パリ・ソルボンヌ大学　理論・応用言語学研究所（ＣＥＬＴＡ）

―ＣＡＳＫ（Computer-aided Acquisition of Semantic Knowledge）プロジェクト―

アンドレ・ヴロダルチック

パリ・ソルボンヌ大学　理論・応用言語学研究所（ＣＥＬＴＡ）

andre.wlodarczyk@paris-sorbonne.fr

1．キーワード：

(1) 言語理論（言語モデリング，記号，意味分野），(2) ＫＤＤ：データベースにおける知識発見（決定論理，形式概念分析，ラフ集合理論，クラスタ分析，要素分析），(3) ＤＢＭＳ：データベース管理システム（ソフトウェア工学），電子辞書

１．はじめに

ＣＡＳＫ（Computer-aided Acquisition of Semantic Knowledge コンピュータを活用した意味知識獲得）は，パリ・ソルボンヌ大学の理論・応用言語学研究所（ＣＥＬＴＡ： Centre for Theoretical and Applied Linguistics, http://www.celta.paris-sorbonne.fr/）が行っている研究プログラムである。理論・応用言語学研究所では，66名の研究者，及びほぼ同数の博士課程の大学院生が，ヨーロッパ言語10ヶ国語と日本語に関する研究を行っている。同研究所は，スラヴ言語学のエレーヌ・ヴロダルチック教授によって,“Forme-Discours-Cognition”の名の下に2000年に設立され，2002年にフランスの教育研究省に正式に認可された。同研究所は六つの研究チームに分かれており，ＣＡＳＫは筆者の率いる第３チーム（意味分析）による研究プログラムである。

ＣＡＳＫ（コンピュータを活用した意味知識獲得）は，ＫＤＤ（Knowledge Discovery in Databases データベースにおける知識発見）の技術を利用して複数のヨーロッパ言語の様々な意味分野を記述することを目的としている。そのため,このプロジェクトは本質的に学際的であり，理論言語学と情報技術という異なる分野の熟練の専門家たちによる研究協力を前提としている。そこでの言語学者の仕事は，ＳＥＭＡＮＡ（Semantic Analyser）というソフトウェアを用いて，素性構造のオントロジカルな階層定義を相互作用的に発見することである。ＳＥＭＡＮＡは，意味知識の言語資源データベースを構築すべく，ＣＡＳＫプロジェクトのために特別に設計されたソフトウェアである。ＳＥＭＡＮＡは複数のＫＤＤアルゴリズムを統合したプラットフォームの形をとっており，現在ジョルジュ・ソーヴェ（Georges Sauvet）とアンドレ・ヴロダルチック（André Wlodarczyk）がTranscript[1] による設計・ インプリメンテーションを行っている。

ＣＡＳＫプロジェクトは，記述の関連性と相対的な重要性を決定するためにコンピュータ計算（近似値に基づいたものも含めて）を利用した初の試みとなる。意味概念の言語横断的な定義を，立証された（すなわち，実験的に検証された）比較可能な形で実現する上では，複数の異なる言語の詳細な形式的記述をデータベースに蓄積することが欠かせないのである。

２．研究の背景：言語学とオントロジー

言語現象の意味分析の分野においてよりよい成果を得るためには，現在用いられている基礎的概念のいくつかを形式的に再構築する必要がある。言語学的な（より一般的には記号論的な）観点からは，意味概念（内容）は記号から切り離して考察してはならない（「形式」と「内容」のペアとしてもともと定義されている単位がある)。それゆえ，我々が現在とっているアプローチは，人によって造られた記号の意味はそれ自体では検査不可能であり，唯一の理にかなった意味研究の方法はモデリングだという前提に立っている。

加えて，記号は，オントロジーを基礎とする意味的な対象である。オントロジーは，記号の意味特性のモティヴェーション（階層構造を持つ基盤）と見られる。人間言語における意味は，適用やドメインに応じて固有である（すなわち，特に局所的なドメインに対応することができる)。さらには，言語単位（記号）は，その特性を複数のオントロジーから受け継いでいる。例えば，動詞はその特性を，音素構造，結合価スキーマ，役割，状況フレームなどから同時に受け継ぎ得る。しかしそれでも，特定の意味解釈を導くメタオントロジカルな（普遍的な）概念の階層性を構築することは可能と思われる。

人類の歴史において，語彙目録や辞書は，言語資源を注釈や翻訳の目的で利用しようとした最初の試みである。中でもシソーラスは，最も構造化された語彙集成である。しかしながら，記号は本質的に多義的なので，シソーラスがとらえ得る記号間の関係も大まかなものにとどまっている。このことからも，以下に示すような動的な意味のマップやラティスは，研究における記述の段階においても，将来コンピュータ化された辞書を活用する際にも，非常に有用と考えられる。

●**セマンティック・マップ（Ｓ－マップ）**：類似関係によって配列された記述子のついた，（用法タイプに関して）類似の記号の集合

●**セマンティック・ラティス（Ｓ－ラティス）**：含意関係によって配列された記述子のついた，（用法タイプに関して）類似の記号の集合

上述の「記号」やその「体系」は，一価のブール属性の表である形式概念文脈(formal concept contexts[2])でも，多価の属性の表である情報システム(information systems[3])でも，表すことができる。

３．計算ツール：ＳＥＭＡＮＡプラットフォーム

近年，コンピュータの利用により，言語学はますます「実験科学」の様相を増してきた。大量の用例をデータベースに蓄積し，それらのデータを記号処理により，そして統計的なＫＤＤの手法で記述・分析するというやり方は，限られた数の例示に基づく方法論の「仮説演繹的説明法」に重きを置くタイプの言語学とは明らかに対極をなしている。

しかしながら，意味分析のためのデータの入力が大変な作業であることは強調しておかなけれ

ばならない。言語データを集め，注釈を付ける作業段階では，言語研究者の直観力（言語話者としての能力に,その言語に関する学術的な知識が加わったものに基づく直観力)は不可欠である。しかし，ＳＥＭＡＮＡは動的な性格を持つものであり，人とコンピュータの相互作用によって明示的に定義された属性のリストの作成や利用が行われる。それらのリストは，変更修正も容易なものになっている。これによって，異なる使用コンテクストにおける表現の意味について人間の判断が主観的になったりゆれたりすることの影響を防ぐことができる。

しかしその一方で，データ入力が難しいことの原因はまた別のところにも求められる。文脈における言語表現は非明示的な言外の意味も併せ持っており，前提的な知識と推論で導かれる知識の両方を内含している。それら２種類の知識のどの部分を記述において考慮に入れるべきかを確定するのが難しいのである。多くの場合，非明示的意味のどの部分を明示化すべきかは，対照の準拠枠となる言語がどのようなものかによって決まってくる。ある言語を複数の他言語と対照することで，それら諸言語それぞれの表現単位の意味内容にについて，より詳細な記述が可能になると考えられる。

データベース技術における知識発見の原理は，関連文献においては以下のように挙げられている。

・**タスク**（視覚化，分類，クラスタリング，回帰，など）
・**モデルの構造**とデータの適合（比較や検証の範囲を決定する）
・**評価機能**（適切性／対応関係や一般化の問題）
・**検索あるいは最適化の方法**（データ探索アルゴリズムの中心部）
・**データ管理技法**（データの蓄積と索引付与のツール）

ＳＥＭＡＮＡ(Semantic Analyserの略称)ソフトウェアには,動的なデータベース構築機能と,言語の意味研究のためにコンピュータを活用してオントロジーを探求すべく設計されたプラットフォームが含まれている。言語研究者は，自分たちの研究対象が途方もなく複雑なものであることをよく分かっている。しかし,ここで強調したいのは,関係の複雑さを反映するからといって,データ構造までが複雑に見えてはいけないということである。以下の図表に示されるように,（木による表現よりもさらに強い力を有する）ラティス表現を用いることで，単純な表の表現による記述を集めたものでは見えない（「隠れている」）ような複合的な関係も，計算によって明らかにすることができる。

表１　表の形で示した属性付与

	空中	ゆっくり	転置	速く	徒歩	地表
「飛ぶ」	○		○			
「歩く」		○	○		○	○
「走る」			○	○		○
「行く」			○			

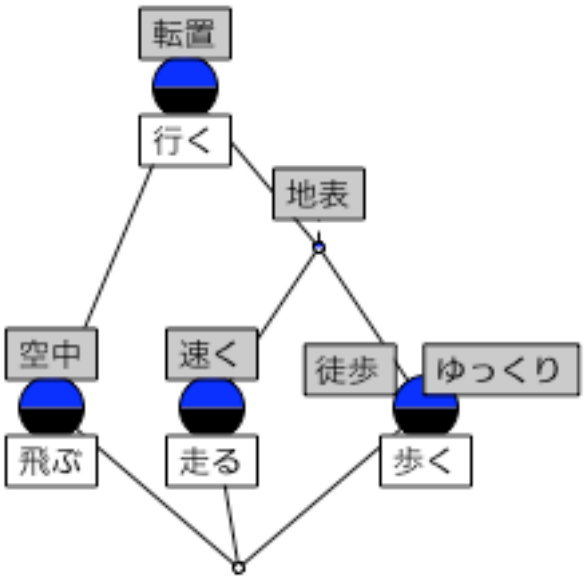

図１　ラティス（Ｓ－ラティス）の形で示した属性間の含意関係

表１は，日本語の四つの動詞を６個の属性だけで意味記述したものであるが，データの見方を変えて図１のＳ－ラティスのようにできると，関係する列や行の数が多くなった場合に特に有効である。

3.1. ＫＤＤツール付きの動的なデータベース管理システム

記号のように多様な異質の成分を含む対象を研究するためには，極めてフレキシブルなシステム環境でデータを収集する必要がある。我々の“db Builder”（Database Builderの略称）は，アプリオリに構造化された知識の少ない言語データの研究を目的として特別に設計されたものであり，意味知識の獲得や実験に適している。“db Builder”の機能として，以下のことが挙げられる。

(1) 研究対象とする記号のサンプルを含む発話例を，文脈環境，他言語への翻訳，自然言語での（定型化されていない）自由記述付きで蓄積する。

(2) 属性と価（パラメータ表示した素性）を用いて意味を記述する。

記号の用例データに用いられる属性のセットは不定である。しかし，一つのカテゴリーを記述する属性の数は有限と考えられる。そこでの言語研究者の仕事は，所定の意味ドメイン（フィールド）に関して属性の配置・構成を安定化させることであり，すべての属性は，いわゆる「素性構造」を構成するオントロジカルな階層の形で定義されねばならない。

発見手順は，以下のようになる。

１．ある言語記号（または表現）の用例を大量に収集し，（必要に応じて）オントロジーを基礎とする属性と価による記述を加えて情報システムを構築する。

２．データベースを必要な数の情報システムに自動分割する。

３．各々の情報システムに含まれる知識を縮小・安定化させる。

４．確定した情報システムを合併させて，一つの巨大な形式概念にする。

このようにして得られた構造は，言語単位の意味構造記述である。属性スペースにおける記述の実験を可能にするＫＤＤの専門機能には様々なものがあるが，中でも特に有用なのが，記号間の関係を明らかにする上述の二つの機能である。形式概念（eーディクショナリー）の集合からＳーマップやＳーラティスを構築する手順の自動化については，現在研究が進められている。

3.2. ＳＥＭＡＮＡプラットフォームのアーキテクチャの概要

ＳＥＭＡＮＡプラットフォームは，(1) データベースの作成と動的な維持，(2) 記号的・統計的なデータ分析のためのＳＥＭＡＮＡ固有のアルゴリズムという２種類のオペレーションから成る。

(1) Data Base Builder

動的にデータを再構成できるデータベース構築環境

- Editor of Records
- Tree Builder Assistant
- Attribute Editor

(2) SEMANA Editor

ＳＥＭＡＮＡのモニターであり，ファイルを開く，作成する，編集するなどに加え，意味分野などの構築に役立つ類似や類推を発見することもできる。

a) Symbolical Data Analysers

- Formal Concept Analyser - FCA （cf. Wille, R. 1982, 1997; Ganter, B. & Wille, R. 1999）
- Rough Set Analyser - RSA （cf. Pawlak, Z. 1982）
- Formal Rough Concept Analyser - FRCA （cf. Saquer, J. & Deogun, J.S. 1999）
- Rough Decision Logic Analyser - RDLA （cf. Bolc, L., Cytowski,J. & Stacewicz, P. 1996）

b) Statistical Data Analysers STA3

- Factor Correspondence Analysis
- Ascending Cluster Analysis

ＳＥＭＡＮＡ Vers.Ⅱのデータベースは HTML と XML でフォーマットされている。

４．日本語を「対照の軸（ピボット）の言語」としたヨーロッパ言語の研究

現在，理論・応用言語学研究所（ＣＥＬＴＡ）では，ＣＡＳＫプロジェクトの枠組みにおいて，ＳＥＭＡＮＡプラットフォームを用いたヨーロッパ諸言語の研究が行われている。ＣＡＳＫプロジェクトのメンバーである言語研究者たちは，研究の第一期（アスペクト，モダリティ，移動）のために選ばれた専門家であり，それらのテーマについてのモノグラフや論文，博士論文を執筆している。

アスペクトは，文法と語彙の境界をまたぐカテゴリーであり，動詞が表現する「状況の意味的

タイプ」[4]だけでなく，様々な文法的・語彙的手段によって表わされ得る「状況の意味的タイプ」にも関わっている。現在，スラヴ言語（ポーランド語とロシア語）のアスペクトの研究[5]がフランス語，英語，ドイツ語との対照において進められている。これによって，２種類の異なるタイプの言語におけるアスペクトの文法的・語彙的表現方法の比較が可能になっている。すなわち，すべてのスラヴ言語のようにかなり複雑な文法的動詞アスペクト（しかし，より非明示的な名詞判別システム）を持つ言語と，動詞アスペクトはそれほど複雑ではないが，より複雑な名詞判別システムを持つ言語（フランス語，英語，ドイツ語のように冠詞を持つ言語）との比較である。

研究の現段階では，我々が蓄積している素性構造は発話における様々なアスペクト用法を記述する上でまだ網羅的とは言えないが，ＳＥＭＡＮＡを用いて記述の一貫性を検証しつつ，アスペクト理論を精密化しているところである。

数多くの異なる言語の情報を含むＳＥＭＡＮＡデータベースは，アスペクトの定義の一般化を可能にすることであろう。より多くの言語からのデータを得ることで，アスペクトのオントロジー構造の木に新たな属性が加えられ，場合によってはその再構成につながる可能性もある。このことは，我々の実際のアスペクトのメタオントロジーについて確証を得ること，そして異なる複数言語に対する単一の記述を可能にすることにもつながる。この点において，日本語のアスペクト用法に関する新たなデータを得ることは歓迎すべきことである。既に，研究の最初の成果は，伝統的な，しかし明確に定義されていなかったアスペクト概念（Aktionsartと呼ばれるもの）の形式化に寄与している。この概念は，異なるヨーロッパ言語の記述において異なる意味で用いられてきたものである。同様の研究が，モダリティについて，ポーランド語，フランス語，ロシア語の間で行われている。モダリティのオントロジカルな木について，最初の大まかな図は既に得られているが，さらにデータを増やして検証・修正が行われる必要がある。

ＣＡＳＫプロジェクトが前提とするのは，多言語間対照というアプローチによって，他の言語との比較から素性を補充したり修正したりしつつ，一つの言語の意味記述を深めていける，という考え方である。同時に，実際の言語データから生まれるオントロジーの構築にも対照研究的アプローチは適している。対照に基づく記述の有効性は，既に異なるタイプのヨーロッパ言語について実証済みであるが，類型論的により遠い日本語のような言語とヨーロッパ言語の対照を行うことで，この方法の効果や重要性がより明らかに示されることであろう。日本語のデータは既に数々の日本の研究機関で利用可能になっており，それらが「対照の軸」としてヨーロッパ言語の研究に用いられることになる。特に，日本語の電子辞書が利用される予定である。この点に関しては，日英対照によって日本語の語彙素の記述をより深い，広がりを持ったものにした鳥取大学の池原教授の研究室による研究が，対照研究的アプローチの成功例と言えよう。

５．国際協力と今後の展望

ＣＡＳＫプロジェクトは，双務的な研究協力を基盤として，国際的に進められている。フランス語とポーランド語の二言語プロジェクトが現在進行中である。ポーランド語チームのメンバーは，ワルシャワ大学（ワルシャワ），ヤゲロンスキ大学（クラカウ），スラスキ大学に，フランス

語チームのメンバーは，パリ・ソルボンヌ大学（パリ第４大学)，シャルル・ド・ゴール大学（リール第３大学)，エクサンプロバンス大学に，それぞれ所属している。研究協力は，2004年と2005年にポーランドで開かれた２回の予備会議を通して準備された。公式には，ＣＡＳＫ二言語プロジェクトの研究期間は2006年１月から2007年12月までである。2006年には，５月にクラカウで，そして９月にパリで，計２回の会議が開かれた。2006年12月22日には，東京の早稲田大学で原田康也教授を座長に，第３回ＣＡＳＫワークショップが開かれた（主催：早稲田大学総合研究機構情報教育研究所，共催：早稲田大学総合研究機構ことばの科学研究所，後援：国立国語研究所)。次回の会議は，2007年４月にクラカウで，そして９月にパリで行われることになっている。今後，ＣＡＳＫプロジェクトの展開としては，スペイン語，ドイツ語，ロシア語の国際共同研究を統合する予定である。

ＣＡＳＫプロジェクトの中心的な構想は，いくつかのヨーロッパ言語の主要な言語意味分野について日本語との対照をもとにオントロジカルな研究を行い，それによって多言語に共通な意味素性構造の蓄積を作り上げることである。日本では，理論・応用言語学研究所の代表と，以下に挙げる諸研究者との間で，準備的な関係づくりがなされている ― 相澤正夫（国立国語研究所部門長)，荒川直哉（(株）ジャストシステム研究員)， 原田康也（早稲田大学教授)，池原悟（鳥取大学教授)、井佐原均（けいはんな情報通信融合研究センター自然言語グループリーダー)、柏野和佳子（国立国語研究所研究員)，神崎享子（けいはんな情報通信融合研究センター研究員)，黒田航（けいはんな情報通信融合研究センター上席研究員)，村上祐子（国立情報学研究所特任助教授)，アントニオ・ルイズ・ティノコ（上智大学教授)，横井俊夫（東京工科大学教授)，横山晶一（山形大学教授)，吉本啓（東北大学教授）（敬称略，アルファベット順)。

注

１　Transcriptは，Apple社のHypertalkをもとにしたオブジェクト指向のプログラミング言語である。

２　Wille, R. (1982, 2001), Ganter, B. & Wille, R. (1999)

３　Pawlak, Z. (1981), Orłowska, E. & Pawlak, Z. (1984)

４　Wlodarczyk, A. (2003)

５　Wlodarczyk, A. & Wlodarczyk, H.. (2003, 2006)

参考文献

(A) 言語理論

Hartshorne Ch. and Weiss P. (eds.) (1934) *Collected Papers of Charles Sanders Peirce*, *Volume 5 : Pragmatism and Pragmaticism*, Cambridge, MA. : Harvard University Press.

Pogonowski, J. (1993) *Linguistic Oppositions*, Wyd. Naukowe UAM, Seria Językozawstwo Nr 17, Poznań, 1-136.

Putnam, H. (1975) The Meaning of 'Meaning'. Gunderson, K (ed.) *Language, Mind and Knowledge*, Minnesota Studies in the Philosophy of Science, vol. 1, Minneapolis: University of Minnesota Press, 358-398.

Wlodarczyk, A. & Wlodarczyk, H. (2003) Les paramètres aspectuels des situations sémantiques. *Etudes Cognitives / Studia Kognitywne* V, Warszawa: SOW Publishing House, 11-34.

Wlodarczyk, A. & Wlodarczyk, H. (2006) Semantic Structures of Aspect (A Cognitive Approach). *Od Fonemu do Tekstu, in honour of Roman Laskowski*, Krakow : Lexis Pub. Co., 389-408.

Wlodarczyk, A. (2003) Les Cadres des situations sémantiques. *Études Cognitives / Studia Kognitywne* V, Warszawa: SOW Publishing House, 35-51.

(B) KDD – データベースにおける知識発見

Bolc, L., Cytowski, J. & Stacewicz, P. (1996) *O Logice i Wnioskowaniu Przybliżonym (On Logic and Rough Reasoning).* Institute of Computer Science, Polish Academy of Sciences, ICS PAS Report 822 (in Polish), 1-54.

Ganter, B. & Wille, R. (1999) *Formal concept analysis: Mathematical foundations*, Berlin: Springer.

Orłowska, E. & Pawlak, Z. (1984) Logical Foundations of Knowledge Representation. IPI-PAN, ICS PAS Report 537, Warszawa, 1-106.

Pawlak, Z. (1982) Rough Sets. *International Journal of Information and Computer Sciences*, Vol. 11, No. 5, 341-356.

Pawlak, Z. (1992) *Rough Sets: Theoretical Aspects of Reasoning About Data.* Dordrecht: Kluwer Academic Publications.

Saquer, J. & Deogun, J. S. (1999) Formal Rough Concept Analysis. Zhong, N., Skowron, A., & Ohsuga, S (eds.) *Lecture Notes in Computer Science*, Berlin/ Heidelberg：Springer-Verlag, 91-99.

Wille, R. (1982) Restructuring Lattice Theory: an Approach based on hierarchies of concepts. I. Rival (ed.) *Ordered Sets*, Dordrecht-Boston: D. Reidel, 445-470.

Wille, R. (2001) Why Can Concept Lattices Support Knowledge Discovery in Databases ? Mephu, E. N. et al. (eds.) *ICCS 2001 International Workshop on Concept Lattice-based Theory, Methods and Tools for Knowledge Discovery in Databases*. Palo Alto, CA: Stanford University, 7-20.