大規模英語単語辞書の開発

6F−5

有岡 昌子, 石渡 裕美, 河田 康裕, 末松 博, 原田 千秋, 天野 真家
(株) 日本電子化辞書研究所

1 はじめに

(株) 日本電子化辞書研究所 (略称ＥＤＲ) では、自然言語処理用の大規模電子化辞書の開発を行っている。ＥＤＲ電子化辞書は、単語辞書、概念辞書、共起辞書、対訳辞書の4種類の辞書から構成されており、日英二ヶ国語を対象としている [1]。このうち単語辞書は、文の統語構造の解析や生成に用いる語の文法的特性及びその語が表す概念を記述したものであり、現在基本語 (日英各20万語) の辞書データ品質向上のための作業を進めている。本稿では英語単語辞書の開発法と開発の各工程における問題点について述べる。

2 辞書開発の各段階における問題点と方策

大規模電子化辞書の開発には、レキシコグラファのもつ専門知識を大量の語彙を対象にいかに効率的に記述し、しかも高い品質をうるかという課題がある。そこでＥＤＲ単語辞書は図1に示す工程にしたがって、専門家による記述を計算機により支援しながら辞書開発を進めている。以下、英語単語辞書開発の各工程において生じた問題とそれに対して行った方策について述べる。

2.1 記述仕様・ワークシート作成

大規模データの記述仕様には、多様な言語現象に対応できる十分な記述力と記述者による揺れの生じない明解さが同時に求められる。したがって、記述の枠組としてはゆるやかにしておく一方で、ある言語現象の記述結果が記述者により異なることがないよう、仕様の意図を明確にする必要がある。特に電子化辞書の仕様には記述者も不慣れであるため、誤解を生じやすい。

そこで従来の辞書の仕様と混同しやすいものについては、説明を加えて違いを明示した。例えばある名詞が一つの語義において可算、不可算の2つの用法をもつ場合に、従来の辞書では両用法を同じ見出しの下に併記しているものが多い。しかし、本辞書においては各々の見出しごとに可算用法、不可算用法の他にも連接属性、語形変化情報、冠詞との共起関係などの各種の情報を記述しているので、可算、不可算の違いにより、他のデータの値も異なってくる。そこで一つの名詞の同じ語義であっても、可算、不可算の2つの用法に分けて別見出しとして記述するようにした。

また、従来の辞書にない電子化辞書特有の仕様の記述にあたってはワークシートを工夫して記述の便を図った。例えばＥＤＲ単語辞書では、活用語の場合、活用の際変化を受けない部分 (不変化部分) を見出しとし、その見出しの左右に接続可能な形態素を連接属性というコードで示している。この連接属性の品質向上の際にはコードそのものを見直すのではなく、各見出しの連接属性から自動生成した活

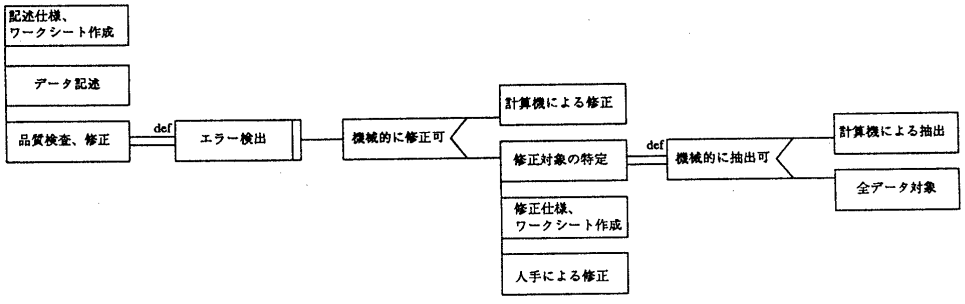

図 1: 単語辞書の開発工程

Developing a Large-Scale Electronic English Word Dictionary
Masako Arioka, Hiromi Ishiwatari, Yasuhiro Kawata, Hiroshi Suematsu, Chiaki Harada, Shin-ya Amano
Japan Electronic Dictionary Research Institute, Ltd.

damnify;1

( ) 原形	damnify
(M) 三単現形	damnifies
(M) 過去形	damnified
(M) 過去分詞形	damnified
( ) 現在分詞形	damnifying

=# : E((of something)to get damaged)

図 2: 辞書記述 (連接属性) 修正後のワークシートの例

用形を修正対象として、図2のようなワークシート上で作業を行った。誤りが検出された活用形は、修正の種類を示すマークを付した上で、直接上書きして修正した。修正結果は形態素解析により連接属性コードに再度変換することで、効率的な修正を行うことができた。

2.2 データ記述

大量のデータを複数の記述者が記述するため、記述者間の判断のばらつきが問題になる。例えば、言語現象自体が複雑で記述が困難であったのは、名詞、形容詞、副詞の活用の有無に関する情報である。名詞であれば数えられるか、形容詞、副詞であれば程度表現可能かという判断を個々の語の表す概念ごとに行ったが、大規模な語彙を対象に統一のとれた記述結果を得ることは非常に困難であった。また形容詞 'extreme' のように比較的高頻度に用いられると思われる語であっても、この語が「極度の」という語義において程度表現可能であるという事実は、必ずしもすべての記述者の直観とは一致しなかった。

また、外来語などで記述仕様作成時には予測できなかった変則的な振舞いをするものがデータ記述の段階で検出されることがある。例えば 'Grand Prix(国際自動車レース)' の複数形には 'Grand Prix'、'Grand Prixes'、'Grands Prix' の 3 つがある。このような外来語の複合語は全体を一語として登録し、英語の屈折規則で対応できるものは英語と同様に活用パタンを記述し、英語にはない不規則な活用をするものは各々の活用形を別見出しとした。上述の 'Grand Prix' の場合は以下のように記述した。

検索見出し	右連接属性	語形変化情報
Grand Prix	名詞単複同形	なし
Grand Prix	名詞語幹後接 es	es 変化型
Grand Prix	名詞単数形	不規則
Grands Prix	名詞複数形	不規則

2.3 品質検査・修正

エラーが検出された場合、単に全数データの目視を繰り返すだけでは全体の品質は大きく改善されない。そこでサンプル中に検出されたエラーの傾向を分析し、エラーの種類ごとに範囲を推定して集中的に修正を行うこととした。エラーの特徴が形態やコードなどで特定できると、計算機により抽出して修正対象の絞り込みを行った。

特定の形態をもつものにエラーがみられたものには、活用語の語形変化に関する情報がある。これは、活用の際に語形変化を起こさない部分 (不変化部分) と残りの語尾との切り分けを誤り、そのまま屈折規則を誤適用したために生じたものと思われる。例えば形容詞では 'piney' のように特に語末が '-ey' で終わるものにエラーが集中していたので、これらは計算機で取り出して修正を行った。

単語	不変化部分	原形語尾	比較級	最上級
piney	pin	-ey	-ier	-iest
(誤)	pine	-y	-ier	-iest

また、名詞の文法情報については名詞の語形、可算・不可算、冠詞との共起、動詞との一致などのコードの組み合わせに制約があるので、組み合わせ自体に矛盾がみられるものは計算機により抽出して修正している。

3 まとめ

ＥＤＲ英語単語辞書の開発方法について、開発の各工程における問題点と方策を中心に報告した。

大規模電子化辞書を効率的に開発するためには、ワークシートの工夫や計算機の利用により記述、修正を支援することが効果的であった。また、良質なデータを得るためには、記述者とのコミュニケーションにより、記述の段階で生じた問題点を記述仕様に反映し、不慣れな電子化辞書仕様での記述をフォローしていくことが重要であった。

今後は実際に英語単語辞書データを利用した実証評価からのフィードバックを通じて、一層精度を上げていく予定である。

参考文献

[1] 日本電子化辞書研究所 : TR-019 英語単語辞書, 日本電子化辞書研究所 (1990).