様式 C-7-2

自己評価報告書

平成２１年　４月２７日現在

研究種目：特定領域研究
研究期間：2006～2010
課題番号：18060012
研究課題名（和文）　多様な目的に適した形態素解析システム用電子化辞書の開発

研究課題名（英文）The development of a multi-purpose electronic dictionary for morphological analyzers

研究代表者
氏　　名　：傳　康晴（DEN YASUHARU）
所属機関・部局・職：千葉大学・文学部・教授
研　究　者　番　号：７０２９１４５８

研究分野：人文学
科研費の分科・細目：言語学・言語学
キーワード：電子化辞書　形態素解析　書き言葉コーパス　音変化　アクセント

１．研究計画の概要
（１）従来開発を進めてきた形態素解析システム用電子化辞書 UniDic を拡充・改良することにより、本研究領域が目指す大規模書き言葉コーパスの構築を支援する。
（２）日本語学・日本語教育学における語彙・文法調査研究、自然言語処理における構文・意味解析研究、音声情報処理におけるテキスト音声合成研究など、多様な目的に適した統合的な電子化辞書およびその利用システムを提供する。
（３）これら目的の達成のために、本研究領域で用いる短単位辞書 10 万語以上と解析精度 98%以上の処理システムを開発する。

２．研究の進捗状況
（１）短単位辞書を関係データベースとして実装した。研究分担者・研究補佐員が、常時、辞書登録作業を行ない、15 万語を超える辞書情報の登録を行った。見出し・表記・品詞などの基本的な辞書情報に加えて、語種・発音・アクセント型および音変化やアクセント変化に関わる情報など、多彩な情報を記述した。
（２）辞書データベースと学習コーパスから形態素解析システム用辞書を作成した。形態素解析システム ChaSen と MeCab で運用・評価を行い、品詞認定 98.9%、見出し認定 98.6%の高い解析精度を得た。さらに、本辞書を用いて形態素解析を実行するための周辺ツールやグラフィックユーザインタフェースを開発した。これらを Web ページで無償公開し、1600 人を超えるユーザ登録があった。
（３）語の複合に伴う音変化・アクセント変化に関するデータを作成し、処理システムを開発した。「数詞＋助数詞」より一般的な音変化（連濁など）を扱うために、学習データの作成を行い、機械学習による処理システムを試作した。同様に、アクセント付きコーパスを作成し、機械学習によるアクセント変化処理システムを試作した。
（４）短単位より長い単位の自動構成システムを作成した。本研究領域のコーパスの一部に長単位情報を付与し、機械学習による手法で単位境界認定 98.6%、品詞認定 97.7%の高い精度を得た。また、音変化・アクセント変化処理を高度化するための中単位自動構成システムを試作した。
（５）ジャンル別の形態素解析システム用辞書を作成した。本研究領域のコーパスに含まれる書籍・新聞・政府刊行白書などのジャンルの文章を対象に、語の分布の統計的性質の異なりを調査した。その示唆をもとに、ジャンルごとに適応した形態素解析システム用辞書を作成する手法を提案し、評価実験によってその有効性を確認した。

３．現在までの達成度
当初の計画以上に進展している。
（理由）
短単位辞書の登録語数や形態素解析システム用辞書の解析精度については、当初目標（10 万語以上・98%以上）を既に達成しており、利用者数も想定以上である。加えて、当初目標にはなかった、ジャンル別辞書の開発でも成功を収めている。

４．今後の研究の推進方策
以下の点に重点をおきつつ、短単位辞書の拡充とその利用システムの開発・改良を引き

続き行う。①長単位構成システムの高精度化（品詞・見出し認定で 98%以上)、②音変化・アクセント変化処理と中単位構成システムの統合、③短単位辞書への語の意味分類の記述、④ジャンル別形態素解析辞書とその開発環境の公開、⑤利便性の高い形式での短単位辞書データベースの公開

５． 代表的な研究成果
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 40 件）

①小木曽智信・伝康晴・渡部涼子, ジャンル別 UniDic作成の試み, 特定領域研究「日本語コーパス」平成 20 年度公開ワークショップ（研究成果報告会）予稿集, pp. 17-22, 2009, 査読無

② Uchimoto, K., & Den, Y., Word-level dependency-structure annotation to Corpus of Spontaneous Japanese and its application, Proceedings of the 6th International Conference on Language Resources and Evaluation, pp. 3118-3122, 2008, 査読有.

③Den, Y., Nakamura, J., Ogiso, T., & Ogura, H., A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation, Proceedings of the 6th International Conference on Language Resources and Evaluation, pp. 1019-1024, 2008, 査読有.

④伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵, コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用, 日本語科学, 22, pp. 101-122, 2007, 査読有.

〔図書〕（計 2 件）

①小木曽智信・中村壮範, 『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装, 特定領域研究「日本語コーパス」特定領域研究「日本語コーパス」平成 20 年度研究成果報告書, 141 頁, 2009.

〔その他〕

① 形態素解析システム用辞書 UniDic 公開ホームページ http://unidic.download.org/