ユニバーサル・レキシコンの工学的メリット

野村直之
NEC情報メディア研究所

1　はじめに

計算辞書は、言語処理システムにおける大部分のあいまい性の知識源であるため、あいまい性の解消を主たる課題とする自然言語処理[注1]にとって、極めて重要である。知識源の表現形式はシステムの要求により自由にアーキテクチャを設計できる。しかし、その内容は、研究者が注目する以前から自然界に存在していた自然知識として、構造や制約の解明に取り組む必要がある。

言語工学の分野では説明的理論の構築を目的としないため、言語現象の記述的妥当性が得られれば良いという立場をとる。語彙知識の記述においても実データとモデルの照合が重視されてきたが個々の素性のレベルになると現象をそのまま裏返しにしたような直裁的な記述指針となる傾向がみられる。たとえば、日本語の表層格素性を辞書に記述する際「『で』をとるなら『デ』という素性を付与する」等の直截的な判断基準に頼らざるを得ないという状況である。これでは記述が収束する保証はなく、再現性という工学の要求を満足させることができない。

一方、計算辞書の設計者としては、語彙知識の総体を分類し、その素性のとりえる自由度の大きさ、素性間の依存関係、制約の分量などを見積もり、その妥当性を評価する必要がある。語彙に内在する素性は多くの場合語彙・統語素性と意味・概念の素性の2種類に分けて把えられる[Jackendoff 90, 岡田&中村94]。これらは互いに独立・不干渉のものとして扱われ、モジュラー性の向上によって扱いやすい知識表現が得られる。では、こうした場合の記述の自由度の大きさはどうなっているだろうか。シソーラスの導入によって概念側の自由度を適切に減少させる試みがあるとはいえ[岡田&中村94]、統語素性と概念素性の依存関係を個別に記述する枠組では、実在しない素性の組み合わせを許してはいないだろうか。それがレキシコグラファ間の判断の再現性を妨げているのではないだろうか。

いわゆる表層格と深層格の対応のヴァリエーションを例にとってみる。仮に格スロットの最大数を4、深層格数を50で計算すると可能な組み合わせは約2５万となり[Nomura94b]、我々の経験値であるところの高々数百というオーダー[野村89]と大きな食い違いがある。この違いに潜むなんらかの広範な強い制約が新たに明示されなければ、レキシコグラファ間の判断が収束する見込みはないのではなかろうか。

では、統語側と概念側が互いに依存関係があるならば同義語、類義語どうしが殆どの場合、互いに同じ格フレームをとるかといえば、それは事実に反する。動詞 *飾る* が、「窓に花を*飾る*」「窓を花で*飾る*」の2種をとるのに対し、*よそおう*、*飾り付ける*、*装飾を施す*、*アレンジする*、などの類義語は何れか一方しかとらない。その一方で、意味の離れた *巻く*、*塗る*、*(酒を筒に/筒を酒で)満たす* などは、*飾る* と同じ振る舞いをする。Nomura94bでは、日本語以外の4つの言語でも、全く同じヴァリエーションの存在と類義語間の不一致がみられ、この現象が言語普遍に存在するらしいことを指摘している。さらに言語間の同義語間で一致／不一致の様々な組み合わせが存在することも指摘されている[注2]。辞書を設計する立場からは、深層格の一貫性をとりつつ2種の格フレーム間で概念(語義)を共有させられない、などの困難も指摘される。

本稿では、上記の様々な問題点を解決するための、説明的理論を手段とした1つのアプローチを紹介する。また、実験により、複数言語にまたがる既存の資源を用いることで個別言語の統語素性を開発・検証するという新しい手法の有効性を示す。

2　ユニバーサル・レキシコン

Nomura94bで提出されたユニバーサル・レキシコン（以下ULと略す）は、個別言語の統語素性が言語普遍の原理によって導出されることを仮定する説明的理論である。図1にその導出の過程を示

[注1] 江原&田中93には、「機械翻訳の計算論的に定式化すると単なるグラフ書き換え問題に過ぎない。曖昧性解消のための知識源の構造と内容の解明こそが真の課題である。」という卓見が述べられている。

[注2] 不一致の例：「太郎が花を飾った。窓に。」の前半を "Taro decorated the flower." とするのは大きな誤訳。これでは、花にリボンかなんかを巻き付けたとしか解釈できない。

す。概念の側は、概念分類と言語概念素性とにモジュール化される。概念分類は、概念役割がいくつ存在するか(0, 1, 2 or 3)、それらは動作を経て接近するか離反するかという観点で階層的に分類したものである。言語概念素性は、各概念分類に属する個々の概念について {アスペクト焦点, 他動性, 能格性, 与格位置, 動作主位置} のスイッチを指定する。これら互いに独立に決まる分類、素性が特定の1つの語彙に付与され、あるいは複合語が形成される際に、各分類スイッチに対応して可能な格フレームを統語素性生成器が生成する（例えばこの *生成する* では他動性が+でも-でもなく *指定無し* なので他動詞/自動詞に対応する2つの格フレームが生成する)。ここでさらに、対応する表層格マーカの有無などの言語依存のパラメータによるチェックを経て、最終的に格フレームのセットが定まる。たとえば "clear the table of the dishes" にみられる、いわゆる'分離のof' と同じ機能の格マーカが日本語には無いため、そのスロットの生成は抑止され、「テーブルを片づけた」となる。

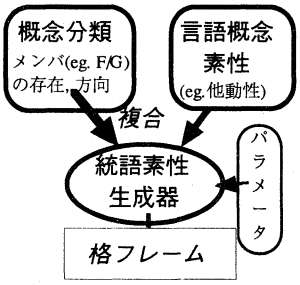


図1 ユニバーサル・レキシコン

先の、*飾る* と *decorate* の違い、それから *覆う* と *掛ける* が複合して *覆い掛ける* となったときの格フレームの説明能力を評価し、モデルを検証してみる。これらの動詞は互いに同義語とは言いがたいが、動作主以外に2つの概念役割をもつ点が共通している。この2つをF, G と名付けると[注3]、動作前に離れていたFとGが動作中に互いに接近し、動作後には同一の場所に存在するという点も共通である。したがって図1の概念分類の中でこれらは同一の分類に属する。一方、言語概念素性のほうでは、アスペクト焦点が際立った違いをみせる。*飾る*、*巻く*, *spray* は焦点がF, Gに固定されておらずどちらも可能であり、*覆う*, *decorate* は焦点がGに固定され、掛けるはFに固定されている[注4]。

統語素性生成器は、アスペクト焦点がFのものに対してその動作方向に基づいて "F を G に" を生成し、Gのものに対しては "GをFで" という格フレームを生成する。こうして、ある動詞が*何故*その格フレームのセットをとるかの理由の一部が説明されるが、それにとどまらず、このモデルは複合動詞の格フレームに対する予測能力ももつ。たとえば「赤ん坊を毛布で *覆う*」が「赤ん坊に毛布を *覆い掛ける*」となるとき、複合動詞全体のアスペクト焦点は *掛ける* のもので上書きされる[注5]。故に、「毛布を *覆い掛ける*」でも「ネクタイをスタンドに *巻き掛ける*」でも、アスペクト焦点がFとなるため、'F を G に' という格フレームしかとれなくなることが予測されたのである。

3 ユニバーサル・レキシコンULの工学的メリット

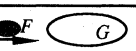
言語工学では、説明的理論自体は研究目的ではないため、前節に記した説明能力・予測能力のみによってULの有効性を認めることはできない。本節では、計算辞書の開発や、中間言語の性能を向上させる局面を中心に、ULの工学的メリットを概観する。

メリット1　「直截的でない語彙素性の判断基準を提供」

UL は、辞書内部の素性のヴァリエーションの実態に即して概念素性側と統語素性側との間に強い制約をもたせたモデルである。また、従来ばらばらに扱われていたこともある複数の統語素性を構造化してランダムな組み合わせ自由度を除去する効果も発揮する。たとえば、1) 1つの格フレーム中のスロットごとに独立した振る舞いを許さないこと、2) 複数格フレームのセット、さらにその間の格役割の交代現象に意味付けし、これらを新たに語彙的素性とみなしたこと、があげられる。さらに、3) アスペクト焦点など基本的な素性のいくつかが前節に記したような独立の判定基準をもつこと、によって新しい検証手段が得られる。これにより「『で』をとるなら『デ』を付与」という域を脱して、'工学' の名に値する言語知識ベースの開発が可能になる。

メリット2　「概念分類に対し統語的テストという検証手段を提供」

ULで想定する、統語素性の知識源としての概念分類という仮定が正しければ、この概念分類は、統語素性との強い依存関係に基づいて統語テストによって一部検証できる見込みがある。

注3　動作前相： 

動作後相： 

FとGの役割は左のように図解できる。

注4 「床を布で覆いつくした」「水を壁に掛けきった」のような完了アスペクトとしたときに全て尽くされた要素が何であるか？ FかGか、どちらでも良いか？ を判定することによって、格関係とは独立に判定される。

注5 この過程を言語学的に説明しようとした試みがNomura94a にある。

メリット３　「単一の木構造で概念分類を構築しようとしたときの組み合わせ爆発を抑止」

概念分類、シソーラスの構築を行うにあたって(たとえばEDR89)、同じ動作概念の自動詞(Event)と他動詞(Action)をどのように扱うかという問題があった。数百ないし数千の中間ノードをもつ１枚の概念体系の構築を考えると、動作主の有無や、アスペクト焦点のある概念役割の種類によって様々に分類が分かれてくる。図２は、動作主の有無をどの階層で分けるか迷っている状態を示した図である。EDR89では図２の右側の対処としたようであるが、その結果、管理は非常に難しくなったようである。左側の対処にしても高位のノード以下が全てEventならEventと、その属性を継承しているかどうか、概念分類を構築しながら管理していくのは困難だと予想される。ULでは、他動性やアスペクト焦点などのオントロジーには無関係の言語概念素性を分離してモジュール化された構成をとることにより、肥大した１枚の概念分類木を管理するための多くの無駄な知的作業を回避できる見込みがある。

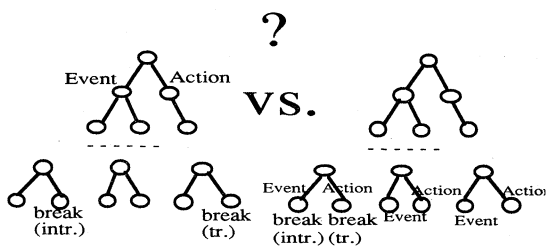


図２　EventとActionをどの階層で区別するか

メリット４　「中間言語の基盤の強化」

ULは、従来のいわゆる「意味的なAgent やObject」を含む深層格に代えて、FとGに代表される概念に内在する概念役割、およびその他の言語概念素性を中間言語の要素として導入することを提案する。これにより、「太郎が花を飾った。窓に。」の前半を安易に "Taro decorated the flower." と訳出する誤りを防止し、正しい語義選択・翻訳を行うための中間言語の理論的基盤を強化する。

メリット５　「辞書登録すべき複合語とそうでない複合語を区別する手段を提供」

動詞複合語については、図１に示した辞書内部での生成プロセスを経た場合、言語概念素性のスイッチの一部が変化することにより、概念役割の焦点が変わったり格フレームの本数が増減したりする。これに該当する複合語は辞書登録の対象とすべきであり[注6]、そうでないものは[注7]、対象から外すべきである。名詞の関係した複合語形成の切り分け方針についてはKageyama93に試論がある。

メリット６　「ある言語の統語素性の開発に異言語で作られたULの核知識ベースを活用できる」

従来、ある言語の文法情報を別言語の文法情報を用いて高精度に開発する試みを理論的に正当化する報告はされていないようである。これを試みた実験の結果を次節に報告する。

4　実験

英語で、F, Gの両概念役割を持つ動詞約200語を対象に、母国語話者によってこれらの格交代現象をを精密に検証したデータがLevin93に掲載されている。これを利用して比較的短時間で、日本語のUL知識ベースを開発するのに、次の手順をとった。

手順１「各英語動詞に対してほぼ対訳とみなせる日本語の訳語を可能な限り多く収集」
手順２「これらの訳語について、『-しきる』『-し尽くす』を付加してアスペクト焦点を判定」
手順３「格フレームの交代現象を文法性の判断によって評価して手順２の予測を検証」
手順４「対訳関係にある英日の動詞間で格フレーム交代のタイプの対応を評価」

手順１を実行した結果、１個以上の日本語動詞が訳語として得られた組み合わせは、複数の市販の英和辞典、英語シソーラス、日本語シソーラス、さらに１名の日本語母国語話者の知識を駆使した結果、127個見つかった。この手順は、本来ならば、伝統的な紙の辞書の電子版が十分揃っていれば、１名の日本語母国語話者がワークステーションに向かって各種資源からの検索結果をBootstrapping方式で積み上げていくことにより、より効率よく実行できることが期待される。

手順２、３では、日本語の格交代現象が、日本語母国語話者の文法性判断により抽出された。これらの大半は、おそらく今回初めてコーディングされた知識である可能性が高い。なぜなら、た

注6　あるいは、ULそのものを実装して語形成過程をシミュレートする機械を作るか、である。

注7　態の助動詞の付加による表層格シフトなども該当する。実装上の経済性のみならず言語理論のほうでも態の助動詞の付加は、辞書の内部で起こるのではなく、文法的導出によって起こるとする理論が優勢である[Kageyama93]。

とえば、10年前にMIT Lexicon Projectで先駆的に行われた研究報告によればFM&T85、３人の言語学者が２週間を費やして語彙調査をした結果、「日本語でFとGの両方がアスペクト焦点にくる動詞は高々３個くらいしか存在しないのではないか」と予想し、「この事実が英語と日本語の語彙システムの著しい違いを象徴」しているかもしれないとのコメントを付している。Gをアスペクト焦点とする動詞に至っては、FM&T85には全く収録されていない。では、果たして本当に英語のほうに、全体アスペクト的な動詞[注8]が日本語より多いのだろうか。その回答は、図３にまとめた結果の中に見ることができる。

	F & G (Eng)	G only (Eng)	F only (Eng)	sum.
F & G (Jap)	*16*	17	0	33
G only (Jap)	8	*16*	0	24
F only (Jap)	3	4	*63*	70
sum.	27	37	63	127

図３　実験結果　～２言語間の格フレーム交代タイプの一致／不一致

図３で、F & G (Jap) が、「日本語でFとGの両方がアスペクト焦点にくる動詞」を意味する。対角線上にある数字は、日英の対訳語どうしで同じ格フレーム交代タイプを共有する組み合わせをカウントしたものである。その総数は95で、交代タイプの異なる組み合わせの総数32のほぼ３倍にあたる。この数字は、互いに同義の概念は、やはり同じ言語概念素性の組み合わせをもつ可能性が大きいことを意味するだろう。その一方で、約１／３も存在する不一致は到底無視するわけには行かない(e.g. 飾る=decorateは、17個存在するF & G (Jap)=G only (Eng)という対応の一例)。この事実は、言語概念素性が語彙毎に恣意的に決まっている可能性を示唆する。もしそうであれば、ULにおけるモジュール化された概念表現は統語素性との依存関係において言語事実と合致すると結論できるだろう。

F & G (Jap) の数は、本実験で検証されただけでも33で、これは同じ意味カテゴリーにあった英語の語彙F & G (Eng) =27より大きい。またG onlyの動詞の数にも有為な差は認められない。ただ、英語でF onlyの語で同義の日本語の動詞がGの可能性を持つものが１つもみつかっていないことから、「日本語は英語に比べてアスペクト焦点がFになりやすい言語」というFM&T85のコメントはまだ意味を失っていない可能性はある。

実験の結果を総括すると、ULに裏付けられた、異言語の知識ベースによる高精度な統語素性の開発が決して荒唐無稽な試みではないことを示していると言って良いであろう。同じ試みがさらに大規模なスケールで検証され、高精度な計算辞書の構築に貢献することができれば、ULの言語普遍性の仮定はかなり真実味を増すと考えられる。

5　おわりに

個別言語の統語素性を言語普遍の要素から説明する、新しい計算辞書のモデルULを提案しその工学的メリットについて考察した。特に概念の類似性と言語普遍性が実際に統語素性のヴァリエーションを制約している可能性を検証するため、英語の格フレームセットをもとに日本語の格フレームセットを高精度で効率よく構築する実験を行った。この結果、ULの予想を裏付けるデータが得られた。今後もULの検証と応用を模索すると共に説明モデルを工学的応用に活かしていきたい。

参考文献

EDR89:「概念辞書　～第２版」より付録, EDR Technical Report No.13, 1989

江原＆田中93 : 江原暉将, 田中穂積,「機械翻訳における自然言語処理」, 情報処理 Vol.34, No.10, Oct., 1993

FM&T85: Fukui, N., S. Miyagawa, and C. Tenny. (1985). Verb Classes in English and Japanese: A Case Study in the Interaction of Syntax, Morphology and Semantics. , Lexicon Project Working Papers #3, MIT.

Jackendoff90: Jackendoff, R., Semantic Structures, Cambridge, MA: MIT Press, 1990

Kageyama93: Kageyama, T., Bunpoo-to Go-keisei (Grammar and Word Formation; in Japanese), Hitsuji-shoboo, 1993

Levin93 : Levin, Beth, English Verb Classes and Alternations, English Verb Classes and Alternations ~ A Preliminary Investigation" The University of Chicago Press. , 1993

野村89: 野村直之, 村木一至 :'機械翻訳システム PIVOTの格フレームモデル', 情報処理学会第38回大会予稿集, 1989

Nomura94a: Nomura, N. A Lexical Syntactic Analysis of Verb Alternation Types Changing, ms. , Workshop on "The Lexicon and the Computation," Cambridge, MA, April 30rd, 1994

Nomura94b: Nomura, N., Jones, D., & Berwick, R., An Architecture for a Universal Lexicon, in Proceedings of COLING94, 1994

岡田＆中村94: 岡田直之, 中村順一,「文法と辞書を作ろう」~自然言語処理入門V, 情報処理 Vol.35, No.3, Mar., 1994

[注8] Gに焦点があるとは、FとGが重なった*全体*の状態に焦点があるともみなせるから。またGは動作全体を通じて動かずにじっとしており F は動いていく。ゆえにFに焦点があるとは*動き*に焦点があるのと同等とみなせるかもしれない。