# 多言語平行コーパスのための言語横断的な構造記述

山崎直樹
関西大学外国語学部
ymzknk@kansai-u.ac.jp

この研究の目的は，言語類型論的研究の資料として使うことを前提とし，統語情報をメタ情報として付与した多言語平行コーパスを制作する[(i)]ために，言語横断的構造記述の枠組みを開発することである。ただ，「言語横断的枠組みの開発」は，非常に遠大な目標であるので，さしあたり，日本語と中国語を同一のフォーマットで記述することを目指した枠組みを考え，その一端を紹介して，諸賢の批判を仰ぎたい。

統語情報を付与したコーパスとしては，*Penn Treebank Project*[(ii)]によるものが，すでに存在する。ここで使われている記述方法は，汎用性のあるものだが（(1)参照），ここにある情報（主に文法機能範疇に関する情報）以上に詳細な情報を付与したいと考えても，それを反映できる仕組みをもたない。

(1) (NP (CP e-1 (NT 昨天) e-2 (VV 来) (DEC 的))
(NP (DT 那) (M 个) (NN 人)))

また，機械処理を前提とした言語学的情報付与の枠組みとして，*Global Document Annotation (GDA)* [(iii)]というフォーマットも開発されている。これは，XML というマークアップ用言語を用いており，非常に多くの情報を付与することができ（(2)参照），なおかつ拡張も可能である。

(2) <su><n id="Z">象</n><ad opr="topic">は</ad>
<adp><np arg="Z">鼻</np><ad opr="obj">が</ad></adp><aj>長い</aj></su>

ただ，GDA は，自然言語を機械処理する前段階として，入力となる自然言語の文の統語・意味情報を機械可読な形で提供することを目的とした記述方法である。よって，このままでは，本研究の目的（言語類型論的研究のための資料を作成する）にはそぐわない面もある。

本研究では，GDA を参考にしつつ，以下の諸点に特に留意して，記述の枠組みを考えたい。

(a) 客観的に検証可能な「統語現象」だけを記述する。例えば，“我明天城里有事”を[我[明天[城里[有事]]]]と階層化するのは，意味解釈を反映させるための措置であると考え，このような階層構造を記述に採用しない。

(b) 句レベルの範疇の終端記号を，語彙範疇の最大投射（例：NP, VP…）にはしない。文の中での機能は同じであっても後置詞の有無で終端の名称が異なったり，時点や場所を表すフレーズの名称の判断に迷ったりする枠組みは，統語構造の対照研究に使いにくいからである。例えば，「お店に行って買い物をする」と“去商店买东西”はどちらも動詞句が連続しているが，前者は，最初の動詞句に接続のための後置詞があるという点で後者と異なる。このような共通点と相違点を明示できるようにする。

(c) 「文」を特定の語彙範疇の最大投射と考えない。用言性の語彙（動詞，形容詞など）を述部に含まない構造をもつ言語は多いからである。

(d) 範疇間の連続性を損なわないような語彙範疇の分類法を採用する。中国語では，ある種の動詞と側置詞（中国語の場合は前置詞）の間には連続性がある。これに，「側置詞」というラベルを与えると，その連続性を分断してしまうことになるので，できる限り離散的な分類は避ける。いっぽう，日本語の側置詞（日本語の場合は後置詞）は，他の範疇との連続性がない。日本語の側置詞と，中国語の側置詞的な働きをする語彙との共通点と相違点を示せるような枠組みを考えたい。また，言語によっては，動詞と形容詞の区別がなかったり，名詞と形容詞の区別がなかったりする。そのような差異を反映する枠組みを考えたい。

(e) 冠詞や決定詞の使用が義務的でない言語のための照応表示の枠組みを考える。「この樹は，花がみっともないね」という文の中の「花」は，形式的な標識は何もないが，先行する「樹」と何らかの関係をもつ。このような広義での「照応」（日本語，中国語を始め，このタイプの言語は多い）を表示する手段を考える。

**注**

i) この作業は，大阪外国語大学（現在の大阪大学外国語学部）で始まった『多言語処理プロジェクト』の活動の一環である。本研究の成果も，ここでの議論に負うところが大きいことを付記しておきたい。

ii) Penn Chinese Treebank Project: http://www.cis.upenn.edu/~chinese/

iii) *Global Document Annotation (GDA)*: http://www.i-content.org/gda/

**キーワード**：多言語平行コーパス，統語情報付きコーパス，通言語的文法記述，汎用タグセット，言語類型論