

# 自然言語テキストの共通的概念記述

## Common Concept Description for Natural Language Texts

石塚 満

Mitsuru Ishizuka

東京大学 情報理工学系研究科 創造情報学専攻/電子情報学専攻

School of Information Science and Technology, University of Tokyo

ishizuka@i.u-tokyo.ac.jp, <http://www.miv.t.u-tokyo.ac.jp/ishizuka/>

**Keywords:** natural language, concept description, knowledge representation, semantics

### 1. まえがき

自然言語は概念や意味を表現し伝達する人間に特有の柔軟で最も自然な、かつ最も重要なメディアと言える。拡大し進化を続ける Web 情報空間においてコンピュータ、あるいはエージェントにもその意味（セマンティクス）を理解できるようにし、人間と情報/知識の共用ができるようにすることは、次世代 Web 基盤構築に向けての重要な観点である。コンピュータによる自然言語意味理解は大きな目標であり、着実に進歩しているものの、人間による意味理解レベルはもとより、必要とされる理解レベルとのギャップは依然として解消されない。この成熟を待つとは別に、自然言語テキストの意味を十分近似的に表現するのに適切なレベルを設定し、共通性のある語彙と表現形式を定め、これを介して人間とコンピュータ間で情報/知識の共有、操作を可能にすることが重要となる。

論理も元来、自然言語表現に対応する概念を記号化して表現する（そして推論する）目的で生まれ、発展してきた経緯があるが、述語論理を例にとると、何を述語とし、何をどのように引数にとるかという任意性があり、この面においても（そのままでは）共通性のある表現には不十分である。1990年代からは、自然言語テキストを知識として利用する試みが行われ、自然言語表現に対応する幾つかの知識表現法が発表されてきている [Iwanska00a]。しかし、語彙や記述に必要な関係の整備が不十分で表現の対象範囲は限定的であり、十分でない。

これらに対し、国連大学研究所において開発された UNL (Universal Networking Language) [Uchida05, UNDL] は、単に表現形式を定めただけでなく、実用規模の基本語彙オントロジー（関係を表す語彙も含む）も併せて提供されている。この UNL は、エンコンバータ/デコンバータを介して、各国語（英語、日本語、中国語、スペイン語、フランス語など）との間の変換

が可能であり、多国語 Web を相互変換する中間言語の役割を果たし、国際的に一定の支持を得ている。

CDL.nl (Concept Description Language -- natural language version) は、次世代 Web 時代に向け、自然言語だけでなく他のメディアの意味内容も含めて汎用的に表現できるように設定された CDL の枠組み [Yokoi 05, 横井 06] に基づき、UNL の成果を活用して設計された、Web を中心とする自然言語テキストが表す概念を対象とする共通性のある記述言語である。簡素な 3 つ組表現を基礎としているが、自然言語表現をカバーする実用レベルの基本語彙オントロジーを備えている。自然言語テキスト自体のレベルを表層的表現、人間が理解・解釈するレベルを深層意味とすると、CDL.nl は深層意味までは達しないものの、自然言語テキストの意味する概念を表現するレベルと位置づけられる概念記述言語である。

セマンティック Web が Web コンテンツに関するメタデータをコンピュータにも理解できる共通形式 (RDF, RDFS, OWL, .....等) で記述し、コンピュータも人間と共同で働く次世代 Web 基盤を築こうとしているのに対し、CDL.nl は主要ソースデータである自然言語テキストの概念記述まで踏み込んで、この共働の地平を格段に拡大しようとするものである。ただし、セマンティック Web と競合するものではなく、セマンティック Web と融合する形で設計されている。セマンティック Web 基盤作りは米欧中心で進められており、我が国はそれを利用する研究開発は行われているものの、基盤構築への貢献は皆無に近い状態であるが、CDL.nl は今後の情報化社会で重要性を増す Web の新基盤構築に我が国発の技術により国際的貢献を果たそうとするものでもあり、W3C での国際標準化活動も開始している。

本稿は、以上のような自然言語テキストの共通的概念記述に関し、CDL.nl の研究開発と関連する研究開発について紹介する。

## 2. 自然言語を知識表現として利用 する幾つかの試み

豊富に存在する自然言語文をそのまま知識として利用でき、人間が行うように複数の断片的知識から結論を導く推論操作を行えることは一つの大きな目標である。知識表現という場合、演繹のような推論機能が付随することを重視する一方の立場があり、他方に共通性のある幅広い意味表現を重視し、推論機能は二次的と考える立場がある[Iwanska 00a]。3章、4章で紹介する UNL, CDL.nl は後者の立場に立つものだが、まず関連研究として前者の立場の研究から紹介する。

### 2.1. UNO モデル

Iwanska は汎用知識表現として自然言語表現は有用であるとの考えの下、以下の UNO モデルを開発している[Iwanska 93, 00b]。概略を紹介すると、まず最もプリミティブとなるのは名詞、動詞、形容詞の単項表現であり、以下がその例である。

[woman], [walk], [sweet]

これらに対する形容修飾、副詞修飾はラベル属性・属性値として与え、元の概念を限定するサブセットを表現すると見なす。例えば複合名詞“very happy woman”を次のように表す。

[woman (happy => happy (degree => very))]

複合名詞“walk fast”は次のようになる。

[walk (speed => fast)]

ここで、happy, very, fast のように性質が判明している語については、その性質としてラベル属性を与える（ここではそれぞれ、happy(幸せラベル属性で sad などこの属性値となる)、degree, speed)。性質が分からない形容詞の場合はラベル属性を単に adj とし、副詞の場合は単に adv とする。

ブール代数演算の not (否定), and (連言), or (選言) の表現も勿論提供されており、例えば“dog or cat”は 2 要素項の [dog, cat] となるが、詳細は省略する。限定詞修飾の名詞句については、例えば“some students”は次のように表す。

np (det => [some], n => [student])

目的語をとる動詞句は、例えば“love Mary a lot”を次のように表す。

[love (np => mary, adv => a lot)]

以上を用い、文は主語部と述語部を == で結びつけ、例えば次のように表現する。

“John walks very fast, and speaks loudly.” に対して、

john == {[walk (speed => fast (degree => very)), [speak (adv => loudly)]}.

“Some students work hard.” に対して、

np (det => [some], n => [student]) == [work

(adv => hard)].

推論は基本的に集合の包含関係を基礎にして行う。“walk very fast”であれば“walk fast”, 更に“walk”が伴意(entail)されることなどは直接的であるが、異なるが関連する語彙間についても以下の例のような伴意関係を与えている。

- ・数について：two であれば one を伴意する。
- ・限定詞について：all であれば some を伴意する。
- ・副詞について：always であれば sometimes を伴意する。
- ・動詞について：love であれば like を伴意する。
- ・形容詞について：excellent であれば good を、certain であれば likely を伴意する。

これらは非常に少数の例だが、限定詞、頻度を表す限定詞、評価 (excellent, good, ok, bad, terrible など)、類似度 (same, similar, opposite など)、確度 (certain, likely, possible など)、長さ、程度 (extremely, very, somewhat など) の形容詞について、各語彙に [+1.0, 0.0] あるいは [+1.0, -1.0] 間の区間を対応させて、異なる語彙間の包含関係を計算し、伴意関係を求めるようにしている。どの程度の範囲の語彙をカバーしているのか、あるいはカバーできるのかは必ずしも明らかでない。

yesterday, young などの時間エンティティと before, after のような時相関係を含む時間関係についても、基本的には時間軸上の区間に対応させて伴意関係を求めるとしているが、どの範囲の語彙までカバーできるか必ずしも明らかでない。地理的關係も空間的包含関係に基づき、例えば“Detroit is in Michigan.”と“John is in Detroit.”から“John is in Michigan.”が伴意されるとする。記述してある範囲なら可能であろうが、常識的なレベルの推論を可能にするには大量の背景知識を整備する必要があること、空間関係を表す多様な表現への対処をどうするかなどの課題がある。

自然言語表現は一般に不十分規定(underspecified)の性質を持ち、多くの場合にコンテキスト依存で意味が定まるという性質をもつ。UNO モデルでは文を順序付きの並びの集合として、ある文のコンテキストを定めてその文の意味を規定するのに必要な情報を求める。1)陽的記述コンテキスト、2)陰的記述コンテキスト(文には陽には表れない人の気持ちなど)、3)否定的コンテキストとしての事前知識、4)デフォルトのコンテキスト、などの扱いについて検討がされているが、自然言語意味理解に共通の重い課題であり、まだ不十分な段階である。

以上のように、UNO モデルでは、自然言語文を知識表現として、包含関係を基礎にした推論機能を重視しているため、扱う語彙及び表現の範囲が狭くなってしまう。

## 2.2. 制限英語 ACE

英語については 1)正確なドキュメント記述, 2)(仕様書等での)論理的な文章の作成, 3)(非英語民族向けの)読みやすい英語, 4)機械翻訳の品質向上, 5)コンピュータとのコミュニケーション, 等のために制限英語が開発されてきた。

最初に 1930 年頃にイギリスの C. K. Ogden は英語で意味を記述するのに 850 語で可能なことを発見し, Basic English を提唱した。この Basic English では動詞は次の 16 語だけある: come, get, give, go, keep, let, make, put, seem, take, be, do, have, say, see, send (ただし, 合成語規則により overcome, become などのように幾つかの用語も動詞として使うことができる)。助動詞は, may と will だけである。2 つあるいはそれ以上の語を連結して合成語として使用できる (例えば, law + make + er = lawmaker)。不自然でない限り, 名詞に -er, -ed, -ly, un-などの接頭詞, 接尾詞を付け, 形容詞あるいは副詞的に使用することができる。

制限英語は近年では, 産業界でグローバルな業務を行う上での必要性から, キャタピラー, ボーイング, IBM, GM, XEROX 等で開発が行われてきた。

ここでは 英語文を知識として利用する観点を持ち, チューリッヒ大学を中心に現在も研究開発が進行中の ACE(Attempto Controlled English)[ACE]について紹介する。

ACE の語彙と許される構文は以下のものである。語彙は機能語(限定詞, 限量詞, 前置詞, 接続詞, 否定語, 代名詞, 数値など)と約 10 万エントリの意味語(名詞, 動詞, 形容詞, 副詞)であり, ユーザはドメイン固有の語彙も加えられる。名詞句, 動詞句, 及び文の構文は限定されている。全てを示すことはできないが, 以下に幾つかについて例示する。

- 前置詞句は(名詞でなく)動詞を修飾する。
- 関係節は直前の名詞を修飾する。
- 関係節内で接続を表す時は関係代名詞を繰り返す。(例: A customer inserts a card that is valid and that has a code.)
- 複数名詞句は一般に解釈が曖昧になるが, ACE では複数名詞句は集合的及び配分的解釈をとり 配分的解釈をする場合には each of で表す。(例: A clerk enters each of 2 cards.)
- 前方参照は次の場合のみ許す。正規名称(例: John), 代名詞(例: a card に対して it や itself), 定冠詞名詞句(例: a card や some water に対して the card, the water), 変数(例: a card X に対して the card X や X)。前方参照は accessibility, recency, specificity, reflexivity に基づき一意に決定する。
- 人称代名詞 (he, him など) や所有代名詞 (his など) の前方参照の場合は, 性と数が一致し, 文の主

語でないアクセス可能な直前の人物に決定する。

- 動詞句でモダリティ (can, must など) や意図性 (believe など) は使えない。
- 前置詞付き動詞の前置詞は動詞にハイフンで結ぶ (wait-on, look-up, apply-for など。これにより解釈時に "A student is interested-in a course." と "A student is interested in a classroom." が区別される)。
- 副詞は動詞の直前か直後 (ただし, and や or で結ばれた複数の副詞, あるいは前置詞句との結び合わせは許す)。
- 複合文は接続詞, 否定構文 (It is not the case that ... など), 従属接続詞により構成する。(接続詞 and, or は単に論理的な意味を表し, 時相は表さない)。
- 従属接続については (現状では) 関係節と条件文 (If...then...で構成される) のみ。関係節は直前の名詞句を修飾する (a customer who is rich など)。先読みで構文カテゴリを選択できる look-ahead text editor が開発されているが, (慣れにもよるであろうが) かなり厳しい制限となっている。

このような制限英語 ACE テキストは曖昧性なしで DRT (Discourse Representation Theory)[Kamp 93] に基づく表現 DRS (extended Discourse Representation Structure) に変換される。以下はその変換例である。

ACE テキスト: Every company that buys a standard machine gets a discount.

DRS:

```
drs([], drs ([A, B, C, D, E], [structure (B, atomic)-1, quantity (B, cardinality, count_unit, A, eq, 1)-1, object (B, company)-1, structure (D, atomic)-1, quantity (D, cardinality, count_unit, C, eq, 1)-1, property (D, standard)-1, object (D, machine)-1, predicate (E, event, buy, B, D)-1])) => drs ([F, G, H]), [structure (G, atomic)-1, quantity (G, cardinality, count_unit, F, eq, 1)-1, object (G, discount)-1, predicate (H, event, get, B, G)-1]))
```

更にこの DRS は Flat First-Order DRS, 次いで 1 階述語論理表現へと変換されて知識ベース化が図られる。そして, 質問に対する推論が実行できるようになっているが, 詳細は省略する。

このように ACE も推論機能を重視しているため, 語彙は 10 万程とかなり有しているものの, 構文上の制限はかなりきつめになっている。

### 2.3. その他の枠組み

著者らは、1999 年頃から自然言語表現による知識表現と推論を行う枠組みとして、KRNL (Knowledge Representation for Natural Language)の開発を行った。これは化学者である本木協司氏による「原子が結びつき分子となり、異なる分子が融合して高分子となって高次機能が発現するように、語彙が結びついて文となり、複数の文が結びついて人の頭の中で高次概念が形成される」といった考えを基にシステム化したものであり、当初は CCR (Concept Chemical Representation)と称していた[本木 98, 友部 00]。自然言語テキストを入力するとシステムが可能な複数解釈を提示し、人が正しいものを選択することで半自動的な KRNL 表現の生成機能、表現の範囲を現在形の文のみに制限し Prolog に変換して推論を行うシステムなどが作成された。単語表記の揺れへ対処するためには、語彙辞書を整備する必要があり、更に意味的マッチングを可能とするためには類義語辞書の整備が必要となるなどの課題があり、スケールアップが困難となった。

以上は推論機能を重視したアプローチであるが、これらとは別の系譜として意味ネットワークをベースとするアプローチが存在する。元来の意味ネットワーク [Quillian 68]は概念記憶の認知的構造として生まれたが、その後幾つかの強化、拡張が図られてきた。自然言語文の表現への対応に向けては、リンクに与える文要素間関係の記述機能の追加を中心に拡張と精密化が図られてきている。これらの中で SNePS [Shapiro 00]は特に 1 階述語論理との関係に考慮が払われている。MultiNet [Helbig 06]は自然言語文の表現に必要な関係等の記述の精密化が最も深まっていると考えられるが、反面で多層拡張意味ネットワーク構成となり複雑になってしまっている感がある。

## 3. UNL

UNL は“Universal Networking Language”の略語であり、国連大学で 1996 年にスタートした UNL プログラムの下で開発された[Uchida05, UNDL]。UNL プログラムはコンピュータのための共通言語を作成し、Web を中心に世界の人々が自国の言語で情報や知識を提供し、アクセスすることができるようにすることを目的とした。コンピュータは UNL の意味を理解し、互いにコミュニケーションすることができるようになり、その結果、多言語で情報を発信したり受け取ったり、理解したりするための言語基盤を人々に提供する。

UNL はハイパーノードを含む意味ネットワークの形式で情報や知識を表現し、その表現には曖昧性がない。UNL の意味ネットワーク表現において、ノードは概念を表し、アークは概念間の関係を表す。概念に

は注釈がつけられる。

UNL の語彙を構成するのが UW (Universal Word) である。UW は概念を表す“単語”であり、他の UW と関係子によってリンクされ文の意味を表現する。関係子は文を構成する各々の単語 (UW) の役割を指定することになる。作者が意図した主観的な意味は関係子とともに UNL の文法を構成する属性で表される。この主観性(subjectivity)やモダリティも属性として記述することで、UNL は文のニュアンスも含めて表現可能である。

UW は世界中の人々が認識できる概念を正確に表現し、かつそのような概念が誰でも同じ記号でもって定義できるようになっている。このため意味素には英語の単語を用いるが、意味素の表現する概念の範囲を限定し、表現したい概念の範囲を正確に表現するために、UNL の関係子に基づく他の UW からの束縛によって記述している。これにより多義性をもつ単語も、多義を分解した一義の語彙として定義される。現在、UW の語彙数は約 20 万語である。

UW の意味や役割は曖昧性なく知識ベース (UNLKB) に定義されている。この知識ベースを参照することで、UNL の文の意味づけが曖昧性なく行われる。また、この知識ベースは概念の包含関係に基づく階層関係を含む UW 間の関係の定義も含み、概念の包含関係に基づく推論機構も備えている。UNL と各言語の橋渡しは、UNL と各言語の対訳辞書と文法変換規則を用いて行われる。

コンピュータが知識に基づく処理を行うためには、言語知識だけでは不十分である。例えば、「その本を書いたのは太郎である」ことから「太郎はその本の著者である」ことを推論するためには、文の中に出てくる言葉の意味だけでなく「著者とはその書物を書き著した人である」のような知識も必要となってくる。UNL にはこのような言葉の意味に関する知識を蓄え、利用するメカニズムも用意されている。

UNL システムはすでに実証テストの段階になっている。UNESCO で開発されたエンサイクロペディアの UNL 化に基づく多言語提供や、WIPO と共同で IPC (International Patent Classification) の UNL 化に基づく多言語提供および検索機能強化、各国図書館との図書情報の UNL 化などが計画され、一部開発が行われている。

### UNL システム

UNL システムは主要 3 コンポーネントから成っている。言語資源、それを処理するための言語処理ソフトウェア、それらを管理・運用するソフトウェアツールである。図 1 はその構成を示している。

言語資源は言語独立なものと言語依存なものに分かれる。言葉の概念やその概念に関する語学知識 (Linguistic Knowledge) は言語独立なものとして共

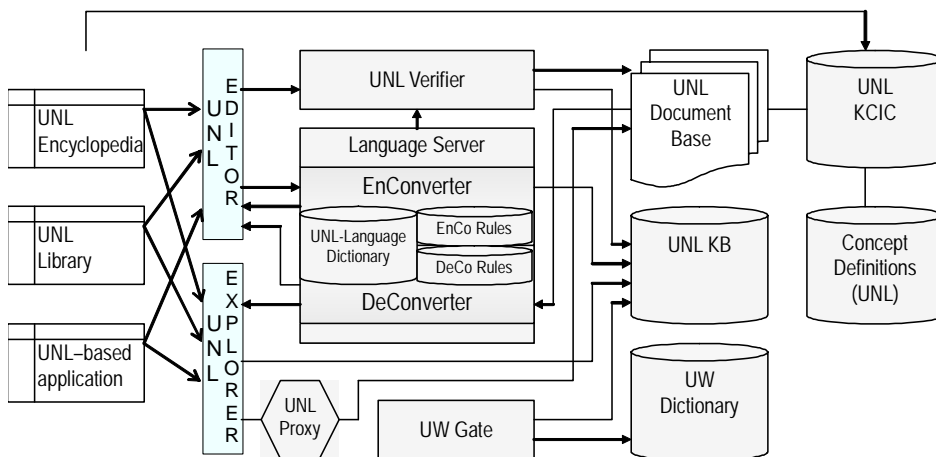


図1 UNLシステムの構成

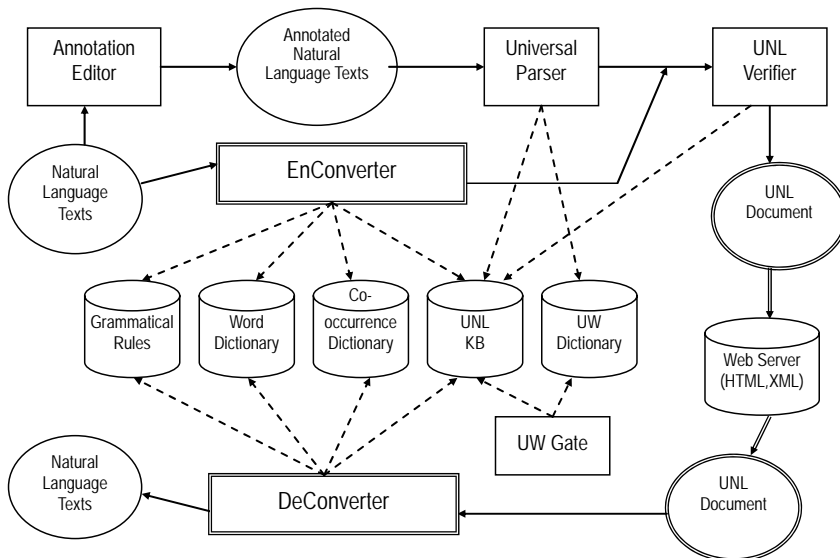


図2 UNLの変換メカニズム

通にUNL知識ベース(UNLKB)に置かれる。言語依存な言語資源、例えば言語処理用辞書やルールは言語処理ソフトウェアとともに言語毎に少なくとも一つは存在する言語サーバで使われ、これらの言語サーバはインターネットを通してUNLを介し相互接続される。このように各言語サーバは統合され、言語障壁を越えた人々のコミュニケーションを可能にする多言語基盤を作り上げる。

図1右上部のUNL KCIC(UNL Key Concept in Context)はUNLドキュメントの主要概念情報を蓄積し、UNLKBでは言語知識はコンテキスト独立な概念間の関係で表されるが、コンテキスト依存の概念間の関係で規定されるような現実の知識を表すことを可能にする。

UNLドキュメント作成のサポート・ツールはユーザのコンピュータで使用可能となっている。言語独立な言語資源であるUNLKBを作成・運用するツールや、概念と各言語とのインターフェースであるUW辞書を作成・運用するツール(UWゲート)も提供されている。自然言語で書かれた情報や知識は、それらをUNLに置換することで多言語化が可能

になる。したがって、UNLの応用範囲は広いが、現在は多言語でのWebページの参照を可能にする多言語Webページシステムや、知識を再帰的に定義するとともに多言語での知識参照を可能にするUNLエンサイクロペディアなどが中心的な応用システムになっている。

UNLの価値は各種言語で記述された情報や知識を共通的に記述、表現できるということであり、言語の壁を越えグローバルな情報・知識の流通と共有の基盤になりえることである。

### UNLの変換システム

UNLシステムの中核となるソフトウェアは、エンコンバータ(Enconverter)およびデコンバータ(Deconverter)からなるソフトウェアと、UNLの資源を作成、管理するソフトウェアからなる。

図2はUNLの変換に関するメカニズムを示している。エンコンバータはパーサ(解析システム)であり、自然言語をUNLに変換する。作成されたUNLドキュメントはベリファイヤツールで正しさを検証され、保存される。デコンバータはUNLを自然言語に変換する。ユニバーサルパーサ(Universal Parser)はエンコンバータの拡張版で、解釈の曖昧性を回避するためのタグを付与したテキストを入力対象とする。

エンコンバータとデコンバータを使用したツールがUNLエディタとUNLプロキシ・サーバである。UNLプロキシ・サーバは言語サーバとの通信を行うことにより、UNLで書かれた文書を各国語で見る

ためのものである。

### UNL における知識の表現

自然言語による知識は、言語知識、概念の語義文、および世界知識の三つのレベルの知識に分けることができると考えられる。

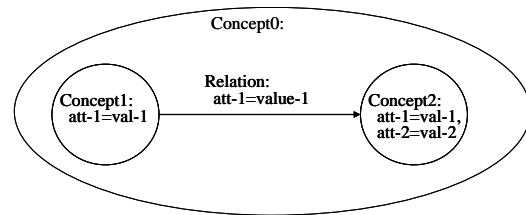
言語知識は自然言語に含まれる言葉の概念に関する知識であり、自然言語の意味を理解するのに必要な知識である。例えば「著者」は「人間」であり、「人間」が備わっている機能を持ってさまざまな行為をすることができるといった知識である。このレベルの知識は自然言語の意味的背景を与えるのに必要である。

語義文 (Concept Definition) はある概念を他の概念との関係に基づいて定義するものである。例えば「その書物を書き著した人」は「著者」の定義文であり、この定義文によって「著者」とその他の概念との関係が特定でき、「その書物を書き著す」との関係が分かる。このレベルの情報は情報検索などにおける推論に不可欠である。最後の世界知識はエンサイクロペディアをはじめあらゆる種類やレベルの知識に関する情報である。このレベルの知識は最終的には人間に必要な知識を提供するものであるが、情報検索の過程においてコンピュータも必要な知識を問い合わせることができる。言語知識は UNL では言葉の概念を UW として UNLKB で定義される。概念の語義文や自然言語に基づく知識は UNL で表現される。

## 4. CDL.nl

CDL (Concept Description Language) [Yokoi05] は本特集号の [横井 06] にも記されているように、自然言語テキストだけでなく他も含む広いメディア一般が表す概念を表現する汎用で基本的な枠組みとして設計されている。簡素な 3 つの組表現 (<実体 1, 関係, 実体 2>, <主語, 述語, 目的語>あるいは<実体, 属性, 属性値>などを表す) を基礎とし、グラフ表現だと実体を表すノードと関係を表すアークから成り、ノードは (更にはアークも) グラフを含むことができる (複合実体となる) のでハイパーグラフとなる。実体と関係には詳細特性記述のために任意個の属性 - 属性値 (attribute-attribute value) を付加できる (関係に準ずるのものであり、もし属性値が複合実体となるなら関係とすることができる)。図 3 はグラフ表現とテキスト表現の例を示している。

CDL.nl (CDL natural language version) は UNL の成果を CDL の枠組みの上に再設計した、自然言語テキストの意味する概念を汎用的に表現する概念記述言語である。UNL の成果を活用し、自然言語表現をカバーする実用レベルの基本語彙 (記述に必要な関係語彙セットも含む) オントロジーを備えていることも大きな特徴である。UNL は単文単位の扱いであるのに



```

    {Concept0; {Concept1: att-1=val-1; }
    {Concept2: att-1=val-1, att-2=val-2; }
    {Relation: att-1=value-1; }
    Concept1 >Relation Concept2 }
    
```

図 3 CDL のグラフ表現とテキスト表現

対し、CDL.nl では複数文に渡る関係 (談話構造) も記述の対象にする。セマンティック Web におけるメタデータの記述に留まらず、Web のソース自然言語テキストまでも含めて共通の意味概念記述を可能とすることで、次世代 Web 基盤としての国際標準化も視野に入れていることから、十分な国際性を持つことを前提にして設計されている。

自然言語の意味に関しては、レベルや側面に関して様々なものがあるが、CDL.nl で記述対象とする意味概念は最も一般性があるものとして、表層表現から多少意味に踏み込んだ概念レベルとしている。幾分具体的に表現対象を規定すると、次のようである。

- 1) 辞書的多義を分解した概念を要素実体概念、事物概念とする。
- 2) 事物概念を要素実体概念として、述語成分と格成分及び述語修飾成分によって、複合実体概念、単事象概念を構成する。
- 3) 単事象 (単文) 概念は命題部分とモダリティ的部分に分ける。
- 4) 単事象 (単文) 概念を要素実体概念として節の修飾関係や主従関係によって複合実体概念、複事象 (複文) 概念を構成する。
- 5) 複事象 (複文) 概念を要素実体概念として、それらの論理的・時間的・因果的關係や相互参照関係によって複合実体概念、状況 (文章・談話) を構成する。

以上のように、文章、文の意味を複合実体として概念化して記述することになる。文章の複合実体は複数の文の複合実体をノードとして含み、それらのノード間は文間の接続構造を表す接続関係か、指示表現や代用関係等を表す参照関係で結ばれる。接続関係 (接続詞) は結び付けられる要素がどのような繋がり関係にあるのかを言語化して示し、文章・談話の全体的構造や成分としての文段・話段をまとめ上げる外的な構成機能をもつ。これに対し、指示関係 (指示詞) は、主として内容面から文章・談話の話題の連鎖を示し、全体的構造や成分を情報としてまとめ上げる内的な構成

機能を果たす。

文は命題部分とモダリティ的部分に分けて記述するが、命題部分は“実体 - 関係”で表し、モダリティ的部分(アスペクト, テンス, ポラリティ, モダリティ)は“属性 - 属性値”で表す。命題部分は、述語成分、述語修飾成分、複数の格成分という実体が格関係で結ばれた構造となる。それぞれの成分が複合語、句、節で構成されている場合には複合実体となる。文の実体内には格関係に加えて、内部での参照構造を表す関係も記述することになる。基本語彙(記述に必要な関係語彙セットを含む)は UNL の UW(universal words)を受け継ぐ形で構成されているが、これが充実していることが大きな特長であることから、ここで全体 [CDL.nl]を示すことは出来ないものの、その一端を示すことにする。

CDL.nl の語彙を分ける最上位オントロジーは、

Entity, Relation, Attribute

である。Entity の下位オントロジーには Thing(事物)と CompositeEntity(複合実体)がくる。Thing の下位に以下のカテゴリーが置かれる。

- NominalThing(抽象物, 具体物, 機能, 有意志体, 場所, 代名詞を含む)
- NominalModifier(限定詞と形容詞)
- VerbalThing(Do, Occur, Be 関係の動詞, 述語)
- VerbalModifier(数量と副詞)
- ValueOfAttribute(属性値)

Relation(関係)については、自然言語で表される概念をどのような観点からとらえるかを示す意味で重要であるので、表 1 にその基本関係語彙を提示する。

以下は自然言語文と対応する CDL.nl 表現の簡単な例である。

```
"The verse bought yesterday is broken."  
{#A008: buy:01:tempo=past >tim yesterday:02;  
  01: >obj verse:03:bquan=def;  
  broken:04:view=entry >aoj :03}  
(ここで#A008 は表記する概念を表す ID 番号 01, 02,  
03, 04 は各要素を示すローカルな ID 番号で同一概念表  
記内ではこの番号だけで参照可。tempo は時間性の属性  
で past はその値, bquan は基本限量・限定表現の属性  
で def は特定の実体であること(この場合は the verse)  
を示す値, view は事象や状況の焦点や話題を表す属性で  
entry はこの文の主概念であることを示す値。tim は時間,  
obj は対象, aoj は属性主の関係を表している。)
```

以上のように CDL.nl は自然言語の様々な表層表現からやや意味に踏み込んだ概念レベルを高い近似度で表現する機能を提供する。より深い意味理解や推論機能等は外部プログラムに委ねられる。すなわち、ありのままの自然言語テキストを対象とするが、一旦 CDL.nl の概念記述に変換し、これを共通の土台として幅広い展開を可能にする。CDL.nl 自体は個別の言

表 1 CDL.nl の基本関係語彙

#### ElementalRelation 要素関係

##### CaseRelation 格関係 (=IntraEventRelation 事象内関係)

[QuasiAgent 準主体] Agt(agent:動作主), Aoj(thing with attribute:属性主), Coag(co-agent:並行動作主), Cao(co-thing with attribute:並行属性主), Ptn(partner:相手)

[QuasiObject 準客体] Ben(beneficiary:受益者), Cob(affected co-thing:並行対象), Obj(affected thing:対象), Opl(affected place:場所対象)

[QuasiInstrument 準方法] Ins(instrument:道具), Met(method or means:方法), Man(manner:仕方)

[QuasiState 準状態] Gol(goal, final state:終状態), Src(source, initial state:始状態), Via(intermediate place or state:経由)

[QuasiPlace 準場所] Plc(place:場所), Plf(initial place:起点), Plt(final place:終点), Scn(scene:場面)

[QuasiTime 準期間] Dur(duration:期間), Tim(time:時間), Tmf(initial time:始時間), Tmt(final time:終時間)

[QuasiBasis 準基準] Bas(basis for expressing a standard:基準)

##### InterThingOrInterEventRelation 間事物・間事象関係

And(conjunction:連結), Con(condition:条件), Coo(co-occurrence:同起), Fmt(range/from-to:範囲), Frm(origin:起点・起源), Or(disjunction, alternative:選択), Pur(purpose or objective:目的), Rsn(reason:理由), Seq(sequence:先行), To(destination:目的地)

##### Quantification&ModificationRelation 限定・修飾関係

Cnt(content:内容), Man(manner:仕方), Mod(modification:限定), Nam(name:名前), Per(proportion, rate of distribution:単位), Pof(part-of:部分), Qua(quantity:量)

#### CompositeRelation 複合関係

Pos(possessor:所有者), .....

語に依存せず、汎用的に自然言語の概念意味表現に使えるものであるが、共通である関係の語彙(機能語)を除く語彙オントロジーを各国語対応に用意することにより、各国語対応バージョンを作成できる枠組みとなっている。

CDL.nl は Semantic Web ととも融合して、次世代 Web 基盤技術として国際標準化を図っているため、RDF、OWL 形式の表現への変換も開発を行っている。

CDL.nl の現状での一つの課題はその記述生成法である。自然言語テキストからの全自動変換は一つの理想形であるが、自然言語理解技術の成熟を待たねばならない。UNL で実現されているように自然言語テキストへ必要なタグを付与し、変換するのが一つの方法である。(ただし、UNL は単文単位の変換なので、文間の関係記述については別途の支援機能が必要にな

る。)システムが複数個の変換候補を提示し,人が正しいものを選択するようにし,次第にシステム側の機能を向上させていく半自動変換も一つの方法である.更に,自然言語テキストからの変換という考え方はとらず,橋田のセマンティックオーサリング[橋田 06]のように,最初からグラフ入力・編集を行うというのが全く別の方法ということになる.

## 5. むすび

自然言語テキストの共通的概念意味記述のレベルや記述法,必要となる基本語彙オントロジー等については必ずしも十分な合意が形成されているとは言えないものの,長年の自然言語理解や機械翻訳,電子化辞書等の研究の積み重ねの上に,一定の見通しが得られてきていると言える.コンピュータやエージェントも自然言語テキストの意味を把握・理解できるようにし,人間との共働の地平を飛躍的に拡大することは,次世代 Web 基盤として非常に重要な方向であり,今日の時点で国際標準として共通的概念記述の枠組みを設定することの現実的意義は大変大きいと言える.このようなプラットフォームの設定は,単に Web だけでは限らずに今後のセマンティックコンピューティングへ向けての共通の土台として,大きな役割を果たすものと考ええる.

本原稿では,以上のような自然言語テキストの共通的概念記述法として,UNL とその成果を活用して設計された CDL.nl を中心に紹介した.我が国発の技術として賛同と協力の輪を広げ,国際貢献を果たして行きたいと考えている.

**謝辞:**CDL, CDL.nl の設計に主要な役割を果たされたのは横井俊夫氏(東京工科大)である.内田裕士氏と朱 美英さん(UNDL 財団)は UNL を設計し実現され,CDL.nl の設計にも貢献された.橋田浩一氏(産総研)と安原宏氏(沖電気及び ISeC)は CDL, CDL.nl の設計に貢献された.記して感謝致します.また CDL.nl に関する国際標準化活動は総務省 SCOPE 国際技術獲得型研究開発の支援を受けている.

## 参考文献

- [ACE] <http://www.ifi.unizh.ch/attempto/>  
 [CDL.nl] Specification of CDL.nl, ISeC 資料 (2005), <http://www.instsec.org/tr/>  
 [橋田 06] 橋田浩一:オントロジーと制約に基づくセマンティックプラットフォーム,人工知能学会誌,Vol.21, No.6, pp. (2006.11)  
 [Helbig 06] H Helbig: *Knowledge Representation and the Semantic of Natural Language*, Springer (2006)  
 [Iwanska 00a] L. M. Iwanska, and S. C. Shapiro(eds):

*Natural Language Processing and Knowledge Representation*, The AAAI Press/The MIT Press (2000)

[Iwanska 93] L. M. Iwanska: Logical Reasoning in Natural Language: It is All about Knowledge, *Int'l Jour. of Minds and Machines*, Vol.3, No.4, pp.475-510 (1993)

[Iwanska 00b] L. M. Iwanska: Natural Language Is a Powerful Knowledge Representation System: The UNO Model, in [Iwanska 00a], chap.1, pp.7-64 (2000)

[Kamp 93] H. Kamp and U. Reyle: *From Discourse to Logic*, Kluwer Academic Pub. (1993)

[木本 98] 木本協司:研究開発における発想とその支援ツール:概念化学プログラミング---構造化中間言語,研究開発マネージメント(1998)

[Quillian 68] M. R. Quillian: *Semantic Memory*, in *Semantic Information Processing* (M. Minsky, ed), pp.227-270, The MIT Press (1968)

[Shapiro 00] S. C. Shapiro: SNePS: A Logic for Natural Language Understanding and Commonsense Reasoning, in [Iwanska 00a], chap.5, pp.175-195 (2000)

[友部 00] 友部教博,木本協司,石塚満:知識表現 CCR による知識統合,人工知能学会知識ベース研究会資料 SIG-KBS-A003, pp.61-66 (2000)

[Uchida 05] H. Uchida, M. Zhu and T. G. D. Senta: *UNL—Universal Networking Language*, UNDL Foundation (2005)

[UNDL] <http://www.unndl.org/>

[Yokoi 05] T. Yokoi, H. Yasuhara, H. Uchida, M. Zhu and K. Hashida: CDL (Concept Description Language): A Common Language for Semantic Computing, *Online Proc. WWW2005 Workshop on the Semantic Computing Initiative (SeC2005)*, Makuhari, Japan (2005.5)

[横井 06] 横井俊夫:セマンティックコンピューティング—知的システム・知的環境の設計原理,人工知能学会誌, Vol.21, No.6, pp. (2006.11)

2006 年 9 月 8 日 受理

## 著者紹介

### 石塚 満 (正会員)

1971 年東京大学工学部電子卒,1976 年同大学院博士修了.工博.同年 NTT 入社,横須賀研究所勤務.1978 年東大大学生産技術研究所・助教授,(1980-81 年 Purdue 大学客員準教授),1992 年東京大学工学部電子情報・教授,2001 年情報理工学系研究科・電子情報学専攻,2005 年同創造情報学専攻(電子情報学専攻兼任).研究分野は人工知能,Web インテリジェンス,次世代 Web 情報基盤,生命的エージェントによるマルチモーダルメディア.IEEE, AAAI, 情報処理学会,電子情報通信学会,映像情報メディア学会,画像電子学会の会員.本会前会長.