2N－5

EDR形式の日本語コーパスを対象とする編集支援システム*

潮 靖之‡　本間 健一§　上原 徹三‡　石川 知雄‡

武蔵工業大学工学部‡　キヤノンソフトウェア株式会社§

1　はじめに

自然言語処理研究において，コーパス，特に解析済みコーパスは，実際の例文とその文法情報を与えるという意味で重要である．その中でも，EDR日本語コーパス [1] は約22万文の例文とその解析情報を持つ最大規模の電子化日本語コーパスである．

しかし，研究対象とするためにはコーパスの例文は更に大量に必要であり，また，解析情報を誤りなく作成し保守することが困難であるなどの問題がある．そこで，既存コーパスに対する編集や例文の追加，及び誤りのない新規コーパスの効率的作成を実現するための支援システムが望まれる．

現在，EDR電子化辞書を編集するシステムとして，日本電子化辞書研究所よりEDR辞書管理システム [2] が公表されている．このシステムでは，検索処理全てにインデックスを用いることによって検索を速くし，多くの辞書を統合して管理できる特長を持つ．しかし，ユーザーはEDR形式の格納形式と意味の理解が必要であり，またシステムの機能をコマンド形式で呼び出さなければならない．

そこで，EDR形式の日本語コーパスを対象として，内部構造の知識が不要で，メニュー形式の例文編集追加機能を持つ編集支援システムを試作した．

2　本システムの対象

本システムの対象は，EDR日本語コーパスだけでなくEDR形式に従うコーパスであればよい．本システムでは文構造を示す構文情報を文節間係り受け関係として捉える (図1の [a])．これは，EDR日本語コーパス本来の形式 (図1の [b]) とは異なる．従って，既存のEDR日本語コーパスの例文の一部を更新追加すると，両形式の例文が混在することがある．但し，EDR日本語コーパスの構文構造を本システムの文節間係り受け構造として読むことは可能である．(EDR形式の構文情報から文節間係り受け関係の構文情報への変換について言及した報告 [3] もあるが，ここでは辞書データは変換しない.)

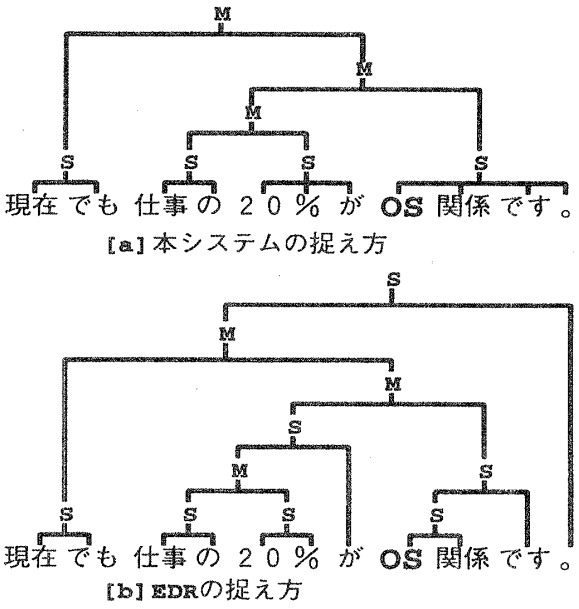

図 1: 構文情報の表示例

3　システムの主要な機能

本システムの主要な機能は次のようなものである．

3.1　検索機能

本システムで提供している検索機能では，品詞と表記による検索，レコード番号による検索，係り受けによる検索などが行なえる．また絞り込み検索や検索範囲を指定した検索も行なえる．

検索範囲を指定したり，レコード番号により高速に検索するために，コーパスに対してインデックスを作成している．インデックスはレコード番号をキーとする $B^{+}Tree$ で作成してある．このインデックスは例文追加や例文更新機能にも対応している．

3.2　表示機能

検索結果の表示は，3段階の表示機能より成る．第1段階では検索該当例文数などの結果を示す．第2段階でそれぞれの例文とレコード番号などの情報を表示す

*Edit System for EDR style Japanese Corpus

†Yasuyuki USHIO, Kenichi HONMA and Tetuzou UEHARA and Tomoo ISHIKAWA

‡Faculty of Engineering, Musashi Institute of Technology

§Canon Software Inc.

る．キーワードがある場合には，KWIC(KeyWord In Context) 方式により表示する．第3段階でそれぞれの文の構文情報を含めた詳細情報を表示する．構文情報は上述のように，文節間係り受け関係として捉えるので，図2に示すような係り受けの木で表示する．

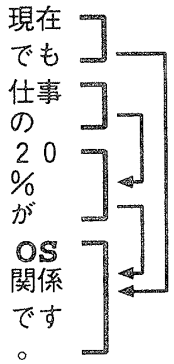

図 2: 構文情報の表示例

3.3 更新機能

更新機能は，既存コーパス中の例文の内容を更新 (修正) することができる．品詞の修正や係り受けの木の表示による係り受け関係の修正も行なえる．後者は，EDR コーパス中の構文情報の括弧による入れ子構造に反映することで実現する．図3の [a] に更新前の構文木を [b] に更新後の構文木を示す．これは，1 文節目が 2 文節に係っていたのを 3 文節目に係るように修正した例である．ユーザーは変更する文節の番号の 1 と係り先の 3 を指定すればこのような処理を行なうことができる．

4 コーパスの内部状態

コーパス (データファイル) とインデックスは別ファイルになっている．データファイルに関しては，コーパスのレコードは可変長なので，更新したデータを前と同じところに格納することはできない．本システムはデータの更新時に，更新前のデータにフラグを付けて，更新後のデータはデータファイルの最後尾に更新前のレコードの位置と共に格納する方法をとっている．すなわち，データファイルはヒープ構成をとっている．これにより，既存の EDR コーパスのファイルをそのままの形で編集対象とすることができる．

インデックスファイルは，データを更新または追加するごとに更新していく． 22 万文の例文に対する $B^{+}Tree$ の段数は 8 段で作成している．

以上のファイル構成により，レコード番号による検索の効率化を実現すると共に更新レコードの復元等のアンドゥ処理に対処できる．編集作業中のデータファイルは無効レコードを含むが，編集作業が完了したときに，データファイルを整形すれば，EDR 形式そのままのコーパスを作成することができる．

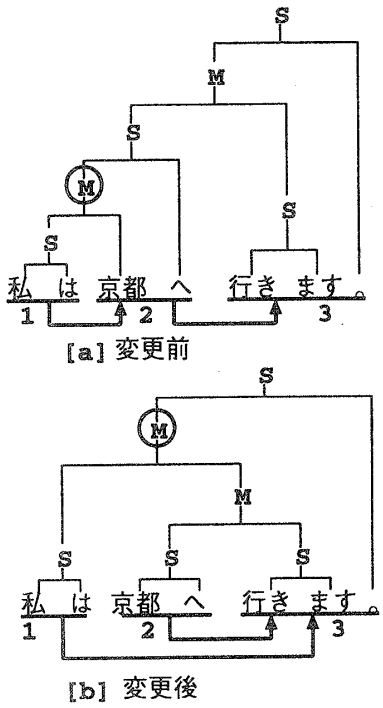

図 3: 構文情報の変更

5 おわりに

EDR 形式の日本語コーパスを対象とする編集支援システムの機能と実現法について述べた．今後は，単語の表記や品詞をキーとしたインデックスを作成して，それらによる検索速度の高速化を行う．また，これを用いて，例文の追加や新規コーパスの作成を支援する機能を作成する予定である．

参考文献

[1] 日本電子化辞書研究所．「EDR 電子化辞書仕様説明書 第 2.0 版」.(1996)

[2] 日本電子化辞書研究所．「EDR 辞書管理システム」.(1997)

[3] 植木, 白井, 徳永, 田中．「構造つきコーパスの共有化に関する一考察」．情報処理学会 自然言語処理 128-9 pp.61-66,11 1998.