている[6]．しかし，このような命名規則では，実際のシステムで用いられるデータ項目名称に対処できない場合があることが，2.3 節の問題 3 で明らかになった．ここで提案する命名規則は，企業で共通に用いる日本語のデータ項目名称として自然な表現ができるような命名規則でなければならない．

4.1 問題の対処

2 章の問題分析で，複合語の解釈は人間がみてほぼ正しいと思われるが，命名規則に従っていないためツールでエラーとなるデータ項目名称について調べると次のような三つの問題があった．

1. 区分語の前に修飾語がおかれる
2. 区分語の後に括弧付きで単位を表している
3. 構造体の名称チェック

以下，4.1.1〜4.1.3 項においてそれらについての対処法を検討した．また，わかりやすい名称を付与するという観点から，さらに命名規則を追加した．これについては 4.1.4 項に示した．

4.1.1 区分語の前の修飾語

Durell の命名規則に制約を加えたもの[5]では，「主要語＋区分語」という語順が必須となっている．しかし，

データ項目名称において，区分語の前に修飾語が置かれても意味解釈が妥当なものがあった．

区分語の前に修飾語が置かれるものについて，具体的には次のような二つの種類がある．

例1：

区分語の前の修飾語が前の主要語にかかるもの

収容＿可能＿条数
（主）（修）（区）

例2：

区分語の前の修飾語が後ろの区分語にかかるもの

発信＿局＿上位＿番号
（修）（主）（修）（区）

例1のケースは，形容動詞型名詞の修飾語が主要語を修飾しているものである．しかし，他の品詞で後ろから前の語を修飾するような用法はなかった．例2のケースは，一般名詞の修飾語が区分語を修飾しているものである．しかし，他の品詞で，区分語の意味を限定する用法で用いられるものはなかった．よって，区分語の前の修飾語を，たかだか一つ，形容動詞型名詞か一般名詞に限って許すことにした．

4.1.2 区分語の後の括弧付き単位

括弧については，数値を扱うデータ項目に対する名称で，単位が本来のものと異なるものに対して付与すると内容がわかり易くなる．そこで，データ項目名称だけで，データの内容がわかるように，括弧を許す命名規則とした．

例3：補償＿金額＿(万円)
（主）（区）（単位）

なお，単位を表す用語は，区分語を修飾し，その基数となる単位を明確にするものであることから，括弧内で許される用語は「数詞＋助数詞」のものとした．数詞は省略可とする．

4.1.3 構造体の名称チェック

設計の容易さ，文書化の容易さから，階層的な日本語名の表現をする場合がある．このようなものを構造体と呼びその例を以下に示す．この場合，複数のデータ項目名称に共通な修飾語と主要語の組が名称の左に存在すれば，それをくくりだして構造体名としている．そこで，構造体の最上位から最下位までつなげた名称をデータ項目の名称と考えて，命名規則を適用することにした．

規則1： 用語の語順は以下の並びに従うこと。

{修} ＋ 主 ＋[修] ＋ 区 ＋ ＜(単位)＞

ここで、＿は必須を表す。
{ } は0回以上の繰り返しを許す。
[ ] は一般名詞、形容動詞型名詞の修飾語をたかだか一つ許す。
＜ ＞は括弧付きで単位を表すことを許す。
単位は「数詞＋助数詞」で表されるものとする。
ただし、数詞は省略可とする。

規則2： 主要語がサ変動詞型名詞のときは、主要語の前に修飾語を付けること。

規則3： 構造体の名称のときは、最上位から最下位までのデータ項目名称をつなげた名称が規則1に従うこと。

図 10 命名規則の改善案
Fig. 10 Improved naming conventions.

例4：

加入者
企業コード →加入者企業コードと解釈
事業所コード→加入者事業所コードと解釈

4.1.4 サ変動詞型名詞の主要語

主要語がサ変動詞型名詞の場合は，主格あるいは目的格となる一般名詞を前につけるとデータ項目の意味がさらにわかりやすくなるので，主格あるいは目的格となる一般名詞を修飾語としてつけることにした．例5の場合，「継続＿時間」のみであると何の継続時間であるかわからない．そこで，「インパルス」という修飾語を付与することによりデータ項目の意味を明確にする．

例5：継続＿時間
↓
インパルス＿継続＿時間

4.2 命名規則の改善案

上述のような，品詞の係り受け関係を考慮した命名規則の改善案を**図 10** に提案する．

以上のような改善案により，2.3 節の分析で問題3に分類された名称（7.5%）をすべて問題なしとすることができた．

5. おわりに

本稿では，データ標準化のための用語辞書構築方法を示し，命名規則の改善案を提案した．これにより，用語辞書の構築が，用語区切り判定については 97%，用語種別判定については 88% の精度で，容易に行えるようになった．さらに，人の解釈では問題がないにもかかわらずツールでエラーとなるデータ項目名称をなくすことができた．これらの成果により，今後，データ項目名称などに関するデータ標準化作業が進展するものと思われる．

データ項目名称に使用される用語は専門的な用語が非常に多いので，あらかじめ用語を登録しておくことは不可能である．そこで，データ標準化作業途中の用語登録が必要となる．本検討により，専門家でなくても容易に作業途中に用語を登録することが可能になった．また，このアルゴリズムをツールで実現することにより，複合語解析結果から自動的に用語を生成できるので，用語区切りと用語種別設定の工程が自動化できた(**図 11**)．これにより，手投入による工数を削減し用語辞書構築に要する時間を短縮することができ，それによって構築された用語辞書の信頼性も向上した．

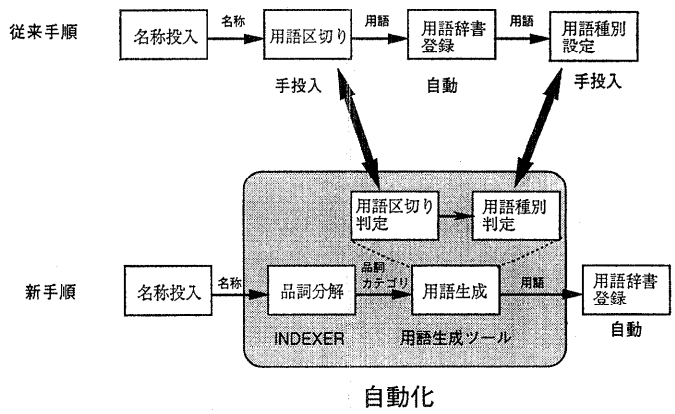


図 11 用語辞書構築の自動化
Fig. 11 Automatic construction of thesaurus.

今後の課題として，用語種別判定法の精度向上のために，INDEXER が出力した三つの意味カテゴリの候補から最適なものを選択する方法を検討する必要がある．

謝辞 本研究を進めるにあたり，INDEXER 使用の承諾および助言をいただいた NTT 情報通信網研究所メッセージシステム研究部の木本晴夫主幹研究員に感謝いたします．また，ご指導いただいた NTT 情報通信網研究所データベース研究部伊土誠一部長，および田中豪主幹研究員に感謝いたします．

参 考 文 献

1) 堀内 一：データ中心システム設計，オーム社 (1987).
2) 山田 進：情報資源管理概説，オーム社 (1987).
3) Newton, J. J.: *Guide on Data Entity Naming Conventions*, NBS Special Publication, 500-149 (1987).
4) Durell, W. R.: *Data Administration*, McGraw-Hill (1985).
5) 関根 純，川下 満，鈴木健司：ネーミング手法と支援ツール，信学技報，DE 89-4 (1989).
6) 中村正弘：先進ユーザがつかんだデータ項目名標準化の糸口，日経コンピュータ，1987. 9. 14 号，pp. 105-113 (1987).
7) 木本晴夫：日本語新聞記事からのキーワード自動抽出と重要度評価，信学論，Vol. J 74-D-I，No. 8, pp. 556-566 (1991).
8) 宮崎正弘：係り受け解析を用いた複合語の自動分割法，情報処理学会論文誌，Vol. 25, No. 6, pp. 970-979 (1984).
9) 国立国語研究所：分類語彙表，秀英出版 (1964).

(平成 4 年 2 月 27 日受付)
(平成 4 年 12 月 10 日採録)

黒川 清（正会員）

1965 年生．1988 年九州工業大学工学部情報工学科卒業．1990 年同大学研究科電気工学専攻博士前期課程修了．同年，NTT 入社．主にデータベース設計支援，情報資源管理の研究実用化に従事．現在，ネットワークオペレーションシステムの開発に従事．

中川 優（正会員）

昭和 22 年生．昭和 45 年大阪大学基礎工学部制御工学科卒業．昭和 47 年同大学院修士課程修了．同年，日本電信電話公社武蔵野通研入所．OS, DBMS の実用化，および，自然言語理解，知識処理の研究に従事．現在，NTT 情報通信網研究所データベース研究部にて，データベース設計法，情報資源管理の研究実用化に従事．主幹研究員．工学博士．人工知能学会，電子情報通信学会各会員．

関根 純（正会員）

1958 年生．1980 年東京大学工学部計数工学科卒業．1982 年同大学院修士課程修了．同年，日本電信電話公社横須賀通信研究所に入社．データベース管理システム，マルチメディアデータベースの研究実用化に従事．現在，データベース設計支援，情報資源管理の研究実用化を行う．情報規格調査会 SC 21/WG 3/RMDM+IRDS サブグループ委員．ACM 会員．