研究部会報告ー抄録

情報知識学会 専門用語研究部会主催 第14回専門用語研究シンポジウム 「多言語用語集に関する諸問題」

Abstracts of the 14th Terminology Symposium "Some Problems on Multilingual Terminology"

山本 昭
愛知大学文学部

第14回専門用語研究シンポジウムが、2001年12月1日（土）、東京、飯田橋レインボービルにて行われた。

"Some Problems in Multilingual Terminology"という副題のとおり、各国における多言語辞書・専門用語にかかわる事例、研究報告が行われた。参加者は約60名であった。

開会にさきだち、中国でのターミノロジー発展に大きく寄与した栗武賓氏の死去が伝えられ、出席者により黙祷が捧げられた。藤原鎮男氏の開会辞に続き、Galinski氏の基調講演から始まり、11件の発表が行われた。発表は、日本語または英語で行われた。以下に発表の概要を記す。なお、日本語で発表が行われたもの以外のタイトルは、山本が和訳した。

Networking of terminology organizations for multilingual terminology -- Networking of terminology networks in cyberspace.（多言語ターミノロジーのためのターミノロジー組織のネットワーク化ーサイバースペースのターミノロジーネットワーク）

Christian Galinski (INFOTERM)

ISO/TC37は、2001年からタイトル変更して、「およびその他の言語資源（and other language resources)」が付け加わった。今後は、TEI(Text Encoding Initiative)、W3C、MPEG等の組織の活動とも関連が強くなるであろう。専門用語は専門家によって使われるのであって言語学者によって使われるのではない。専門家（科学者）の数は増えている。6000学会あってそのうち1／3はターミノロジーを作成している。しかしそれらの多くは外部で利用されない。ネットワーク化が必要である。また、cost-benefit比を上げるためにもネットワーク化は必要である。電子商取引の世界においてもターミノロジーはきわめて重要である。TC37の成果は各分野において広く利用されている。ターミノロジーのユーザの数は多く、かつすべてのユーザは潜在的な作成者(creator)である。以上のような状況を提示して、ISO/TC37のタイトルの「ターミノロジー」に「その他の言語資源」を加える意義や新設されるSC4の可能性を開陳した。

さらに、Romary氏（仏ナンシー大）により、"Knowledge Structure"、"Access Protocol"、"Primary Resources"、"Natural Language Processing"、"Meta Data"、"Lexical Structure"などの、言語資源に関わる諸要素の関連があげられ、それらに対してターミノロジーの方法論でアプローチしようという方向が示された。また、SC4の対象領域(Scope)、戦略(General Strategy)等が紹介された。（発表はいずれも英語で行われた。）

An Examination on Characteristic Words of Disease appeared in Medical Texts（医学テキストに出現する「病名」語の特徴分析）

Haesung Paik, Young-Soo Kang, Hyo-Sik Shin, Hee-Sook Bae, Key-Sun Choi (KAIST)

韓国語の百科事典的テキストから病名に関する項目と、その記述を抽出し、構造的特徴(structural feature)、内容的特徴(content-based feature)、言語学的特徴(linguistic feature)を分析した。構造的特徴は、ID番号<id>、見出し語<title>、本文<contents>、参照<seealso>などの、記述の形態に関するものである。内容的特徴は、本文中に現れる記述の意味に関する特徴である。病名に関する特徴的な「定義」「原因」「症状」「治療」などのタグを人手で付与し、その出現位置(location value)、出現頻度(relative frequency)を計算した。言語学的特徴では、病名の記述に特徴的な文体、述語（動詞)、名詞、名詞－述語の共出現（日本語のサ変動詞に相当)、「接続終了」が分析された。さらに、18の病名について、Feature Cognizerを用いて「原因」「症状」「治療」のテキストについて、重みの計算を行った。「原因」と「症状」のテキストは、文体と述語のみによって判別可能であるのに対し、「治療」のテキストは特定の医薬品名のような名詞表現が重要であった。（発表は英語で行われた。）

Building a platform of automated terminology extraction and analysis based on large-scale true corpus（自動ターミノロジー抽出および大規模コーパスに基づく分析プラットフォーム構築）

YU Xinli, LI Mingfei (CNIS)

ＩＴ分野の特性である急速な更新に対応できるような「自動抽出システム」の構築の基本構想と開発要件が示された。この計画には、ＩＴ分野のコーパス収集と、用語抽出ソフトウェアの開発の２段階がある。コーパス収集では、インターネット上の電子テキストからのコーパスの自動収集を行う。分野、データソース、地域、などの情報を付加する（機械支援により人手で行う)。ソフトウェア開発では、他分野へも応用可能な用語抽出にかかわるＮＬＰ技術、高性能で汎用的な用語抽出システムの実装、用語分類システムを応用した分野の最適化、などの技術が応用される。中国語の用語抽出システムには、大規模な合成的語彙知識ベース、単語分割と品詞判別のための統語論的形態論的ラベリング技術、フレーズ分析、適切な文・パラグラフの特定技術が必要とされる。（発表は英語で行われた。）

魚名をめぐる日中韓三国の共同研究成果「東シナ海・黄海魚名図鑑」

入江隆彦（水産総合研究センター中央水産研究所）

日中韓三国の研究者による共同研究成果、「東シナ海・黄海魚名図鑑」の作成過程が紹介された。ターミノロジーの問題だけでなく、タクソノミー上の不整合を専門家間で解消しようという困難な試みでもあった。魚類分類上の問題点として、国間で、「２種を混同し、１種のみの学名を与えている」「同一種に異なった学名を与えている」といった例が認められた。魚種名の整理・統一に関する研究は以下の手順で行われた。(1)各国の専門家が集まって魚種文献を相互に交換し、不一致点を整理する。(2)対象を漁業重要種にしぼり、日本側からそれぞれの種のカラー写真、和名、学名、形態記載をした例を相手国に送り、相手国でも同じ手法で調査、記載してもらう。(3)不一致の種については、標本の交換を行う一方、実際に相手国にいって標本や生体を調べ、必要に応じて遺伝的手法を用いて種を確定する。日韓、日中韓の研究者で３回の「魚種名整合会議」が行われ、最終的に、対象種約360種について一部を除き不一致が解消された。この成果として「東シナ海・黄海魚名図鑑」（海外漁業協力財団. 東京. 1995.10. 299p.）が発行された。（発表は日本語で行われた。）

Development of spelling checking of Mongolian Language（モンゴル語におけるスペルチェックプログラムの開発）

Nachin SORONZONBOLD (Computer Science and Technical Management School, Mongolia)

モンゴル語の自然言語処理のために必要なことは、モンゴル語の共通ターミノロジーデータベースを作成すること、モンゴル語の音声学的、形態論的、および統語解析のモデルを作成すること、そのモデルを用いて、モンゴル語で書かれたテキストの処理を行うこと、モンゴル語→外国語、外国語→モンゴル語の翻訳ツールを作成すること、である。ここでは、モンゴル語のターミノロジーデータベースが作成された。データベースは原語、翻訳語、語属性、品詞からなる。現在 90000 語が収録されている。用語は、モンゴル語の語の特性から、語幹、接尾辞、前置詞、後置詞に、一定の規則に従って分解する必要がある。また、モンゴル語の文法規則に従い、構文解析を行うため、30 の接頭辞、接尾辞を含む形態論データベースを構築した。これにより、よく使われる語の派生型を発生することができる。これらの規則を実装して、"Spelling Checker of Mongolian language"が開発された。モンゴル語のテキストに対し、スペルチェックと文法チェックが同時に行われる。(発表は英語で行われた。)

Language resource management and special domain language in Korean - Predicates in domain-specific corpus（言語資源のマネージメントおよび韓国語の専門領域言語－特定領域コーパスにおける述語）

Haeseung Paik, Hee-Sook Bae, Key-Sum Choi (KAIST)

韓国語の化学、生物学、物理学の各分野の文献において使用される述語の特性を発見した。各分野における大学学部教科書レベルのコーパスと、一般コーパスを用いた。品詞の分析を行い、述語を取り出した。分野間において、以下の比較を行った。(1)述語の頻度、(2)韓国語には４種の述語形態があるが、各形態の相対頻度、(3)分野に均等に使われる「共通述語」と、分野特有に使われる「独立述語」との比率、(4)独立述語における４形態の相対頻度。さらに、共通述語について、特定分野での使用頻度を得た。分野を特徴づける独立述語、「特徴述語」を用いることにより、文献の分野を決定し、さらにはターミノロジーを用いずに、特定分野の文献を分類する可能性を示唆した。(発表は英語で行われた。)

Corpus Based Word Meaning Analysis of Chinese Ancient Poetry （コーパスに基づいた漢詩の語義分析）

Hu Junfeng, Yu Shiwen （北京大学）

漢詩は独自の表現を持ち、数多くの比喩が用いられるため、語彙を知ることが困難である。600 年から 1400 年前の漢詩の 620 万文字について、統計的手法によって語の特定と、類似機能を持つ語の抽出を行った。それにより、ある語が異なった意味に用いられているかどうかを自動的に判別できた。唐代と宋代とに分け、共出現と対句を計数した。二語の共出現回数をそれぞれの語の持つ全共出現数の対数の積で除したものを内的関連度(Intimate Relation Degree)とした。この内的関連度が一定の閾値を超えるものを「内的関連語」とした。さらに、３語からなる内的関連語組を抽出し、ダイアグラムを作成した。(発表は英語で行われた。)

A contrastive analysis, and related problems, of information contents in Japanese and French abstracts of chemistry

「科学者による要約が同一内容を含むかどうか」日本語・フランス語コーパス研究から見た、多言語情報の分析などにおける諸問題

Sophie Palvadeau（琉球大学）

日本語およびフランス語で書かれた化学論文の著者抄録において、内容の比較を行った。「日本化学会誌」、"Comptes rendus de l'Academie des sciences-chimie"、"Lettre du Departement des Sciences Chimiques du CRNS"の3誌から、それぞれ104、88、48の抄録を用いた。Introduction Method Results Discussionの4部分に分割、さらに文の役割により、「問題意識に関する研究目的の記述」など9に細分し、内容の記述単位(informative module)がどこに存在するかをタグ付けした。記述の精細度、モダリティ、の情報も付加した。フランス語の抄録においては、2－3のinformative moduleが含まれるものが多いのに対し、日本語では3－4含まれるものが多かった。3つのinformative moduleを含む抄録においては、3つの構成は共通のものが少なかった。コーパス内で、抄録ごとに含まれる要素が異なるものが多いことが示唆された。(発表は英語で行われた。和文タイトルは発表者による。)

データベース統合検索用大規模辞書(JST)の監修経験から

笹森勝之助、内田尚子（当時、日本データベース開発）

異なる情報源を統合的に検索するためのツールとして、「データベース統合検索用大規模辞書」が開発された。その監修行程が紹介された。「JICST科学技術文献データベース」の表題中の語を基本とし、日本語－英語の対応辞書とする。見出し語と関係語（同義語、上位語、関連語）からなるワードブロックを処理単位とした。編集工程では、語の抜き出し、日英ペア作成、関連語付与、チェックが行われた。監修（校閲）工程では、第一次校閲作業として、24の各分野ごとに、各ワードブロック内での見出し語候補、関連語候補の検査、修正、第二次校閲作業として、9の分野グループ内でのチェック、分野にまたがる同義語の調整、さらに、総合バッティング修正を行い、整形、納品工程を経て完成した。29.5万語の概念辞書が作成された。仕様変更に対応できる工程管理の重要性、校閲者の要件として、専門知識、パソコン技術のほか、納期遵守、品質保証意識、柔軟性、協調性が指摘された。(発表は日本語で行われた。)

漢字字典編集の試み／漢和韓越字典瀏覧の編集

伊藤全

元獣医である発表者が、引退後、独力で編纂した「漢和韓越字典瀏覧（かんわかんえつじてんりゅうらん)」と、その編纂手順が紹介された。日本語、中国語、韓国語、ベトナム語において使用される、または語源となった漢字に着目し、それぞれの言語における使用法が一覧できるように編集されている。「しろうと仕事」と発表者が言うように、個人の興味と、独自に発案された方法論によって編纂された。「字統」(白川静、東京、平凡社1984)を中心に、日本と韓国の常用漢字を網羅するように3600字を選択した。それぞれの文字について、漢字一文字を見出しとし、4カ国語の、音、字体、成語、意味に関する情報を付け加えて作成された。配列は和音の五十音順だが、各種の索引も作られる。さらに各国語の微妙な用法の変化から、エチモロジー的興味を抱かせるようにできている。

なお、発表者は、この字典の編纂に関して、助力や、後継する人を求めている。(発表は日本語で行われた。)

ISO/TC37会議報告

高野文雄（科学技術振興事業団）

2001年8月にトロントで開催された表記会議の概要が報告された。ISO/TC37は、ISO（国際標準化機構）の技術委員会(ISO/TCs: Technical Committees)の一つであり、用語に関する原則、手法および応用の標準化を推進している。SC1からSC3の3つのSubcommitteeからなっている。Computer Applicationsを担当するSC3では、ISO/CD16642、TMF(Terminological markup

framework)が扱われた。SGMLやXML等の特定の文書記述言語に基づいて蓄積された用語データ集 TML(Terminological Markup Language)を XML ドキュメントとして記述するためのフレームワークを定義するものである。年来対立してきたフランス案のGeneter、米案の MSC は、ともに TMF の "normative annex"として含まれることになった。2001 年からは、TC37 の対象領域が "Terminology - Principle and Method" から "Terminology and other language resources"へ拡大され、それに呼応する形で SC4 "Language Resource Management"が新設された。(発表は日本語で行われた)

事務局から：シンポジウムの予稿集は、残部が多少あります。(1部 4,000 円、送料事務局負担、振込手数料はご負担ください。) 希望の方は、下記まで E-mail または Fax でお申し込みください。

〒111-0051 東京都台東区蔵前 3-1-10
蔵前セントラルビル
(株) システムソフト・ジスト事業本部
長田孝治
HHG01656@nifty.ne.jp
Tel.03-5821-2567　Fax.03-5821-2539

[振込先] あさひ銀行　市ヶ谷支店（２７０）
普通　１５７７０８５
口座名：専門用語部会
代表　長田孝治