京都大学テキストコーパス・プロジェクト

黒橋 禎夫　　長尾 眞

京都大学大学院工学研究科 電子通信工学専攻
{kuro, nagao}@kuee.kyoto-u.ac.jp

1 はじめに

大量の文章に種々の言語情報を付与したテキストコーパスの重要性は広く認識されており，様々な機関でテキストコーパス作成のプロジェクトが行なわれてきた．コーパスに付与する情報が深いレベルであればあるほど，それを用いて行なえること(文法の自動獲得や用例ベースの解析など)も深いレベルとなる．しかし，本稿ではコーパスに付与すべき最も基本的な情報である形態素・構文情報に限定して話をすすめることにする．

コーパスに形態素・構文情報を付与する一般的な方法は，まず何らかの自然言語処理システムによってそれらの情報の自動付与を行ない，次にそれを人間がチェックして正しい情報に修正していくというものである(この作業者を**アノテータ**とよぶ)．これまでのコーパス作成プロジェクトでは，情報の自動付与の仕方について2つの場合があった．

- 解析システムは確実な名詞句，副詞句などをまとめるだけで，曖昧性の生じる部分については何も解析を行なわない．アノテータの仕事は自動解析結果の部分構造をまとめあげることとなる(Penn Treebankプロジェクト[1]など)．

- 解析システムは曖昧性のある部分についてはあらゆる可能性を保持しながら，文全体の構造を求める．解析結果はいわゆる統語森(parse forest)の構造となる．アノテータの仕事はその中から正しい解を選択することとなる(ATR/Lancaster Treebankプロジェクト[2]など)．

このような場合，いずれにしてもアノテーターの負担は重く，テキストコーパス作成は非常にコスト(費用と時間)のかかるプロジェクトであった．

これに対して，京都大学では日本語形態素解析システムJUMAN[3]，日本語構文解析システムKNP[4]を開発・公開し，その修正を継続的に行なってきた．その結果，これらのシステムによって，現実のテキストに対して一意の解を求める形態素・構文解析がかなりの精度で可能となってきた．妥当な解が一意に得られれば，人手による修正が比較的簡単になるだけでなく，その人手修正過程がシステムの問題点の調査に直接つながることになる．

そこで，我々のコーパス作成プロジェクトでは，これまでのプロジェクトとは根本的に異なる次のような立場をとることにした．すなわち，大量の日本語文章の形態素・構文解析，その人手によるチェック・修正を通じて解析システムそのものの改良を徹底的に行ない，その成果として

1. 高精度な日本語解析システム
2. 正しい形態素・構文情報が付与された日本語文章データ

の両方を公開することをプロジェクトの目的とした．

2 コーパスに付与する情報

コーパスに付与する形態素・構文情報は解析システムJUMAN, KNPに準拠した以下の情報である．

形態素情報

- 形態素の区切り
- 形態素単位の品詞，読み，原形，活用型，活用形

```
* 0 1D
沸点      ふってん  沸点  名詞  普通名詞  *                  *
が        が        が    助詞  格助詞    *                  *
* 1 4P
高い      たかい    高い  形容詞  *       イ形容詞アウオ段   基本形
、        、        、    特殊    読点    *                  *
* 2 3D
多く      おおく    多く  副詞    *       *                  *
の        の        の    助詞    名詞接続助詞 *             *
* 3 4D
物質      ぶっしつ  物質  名詞    普通名詞  *                *
を        を        を    助詞    格助詞    *                *
* 4 10A
溶かす    とかす    溶かす 動詞   *       子音動詞サ行       基本形
など      など      など  助詞    副助詞    *                *
、        、        、    特殊    読点      *                *
* 5 10D
水        みず      水    名詞    普通名詞  *                *
が        が        が    助詞    格助詞    *                *
⋮         ⋮         ⋮
```

図 1: 形態素情報，構文情報を付与した文データの形式

構文情報

- 文節の区切り
- 文節間の係り受け[1] (以下の 3 つのタイプを区別)
  - 通常の係り受け関係
    (格要素 – 用言間，修飾語 – 非修飾語間)
  - 並列関係 (「本と鉛筆」など)
  - 同格関係 (「A 社社長，太郎」など)

このような情報を付与した文データの一例を図 1に示す．各行は一つの形態素の情報である (最初の 3 列は表記，読み，原形)．* で始まる行は文節の区切りを示し，* の次の数字は文節 (次の * 行までの形態素列) の番号，2 番目の数字はその文節の係り先の文節番号を示す．2 番目の数字に続く記号 D,P,A によって通常の係り受け関係，並列関係，同格関係を区別している．

3 解析結果修正用インターフェース

前節で述べた情報はまず解析システム JUMAN, KNP で自動的に付与し，その結果をアノテータが手作業によって修正する．この修正作業にはマウスベースの修正インターフェースを用いる (図 2)．

修正は一文単位で行なう．各文には ID，解析の日付け，修正の日付けなどが自動付与され，また，任意のメモが書き込めるようになっている．

インターフェースの上部は係り受け構造を修正する画面で，その右上領域はすべてボタンになっている．ここで適当なボタンを選択することにより，その左側と下側に位置する文節間の係り受け関係を入力することができる．係り受けのタイプ (通常の係り受け関係，並列関係，同格関係) はマウスのボタン (左，真中，右) によって区別される．修正した結果が係り受けの非交差条件を破る場合には，その部分の色が変わり，注意を促すようになっている．

インターフェースの下部は文節区切りと形態素解析結果を修正する画面である．この部分では，品詞ごとに可能な活用型だけを表示する機能，活用型の修正が原形に自動的に反映される機能などを用意し，できるだけアノテータの負担を軽減するようにしている．

4 これまでの経過

4.1 形態素解析システム JUMAN の修正

プロジェクトは実質的に 96 年 1 月にスタートした (図 3)．プロジェクトのはじめの半年間は JUMAN に

[1] 構文情報は文節間の係り受けとして表現するが，名詞句，動詞句などのいわゆる句構造としての標識は係り受けの情報から 一意に判別可能である．たとえば，ある文節の自立語が名詞であれば，係り受け構造 (依存構造木) においてその文節を根とする任意の部分木は 名詞句と解釈できる．

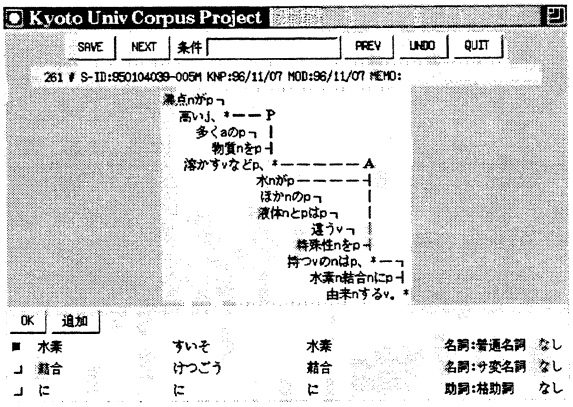

図 2: 解析結果修正用インターフェース

対する以下のような修正を行なった.

- 連語処理機能の追加 [5]
- 標準形態素辞書の整備
- 標準文法辞書の整備

これらの成果は JUMAN3.0beta として 96 年 6 月に公開した.

JUMAN の修正はその後も継続的に行なっており (主に具体的な連語の辞書登録), まとまりがつくごとに新バージョンの公開を続けている (96 年 10 月 JUMAN3.0, 96 年 11 月 JUMAN3.1).

現在の JUMAN の解析精度は, 新聞記事を対象とした場合, 形態素単位で 99.0% 前後である (区切りと品詞付与 (品詞細分類は無視) の双方が正しい場合を正解とした).

4.2 構文解析システム KNP の修正

JUMAN の修正作業が一段落した 96 年 6 月頃から構文解析システム KNP の修正を本格的に開始した. まず行なったことは,

- 文法記述の枠組の修正

である. 従来の KNP では優先規則のかなりの部分が手続き的に記述されていた. しかし, 大量の文を解析しながらシステムの改良を継続的に行なっていくためには, 優先解釈規則を含めて文法を宣言的なものにしておく必要がある. そこで, システムの文法記述形式を抜本的に修正し, 文法をほぼ完全に宣言的に記述できるようにした. 新しいシステムでは, 形態素, 文節に対してパターンマッチング的に属性を与え, その属性を用いて係り受け関係の規則を記述する.

これに続いて, これまでに次のような修正を行なってきた.

- 並列構造解析の強化
- 従属節のスコープ解析の強化
- 用言に準ずる種々の表現への対応

これらの修正については別の機会に詳しく発表する. KNP のシステム的改良, 文法の整備は, コーパスの手修正作業と並行して現在も継続的に行なっている.

現在の KNP の解析精度は, 新聞記事文を対象とした場合, 文節単位, すなわち各文節に対して正しい係り先が特定できたかどうかという基準で 90% 前後である (文末の 2 文節はカウントしていない). これは, おおざっぱにいえば, 12 文節程度の平均的な長さの文であれば, その中で一箇所だけ係り先の誤りがあるという精度である. 97 年 3 月中には, これまでの成果をまとめて KNP2.0 として公開する予定である.

4.3 コーパスの人手修正

96 年 9 月頃から, 人手によるコーパス作成作業 (解析結果の人手修正) を始めた. はじめの数ヶ月間は, 新聞記事 3000 文を対象として, コーパスに付与する形態素・構文情報の基準作りを行ない, ドキュメントを作成した.

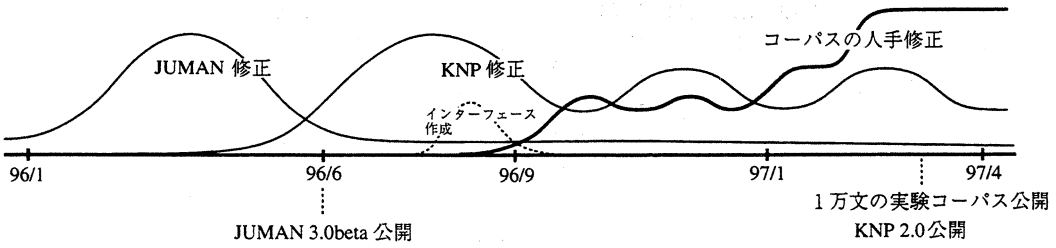

図 3: プロジェクトの経過

定常的な人手修正は，基準がほぼ確定した97年1月頃から開始した．はじめはアノテータ1名，2月からはアノテータ2名で作業を行なってきた．修正作業のペースは1時間あたり約50文(解析誤りに対する分類・整理を同時に行なっている)，一日の修正作業時間は約5時間である．この後，各アノテータはその日修正した解析誤りに対する原因の調査(文法・辞書の不備など)，解決策の検討を行なう．これらの調査・検討に基づいてほぼ一週間ごとにプロジェクト全体のミーティングをもち，そこで実際にシステムの文法・辞書の修正を行なっている．

これまでに，毎日新聞96年データを元テキストとして，約2万文の人手修正済みコーパスを作成した．このうち1万文を実験コーパスとして97年3月中に公開する予定である(ただし，公開するのは京都大学で付与した情報のみで，利用者は毎日新聞CD-ROM 96年版を別途購入する必要がある)．

今後，97年度についても，ほぼ同じ体制でプロジェクトを継続する予定である．コーパスの目標規模は約20万文(Penn Treebankと同程度)である．対象テキストは新聞記事に限定せず，一般教養書(新書など)，科学技術文なども扱っていく予定である．

5 おわりに

本稿では京都大学におけるテキストコーパス作成プロジェクトの概要を述べた．コーパスについては随時，JUMAN，KNPについては修正が一段落するごとに公開していく予定である．これらの情報には長尾研究室のホームページ[2]からアクセスできる．

公開したコーパス，システムについては，できるだけ多くの方々に利用して頂き，種々の指摘を受けて，さらに改良を行なう，という相互関係が生まれることを願っている．

なお，本プロジェクトは95年度，96年度については文部省科学研究費補助金(課題番号07558046)の助成を受けた．97年度については日本学術振興会未来開拓プロジェクトの助成を受ける予定である．

[2] http://www-nagao.kuee.kyoto-u.ac.jp/

参考文献

[1] Marcus, P., Santorini, B., Marcinkiewicz, M. : Building a large annotated corpus of English: the Penn Treebank, Computational Linguistics, 19(2) (1993).

[2] Black, E. et al. : Beyond skeleton parsing: producing a comprehensive large-scale general-English treebank with full grammatical analysis, COLING'96 (1996).

[3] 松本裕治，黒橋禎夫，宇津呂武仁，妙木裕，長尾眞： 日本語形態素解析システムJUMAN 使用説明書，京都大学工学部 長尾研究室 (1992).

[4] 黒橋禎夫，長尾眞： 並列構造の検出に基づく長い日本語文の構文解析，自然言語処理 Vol.1 No.1 (1994).

[5] 山地治，黒橋禎夫，長尾眞： 連語登録による形態素解析システムJUMANの精度向上， 言語処理学会 第2回年次大会 (1996).