文タイプと文間関係の情報を付与したテキストコーパスの作成

黒橋 禎夫　木下 恭子　山田 悟史　長尾 真

京都大学 工学研究科　京都大学 人間環境学研究科　京都大学 工学部　京都大学

1 はじめに

文ではなく文章を単位として，実用に耐えうる言語処理技術を確立する必要がある．文章の解析として，本稿では各文の性質(タイプ)を明らかにする問題と，文間の関係を明らかにする問題を議論する．

文を単位とする形態素，構文解析などでは，コーパスに正しい情報を人手で与え，それをもとにして解析システムを改良することが行われてきた．このような方法は，コーパス中の実際の言語現象を広く吸収することができ，人間の言語直感だけに頼る場合に比べてはるかに頑健なシステムを構築することができる．

同じことは文章の解析にもあてはまるはずである．すなわち，コーパスに文タイプ・文間関係の情報を付与し，その情報に基づいてシステムを構築・改良するのである．ところが，形態素解析・構文解析などの場合に比べて，文タイプ・文間関係などの場合には与えるべき情報が必ずしも明確でないという問題がある．文タイプ・文間関係を整理・分類したこれまでの研究の多くは，典型的な場合を切り出して説明したもので，典型的でない様々な表現を扱う場合，またひと続きの文章全体の現象を扱う場合には必ずしも満足できるものではない．

そこで我々は，実際のテキストへの文タイプ・文間関係の情報付与(タギング)を行いながら，それらの妥当な分類はどのようなものであるかを試行錯誤的に検討した．本稿では，毎日新聞の社説20文章にタギングを行った現時点での文タイプ・文間関係についての我々の分類体系を説明する．

2 文のタイプ

文の構造は，客観的出来事や事柄を表す部分と，話手の判断・態度などを表す部分にわけて考えることができる．前者は事態などとよばれ，後者の主要部分はモダリティとよばれるものである．

事態については，テンス・アスペクトの観点からの多くの研究があるが，実際の文章を扱うと，それよりも無時制で属性(性質・特徴)を示すものの区別，名詞述語文の扱いなどが問題であることがわかった．そこで，事態に関しては次のような分類情報をコーパスに付与することとした．

一般的事態

動態 (例: ～は～に特別拠出を求めた)
継続状態 (例: 食糧庁は～などを行ってきた)
結果状態 (例: 現在は～の定員となっている)
状態変化 (例: ～の役割は大幅に縮小する)
状態 (例: ～の議論が十分でない)
……

属性(性質・特徴) (例: この加速器は～を利用する / 日本の～技術は世界一である)

指定 (例: 理由は～である / 一つは～だ)

一般的事態の状態と属性との区別は難しい場合が少なくない．また，指定についても「理由は～」などの典型的な場合はよいが，それ以外にどのようなものまで指定とみなすかという点がまだ明確ではない．これらは，今後の検討課題である．

モダリティについては，従来の研究では十数個の1レベルの分類を与えたものが多いが[1]，それではコーパスに一貫性のある情報を与えることが，特に典型的な表現でない場合に難しい．そこで，モダリィを大きく3つに分類し，それぞれについてさらに細分類を与えた．

話者の意思に関するもの

意志 (～しようと思う) / 申し出 (～ましょう) / 願望 (～たい / ～てほしい) / 勧誘 (～ましょう) / 命令 / 禁止 / 依頼 (～てください)

当為：一般的常識に照した事態の必要性，望ましさ等

当然 (〜ものだ) / 義務 (〜なければならない) / 必要 (〜べきだ / 〜する必要がある) / 適切 (〜方がよい / 〜がのぞましい)

話者の判断の確らしさに関するもの

推測 (〜だろう / 〜はずだ) / 可能性 (〜かもしれない) / 様態 (〜そうだ / 〜ようだ) / 伝聞 (〜そうだ / 〜という)

モダリティは典型的な表現だけでなく，様々な表現で表される．それらについては，典型的な表現へのパラフレーズを考えるという方法で上記のいずれかの分類を当てはめた．たとえば「〜があってもいいのではないか」という表現は「〜方がよい」とパラフレーズして考え「適切」とみなす．

コーパスへのタギングを通して，上記の3つの大分類はほぼ妥当なものと考えられた．しかし，細分類については，このような離散的な分類ではなくアナログ的な分類，例えば「当為」であれば必要度を強弱で表すような方法が適当ではないかと思われる場合があった．そのような検討は今度の課題である．

3 文間の関係

文間の関係についてもこれまでに多くの研究があり [2, 3]，各研究者がそれぞれ数個から十数個程度の関係を定義している．それらの関係を用いて実際にコーパスのタギングを行ってみると，ある2文の間に従来定義されてきた複数の関係が認められることが少なくなかった．

そこで，どのような関係が共存しうるかということを考慮して文間関係を整理し直し，結果的に次のような分類をえた．すなわち，まず (広義の) 同格の関係とそうでないものに分類し，同格についてはさらに細分類を与えた．一方，同格でない場合は4つの観点 (属性) を考え，それらの組み合わせによって文間関係を表現することとした．

同格 — 同格 / 詳細化 / 要約 / 例示

同格以外

時間経過 — ＋ / − / なし

因果関係 — 原因結果 / 結果原因 / 逆接 / なし

等位性 — 累加 / 対比 / 選択 / なし

主題等の継続性 — 主題連続，主題－焦点連鎖など13パターンに分類

たとえば，次の2文間の関係は { 時間経過：なし，因果関係：なし，等位性：累加，主題：主題連続 } となる．

> 欧州のLHCは水素の原子核である陽子同士を光速に近い速度で衝突させ，宇宙誕生直後の高エネルギー状態を作り出す．
> そして物質に重さを与える役割を担うヒッグス粒子の発見などを目指している．

文間関係は基本的には1文対1文とした．ただし，ある文がそれ以前の一定範囲の文と関係をもつ場合はその範囲を指定した (要約の場合など)．また，ある一文がいくつかの文に対してそれぞれ関係を持つと考えられる場合には，それらをすべてタギングすることとした．また，特に段落の先頭の文では，前のいずれかの文と関係を持つというよりは，文章全体の主題に結び付くだけである，という場合が少なくない．そのような場合には文間関係なしとした．

4 おわりに

現時点の分類体系には細かい点で様々な問題があるが，それらはタギングを行いながら問題箇所にコメントを残し，ある程度の規模のコーパスとなってから再検討する予定である．また，作成したコーパス，分類体系の詳細を書いたタギングのマニュアル等は整備できた時点で公開する予定である．

なお，本研究は日本学術振興会未来開拓学術研究推進事業 (JSPS-RFTF96P00502) の助成を受けている．

参考文献

[1] 森田良行，松木正恵：『日本語表現文型』，株式会社アルク (1989).

[2] 永野 賢：『文章論総説』，朝倉書店 (1986).

[3] Eduard H. Hovy : Automated discourse generation using discourse structure relations, Artificial Intelligence 63 (1993).