談話構造コーパスの提案
A Proposal of Building Discourse Structure Corpus

奥村泰章[†]，白井英俊[‡]
Yasuaki Okumura, and Hidetosi Sirai

[†] 中京大学大学院 情報科学研究科，[‡] 中京大学　情報理工学部
Graduate School of Computer and Cognitive Sciences, Chukyo University

1. はじめに

文章の構造には、三段型、四段型など、文章の型として知られる大局的な談話構造と、文章を構成する命題としての文とそれに先行する文章との関係（白井(2007)[7]に従い、修辞関係と呼ぶ）によって造られる局所的な談話構造の少なくとも二通りの構造があると考えられる。大局的な構造は、その文章の主題の導入や説明、結論という意味単位を表す。それに対し、局所的な構造は、事象間の時間順序や原因結果、事実提示からの帰結、主張に対する証拠提示などの関係というような意味関係のみならず、対照や並列構造という命題の構造的な関係から構成される。そして、局所構造の大きな塊が大局構造の要素を構成するとみなすことができる。

文章の修辞関係が明らかになっていれば、その文章に現れた事象や事物間の関係を推論することが可能になる。また、Asher & Lascarides (2003)[1]で示されているように、語義の曖昧さ、省略や照応の解消、前提のスコープの曖昧さ、さらには会話の含意のような運用論的な問題の解決が容易になる。

このように局所的な談話構造の有用性は明らかであるようにみえるが、実際にはいろいろな問題がある。談話構造の既研究としては、Hobbs(1978)[3]、Grosz & Sidner(1986)[2]、安部ら(1994)[6]などがあるが、実際に言語資料に対してそれらで提案された修辞関係を適用するには困難が多い。RST(Mann & Thompson, 1987)[4]に基づくMarcu(1997)[5]の研究はその中で魅力的ではあるが、本稿では、Asher & Lascarides (2003)[1]で提唱された分節化談話表示理論(以下SDRT)に基づいた修辞関係を設計し、それに基づく談話構造コーパスを作成することを提案する。今回、SDRTを採用したのは、MDC原理を用いた最適な修辞構造の計算機構が用意されており、また依頼・命令や質問のように信念や意図によって説明されてきた「言語行為」に対しても修辞関係として捉えており、従来の談話構造の研究よりも適用範囲が広いからである。本研究では、小学校と中学校の国語教科書に用いられている文章に対して局所的談話構造の分析を行い、その分析によって明らかになったSDRTの問題点と、その対策、また得られた知見について述べる。

2. 作業

これまでに、小学校2年～中学校1年まで、各学年2～5個の計18個の文章を対象に人手による局所談話構造の付与を行った。作業者は二人で、修辞関係の分類は作業者の考えに基づいている。以下に小学校3年生国語教科書上の説明文に対して付与した例を示す。各文の先頭の角括弧の数字は段落番号と段落内での文番号を、文内の角括弧は文内部の各命題を表す。そして各文ごとに、そこで成り立つ修辞関係を記述する。

[2,1] アメリカに、ウイルソンという学者がいます。

[2,2] この人は、[次のような実験をして](1)、[ありの様子を観察しました](2)。

[2,1]->[2,2],記述-主題化(“ウィルソン” =”この人”)

[2,2-1]->[2,2-2],手段-目的

[3,1] 初めに、ありの巣から少し離れた所に、一つまみの砂糖を置きました。

[2.2]->[3,1],詳細化-事象

[3,2] しばらくすると、一匹のありが、その砂糖を見つけました。

[3,1]->[3,2],時間経過-可能化

例えば[2.2]文目に以下のような修辞関係の存在が記されているが、

[2,1]->[2,2],記述-主題化(“ウィルソン” =”この人”)

これは、文[2,2]は文[2,1]で導入された”ウィルソン”を主題とした記述文であることを表している。この「記述」という関係は、結合している二つの命題間において、その下位情報により命題中の要素の一貫性を表す。この情報は、まだ解決されていない照応や、語の多義性、橋渡しの解決に利用できる。

なお、実際の作業では、その修辞関係の同定の根拠(たとえば、[2,2-1]が[2,2-2]の手段であることは、接続助詞「て」とアスペクトが関係している)も記述する。

3. 分析

現在までの分析で、Asher & Lascarides(2003)[1]で定義された修辞関係では十分分類できないということが明らかになった。例えば、例で挙げたような「記述」といった関係はSDRTでは定義されていないが、「同じものについての言及」のような、文の緩やかな結びつきを表現するために必要である。そのため修辞関係を追加するために、分析に基づいて修辞関係の整理を行い、追加すべき修辞関係を選定している。また、どのように修辞関係を推定するかという規則も必ずしも明確でないため、人間が修辞関係の同定に用いている情報を分析、整理している。この情報には、接続詞、接続助詞などの手がかり語句(たとえば、[3,2]の「しばらくすると」)や、述語となる動詞の時制およびアスペクト性(たとえば、[3,2]の動詞「見つけました」のアスペクト性)などが含まれる。このような関係を導く情報を収集し、体系化し、より大規模に談話構造のコーパス作成を行うことを考えている。コーパスの記述法として、橋田らの提唱している大域文書修飾(GDA)[8]などを利用することができる。特に、会話のような話し言葉の談話を対象としたコーパス作成や、機械学習を用いたコーパスの充実を考えている。さらにこのコーパスにより、既研究では因果関係に限定されているような事象間のより広い関係を収集するシステムの構築も視野に入れている。

4. 考察

分析を通して、教科書の文章の修辞関係に関するいくつかの性質が見えてきた。モダリティはほとんど修辞関係に寄与していないこと、低学年の文章では「すると」や「その結果」といった関係を導く手がかりとなる語が多く現れていたのに対し、高学年の文章ではそういった手がかり語句が文中に現れず、述語となる概念に関する知識のみで修辞関係を推定するようなパターンが多いことなどである。また修辞関係同定に用いられる概念間の知識についていえば、低学年では「(花が)咲く」と「閉じる」のように背反事象である関係が用いられるのに対し、高学年では「爆風が市街を襲う」の結果、「市民の命が奪われる」というようなより多くの推論を必要とするものが多いことが明らかになった。このことから、低学年では推論が必要な場合に手がかり語句によって修辞関係を推定させ、それにより概念間の関係を学習し、高学年の文章の読解に用意するという教育方針のようなものが見えてきたことは興味深い。事象間の関係を収集するシステムを構築する際に、より精度の高い学習能力を持たせるために、このような人間の学習過程の知識を利用できると考えている。

参考文献

[1] Asher, N. & Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.

[2] Grosz, B. & Sidner, C. (1986). Attention, Intentions and the Structure of Discourse. *Computational Linguistic*, 12, 175-204.

[3] Hobbs, J. R. (1978) Coherance and Coreference. *Cognitive Science*, 3(1), 67-90.

[4] Mann, W, C. & Tompson, S, A. (1987). Rhetorical Structure Theory: Description and Construction of Text Structures.

[5] Marcu, D. (1997). The Rhetorical Parsing, Summarization, and Generation of Natural Language Text.

[6] 安部純一,桃内佳雄,金子康朗,李光五 (1994). 人間の言語情報処理,サイエンス社.

[7] 白井英俊 (2007). 談話と論理-分節化談話表示理論の紹介- 人工知能学会誌 22(5) 621-629.

[8] 橋田浩一, GDA 日本語アノテーションマニュアル, http://www.i-content.org/gda/tagman.html .