博士論文要旨

題　　名：A Computational Treatment of V-V Compounds in Japanese
指導教官：**郡司 隆男**
著　　者：**橋本 力**
提出年月：2004 **年** 11 **月**

本論文の目的は、日本語複合動詞の言語学的分析に基づいた計算機処理方法を開発することと、それを通じて、自然言語処理がなぜ、そして、いかにして理論言語学の成果を利用すべきなのかを論じることである。

1 章では理論言語学と自然言語処理の目的や方法論の違いを述べ、その上でなお、自然言語処理が言語学的手法を積極的に活用すべきことを論じている。その根拠は、言語学的に「深く」解析することにより意味表示が得られることと、従来の表層的自然言語処理技術が大きく依存していた統計情報への依存度を減らすことができることである。複合動詞は近年自然言語処理において重要課題として認識されつつある Multiword Expressions (MWEs) (Sag et al., 2002) の 1 つである。全ての複合動詞を一様に合成的に扱う方法や、逆に全てを区別無く一語として辞書に登録するような単純な方法では、理論的にも工学的にも満足な結果は得られない。前者の方法では非文法的な複合動詞を排除することも、ある複合動詞のイディオム的な側面を捉えることもできない。Sag et al. (2002) はこれらを overgeneration problem、idiomaticity problem を呼んでいる。後者の方法は flexibility problem と lexical proliferation problem に直面することが知られている。つまり、一部の複合動詞は柔軟、かつ、生産性合成性が高いため、全ての可能な組み合わせを辞書に列挙することは現実的ではなく、同時に、それらの意味や統語的振る舞いを適切に処理するのが困難になる。これらは複合動詞の計算機処理に言語学的知見が欠かせないことの根拠となる。

2 章では、橋田 (1997) による、言語理論が工学に応用され得るための 4 つの基準、現象の重要性 (理論が対象としている問題が工学的にも重要であるか)、設計の単純性 (理論がシステムの設計を単純にするか)、計算の容易性 (理論の予測を導くための計算が効率的に行えるか)、入力の利用可能性 (理論の参照する情報がシステムにも容易に入手可能であるか) についてまず述べている。次に、本論文の複合動詞の分析の基礎をなす、影山 (1993) と Matsumoto (1996) による分析を、橋田 (1997) の基準に照らしながら批判的に概観している。

影山 (1993) は統率・束縛理論に基づき、複合動詞を統語的複合動詞と語彙的複合動詞に区別し、さらに統語的複合動詞を、繰り上げ型、コントロール型、$\overline{\mathrm{V}}$ 補文型の 3 つに分類している。語彙的複合動詞に関して項構造に基づく原則を提案し、いくつかの例外に対しては逆形成に基づく分析と語彙概念構造に基づく分析を与えている。影山 (1993) の分析は複合動詞の MWEs 的側面を計算機上で扱うための有効な理論的基盤を与えてくれているが、いくつかの工学的問題を抱えている。統語的複合動詞の分析では、統率・束縛理論によって立つことから計算の容易性に違反する。つまり、統率・束縛理論、特に移動分析や空範疇分析には、厳密で形式的な数学的基盤が欠けており、効率的な処理技術が確立されていない。

語彙的複合動詞の分析では、例外の扱いが設計の単純性と現象の重要性に違反する。つまり、少数の例外に対して工学的コストが高い説明を与えている。

Matsumoto (1996) は語彙的複合動詞に関して、項構造に基づき、網羅的で示唆に富む観察を示している。Matsumoto (1996) は語彙的複合動詞を Pair compounds、Cause compounds、Manner compounds, Means compounds、Compounds exhibiting other relations, Compounds with semantically deverbalized V_2、Compounds with semantically deverbalized V_1 の 7 つに分類し、前項動詞と後項動詞の意味関係に基づいて詳細な意味論的分析を試みている。しかし、前項動詞と後項動詞の意味関係の認識には文脈や世界知識が大きく関わっており、計算機にそれを行わせることは容易ではない。つまり Matsumoto (1996) の分析は計算機に入手困難な情報を参照している点で橋田 (1997) の入力の利用可能性に違反する。また、Matsumoto (1996) が用いている緻密、かつ、個人間で判断が一致しにくいと思われる意味概念は、大規模な計算機用文法や辞書を構築する際に障害となり、設計の単純性に違反する。

一方、2 章を通して MWEs としての複合動詞の扱いの難しさが明らかにされており、計算機処理においても言語学的手法が不可欠であることが分かる。

3 章では本論文の分析が示されている。まず、橋田 (1997) の基準に沿って文法の設計方針を述べている。本分析では、現象の重要性の基準を満たすため、例外的現象や個人間で文法性判断が揺れるような現象の扱いにより規則が複雑化するようなことは避けた。設計の単純性を満たすため、理論的に高度であるより保守的な記述に努めた。また、計算機に与えることが可能な情報のみを参照するように規則を設計することで、入力の利用可能性を満たした。計算の容易性に関しては、文法の記述に $\mathcal{TDL}$ 言語 (Krieger & Schafer, 1994) を用いることで対応した。

次に本分析の枠組について述べている。本分析は計算機用大規模日本語文法 **JACY** (Siegel & Bender, 2002) の上に実装されている。そのため統語論的枠組は **Head-driven Phrase Structure Grammar** (**HPSG**, Sag and Wasow (1999))、意味論的枠組は **Minimal Recursion Semantics** (**MRS**, Copestake et al. (1999)) に依存している。文法開発には **LKB** (Copestake, 2002) を用いた。

本論文の統語的複合動詞の扱いは影山 (1993) に概ね従っており、A、B、C の 3 タイプに分類している (橋本, 2003)。特に、A、B タイプには、従来のほとんどの計算機用日本語文法では語順の問題のため扱われてこなかったが理論的には重要な、VP 埋め込み構造を採用している。言語学的な正確さを重視した結果、高精度な自然言語処理で重要になる緻密な意味表示を得ることに成功した。さらに本分析は、移動も空範疇も用いない簡潔で明示的な句構造分析であり、影山 (1993) と同等の理論的な説明力を持ちつつも、計算の容易性を満たしている。しかし一方で、VP 埋め込み構造はかき混ぜ現象の扱いを難しくする。JACY の枠組の中でかき混ぜ現象を適切に扱うために、本分析では **Argument Attraction** による分析を採用している。この分析によりかき混ぜを正しく扱え、かつ、適切に制限した結果、他の分析では得難い処理効率を実現している。

語彙的複合動詞は、Matsumoto (1996) を参考に、**Right headed V_1-V_2**、**Argument mixing V_1-V_2**、**V_1-V_2 with semantically deverbalized V_1**、**V_1-V_2 with semantically deverbalized V_2**、**Non-compositional V_1-V_2** に分類している。Matsumoto (1996) の Pair、Cause、Manner、Means の 4 種は Right headed V_1-V_2 と Argument mixing V_1-V_2 によりカバーされ、その 4 つの意味関係は文法においては未指定のままにしている。これは入力の利用可能性の基準から支持される。本分析は今泉郡司 (2000) の項構造に基づく簡潔な

ものであり、さらに高いカバー率を持つ。項構造は従来の大規模計算機用日本語文法では、複雑さを避けるため扱われて来なかったが、VP 埋め込み構造と同様、理論的な正確さを保つ上で不可欠である。これにより Matsumoto (1996) の分析では満たし難い設計の単純性を満たしつつも、同等の理論的説明力を持つことができる。さらに、統語的・意味的な基準に基づく分類により、統語的な生産性と意味的な合成性を適切に捉えることに成功している。意味表示は、統語的複合動詞と同様、緻密なものになっている。

3 章を通して、本分析が、橋田 (1997) の基準を満たしつつも、複合動詞の理論的あるいは MWEs 的な側面を適切に捉えていることが明らかになっている。特に、従来の計算機用日本語文法では扱えていなかった VP 埋め込み構造と項構造の導入が大きな役割を果たしている。

4 章では、本論文の実装の工学的観点からの評価実験について述べている。具体的には、本文法のカバー率、曖昧性の多寡、処理効率を、文法評価システム [incr tsdb()] (Oepen & Carroll, 2000) と **Lexeed** コーパス (笠原他, 2004) を用いて調べた。実験の際には本分析が実装されている JACY-vv とオリジナルのままの JACY-plain を用意して、カバー率、曖昧性の多寡、処理効率の差を調べた。JACY-vv には複合動詞の語彙項目は (non-compositional V_1-V_2 を除いて) 全く与えられていないが、複合動詞規則が実装されている。JACY-plain は 1,325 の複合動詞語彙項目が与えられているが、規則は実装されていない。結果として、JACY-vv は JACY-plain よりも高いカバー率と少ない曖昧性が得られることが判明した。カバー率の向上は主に、複合動詞の高度な生産性によるもの、言い替えれば、MWEs に関わる問題の一つ、lexical proliferation problem によるものだった。つまり、JACY-vv は未知の複合動詞が入力されても適切に解析することが可能だが、JACY-plain は膨大な数の複合動詞が辞書に登録されているにも関わらず、その生産性の高さには追い付けなかった。曖昧性の現象は VP 埋め込み構造に関わるかき混ぜ現象の扱いの違いによるものだった。つまり、本分析の Argument Attraction に基づく分析が正確、かつ、より適切に制限されていることが示されている。また、JACY-vv は言語学的分析により生産性・合成性の高いものとそうでないものを区別して扱っているため、規則による複合動詞の扱いが直面し得る MWEs に関わる問題、overgeneration problem と idiomaticity problem にも適切に対応できている。一方、処理効率では JACY-plain よりも劣っている。一般に、規則が増えれば処理効率は落ちてしまうのだが、本章では、規則の形式化を工夫することで改善し得ることが議論されている。

5 章では、これまでの議論をまとめた後、今後の課題と、言語学と自然言語処理のあるべき関係についての議論が述べられている。今後の課題として、本分析では捉え切れなかった現象をいかに扱うか、non-compositional V_1-V_2 をどうやって自動的に検出、収集するか、日本語複合動詞はどのように翻訳されるべきなのか、の 3 点を挙げている。本分析で扱い切れなかった複合動詞については、それらがどの程度生産的なのかをコーパスなどを通して見極め、新たな規則を追加するか、それらを 1 語として辞書に登録するかを検討すべきであることが述べられている。non-compositional V_1-V_2 の自動検出については、英語の句動詞の自動検出の研究を例に挙げ、それらの日本語複合動詞に対する適用可能性について議論している。言語学と自然言語処理の今後のあるべき関係については、自然言語処理の応用において言語学的手法では賄えない技術には何があるのかをまず述べている。最後に、自然言語処理が言語学の究極の目標にいかに貢献しうるかについて議論し、言語学的手法に基づく自然言語処理を提唱している。

参考文献

Copestake, A. (2002). *Implementing Typed Feature Structure Grammars*. CSLI Publications.

Copestake, A., Flickinger, D. P., & Sag, I. A. (1999). Minimal Recursion Semantics: An Introduction. Manuscript, Stanford University: CSLI.

Krieger, H.-U., & Schafer, U. (1994). $\mathcal{TDL}$ — A type description language for constraint-based grammars. In *Proceedings of the 15th International Conference on Computational Linguistics*.

Matsumoto, Y. (1996). *Complex Predicates in Japanese: A Syntactic and Semantic Study of the Notion 'Word'*. CSLI Publications.

Oepen, S., & Carroll, J. (2000). Performance profiling for grammar engineering. *Natural Language Engineering*, 81–97.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing: Third International Conference*, pp. 1–15 Mexico City, Mexico.

Sag, I. A., & Wasow, T. (1999). *Syntactic Theory: A Formal Introduction*. Center for the Study of Language and Information, Stanford. Japanese edition (two volumes) – translated and edited by Takao Gunji and Yasunari Harada, appeared in 2001.

Siegel, M., & Bender, E. M. (2002). Efficient Deep Processing of Japanese. In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization* Taipei, Taiwan.

橋田浩一 (1997).**「言語への情報科学的アプローチ」**. **大津由紀雄**, **郡司隆男**, **田窪行則**, **長尾真**, **橋田浩一**, **益岡隆志**, **松本裕治（編）**, **『言語の科学入門』**, 3 **章**. **岩波書店**.

今泉志奈子, **郡司隆男** (2000).**「語彙的複合における複合事象」**. **伊藤たかね**, **矢田部修一（編）**, **『レキシコンと統語』**, pp. 33–59. **ひつじ書房**.

影山太郎 (1993).**『文法と語形成』**. **ひつじ書房**.

笠原要, **佐藤浩史**, FrancisBond, **田中貴秋**, **藤田早苗**, **金杉友子**, **天野成昭** (2004).**「「基本語意味データベース**:Lexeed**」の構築」**.**『情報処理学会自然言語処理研究会報告** 2004-NL-159**』**, pp. 75–82 **東京**.

橋本力 (2003).**「長距離受け身の** HPSG **に基づく分析」**.**『日本言語学会第** 126 **回研究大会』**, pp. 256–261 **東京**.