様式 C-19

科学研究費補助金研究成果報告書

平成２４年　５月１８日現在

機関番号：１２５０１
研究種目：　特定領域研究
研究期間：２００６～２０１０
課題番号：１８０６１００２
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 electric dictionary for morphological analyzers

研究代表者　傳　康晴（DEN　YASUHARU）
千葉大学・文学部・教授
研究者番号：７０２９１４５８

研究成果の概要（和文）：
(1) 以下の特徴を持つ形態素解析辞書 UniDic を設計・開発した。
- 「短単位」という揺れがない斉一な単位で設計
- 語彙素・語形・書字形・発音形の階層構造を持ち、表記の揺れや語形の変異にかかわらず同一の見出しを与えることが可能
- アクセントや音変化の情報を付与でき、テキスト音声合成などに利用可能

(2) 辞書データベースを構築しながら、形態素解析システム MeCab 用辞書を随時公開し、最終的に語彙素約 21 万語･書字形約 33 万語の規模と、品詞認定約 98.9%･語彙素認定約 98.6%の解析精度を達成した。
(3) さらに、辞書データベースを XML ファイル群として記述し、ユーザがカスタマイズ可能な辞書作成環境を提供する新しい方式で UniDic2 を設計・開発した。
(4) 中・長単位解析システムを含む、形態素解析の後処理ツール群を作成し、多様な目的に供した。

研究成果の概要（英文）：
(1) An electric dictionary for morphological analyzers with the following characteristics has been developed.
- Lexical entries with uniform unit-size based on *Short-Unit Words*
- Hierarchical representation of lexical entries, consisting of *lemma*, *form*, *orthography*, and *pronunciation*, which enables us to deal with variations in orthography and word form
- Rich information including features for phonological and accentual sandhi

(2) A version for morphological analyzer *MeCab* has been derived from the dictionary database, with several updates, which amounts to 210K lemma and 330K orthographic entries and which achieves an accuracy of 98.9% in part-of-speech tagging and an accuracy of 98.6% in lemma identification.
(3) A version of the dictionary database represented by XML files has also been developed, which enables users to build customized dictionaries for morphological analyzers according to the user's preference and purpose.
(4) Post-processing tools, including *Middle*- and *Long-Unit-Word* analyzers, have been developed for advanced use of the dictionary, such as syntactic analysis and text-to-speech application.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００６年度	17,200,000	0	17,200,000
２００７年度	19,000,000	0	19,000,000
２００８年度	19,000,000	0	19,000,000
２００９年度	19,000,000	0	19,000,000
２０１０年度	17,700,000	0	17,700,000
総　計	91,900,000	0	91,900,000

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．研究開始当初の背景

(1) 日本語の電子化辞書は数多く存在した（EDR 辞書･IPAL 辞書･NTT 語彙特性、各出版社が発行している電子化辞書など）が、これらは検索目的のものであり、形態素解析システムなど言語解析での利用を想定したものではない。

(2) 一方、形態素解析システム（ChaSen や Juman など）付属の辞書は、自動解析に必要な範囲の情報の記述にとどまっており、計量言語学的な研究に必要な単位の斉一性（何を一語とするかについての基準）や語の同一性（異表記や異形態の扱い）の問題を解決していなかった。

２．研究の目的

本研究の目的は、単位の斉一性や語の同一性の問題を解決した形態素解析辞書の開発を通じ、以下を達成することである。

(1) 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．研究の方法

(1) 単位の斉一性・語の同一性・多彩な記述を満たす形態素解析辞書を設計し、辞書データベースを構築する。

(2) 辞書データベースから形態素解析システム用辞書を作成し、本研究領域内外に公開する（目標語彙数 10 万語以上、解析精度 98%以上）。

(3) 複数粒度での単語認定、および、音変化・アクセント変化を処理する後処理ツール群を開発する。

４．研究成果

(1) 以下の特徴を持つ形態素解析辞書 UniDic を設計・開発した。

- 「短単位」という揺れがない斉一な単位で設計
- 語彙素・語形・書字形・発音形の階層構造を持ち、表記の揺れや語形の変異にかかわらず同一の見出しを与えることが可能
- 音変化やアクセント変化の情報を付与でき、テキスト音声合成などに利用可能

(2) 辞書データベースを構築しながら、形態素解析システム MeCab 用辞書を随時公開し、最終的に、語彙素約 21 万語・書字形約 33 万語の規模と、品詞認定約 98.9%・語彙素認定約 98.6%の解析精度を達成した。

(3) さらに、辞書データベースを XML ファイル群として記述し、ユーザがカスタマイズ可能な辞書作成環境を提供する新しい方式で UniDic2 を設計・開発した。

(4) 中・長単位解析システムを含む、形態素解析の後処理ツール群を作成し、多様な目的に供した。

本研究の成果である形態素解析辞書はその規模・記載内容において最高水準のものであり、解析システムは短単位解析から中・長単位構成、音変化・アクセント変化処理までカバーする他に類を見ないものである。

５．主な発表論文等

（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 73 件）

① K. Maekawa, M. Yamazaki, T. Maruyama, M. Yamaguchi, H. Ogura, W. Kashino, T. Ogiso, H. Koiso, and Y. Den, Design, compilation, and preliminary analyses of Balanced Corpus of Contemporary Written Japanese, *Proceedings of LREC2010*, 査読有, 2010, 1483-1486.

② 小木曽智信・小椋秀樹・田中牧郎・近藤明日子・伝康晴, 中古和文を対象とした形態素解析辞書の開発, 情報処理学会研究報告，査読無，2010-CH-85, 2010, 49-64.

③ H. Hirano, M. Suzuki, K. Innami, N. Minematsu, and K. Hirose, Development of

an on-line word accent dictionary of Japanese, *Proceedings of JSAA-ICJLE 2009*, 査読有, 2009.

④ 伝康晴, 多様な目的に適した形態素解析システム用電子化辞書, 人工知能学会誌, 査読無, 24, 2009, 640-646.

⑤ 浜辺良二・内元清貴・河原達也・井佐原均, 話し言葉における引用節・挿入節の自動認定および係り受け解析への応用, 自然言語処理, 査読有, 16 (1), 2009, 3-23.

⑥ 小木曽智信, 形態論情報の自動付与とその問題点, 国文学 解釈と鑑賞, 査読無, 74 (1), 2009, 35-43.

⑦ K. Uchimoto and Y. Den, Word-level dependency-structure annotation to Corpus of Spontaneous Japanese and its application, *Proceedings of LREC2008*, 査読有, 2008, 3118-3122.

⑧ Y. Den, J. Nakamura, T. Ogiso, and H. Ogura, A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation, *Proceedings of LREC2008*, 査読有, 2008, 1019-1024.

⑨ N. Minematsu, R. Kuroiwa, K. Hirose, and M. Watanabe, CRF-based statistical learning of Japanese accent sandhi for developing Japanese text-to-speech synthesis systems, *Proceedings of ISCA Workshop on Speech Synthesis*, 査読有, 2007, 148-153.

⑩ K. Uchimoto, and H. Isahara, Morphological annotation of a large spontaneous speech corpus in Japanese, *Proceedings of IJCAI2007*, 査読有, 2007, 1731-1737.

⑪ 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, 査読有, 22, 2007, 101-122.

⑫ K. Uchimoto, R. Hamabe, T. Maruyama, K. Takanashi, T. Kawahara, and H. Isahara, Dependency-structure annotation to Corpus of Spontaneous Japanese, *Proceedings of LREC2006*, 査読有, 2006, 635-638.

〔学会発表〕（計 59 件）

① 小木曽智信・小椋秀樹・小磯花絵・宮内佐夜香・渡部涼子・伝康晴, 形態素解析辞書のベンチマークテスト―IPAdic・NAIST-jdic・UniDic のジャンル別精度比較―, 言語処理学会第 16 回年次大会, 2010 年 3 月 10 日, 東京大学（東京）.

② 山田篤・伝康晴, UniDic 汎用後処理ツールの設計と実装, 特定領域研究「日本語コーパス」平成 21 年度公開ワークショップ, 2010 年 3 月 15 日, 東京工業大学（東京）.

③ 小磯花絵・田中弥生・小木曽智信・近藤明日子, テキストの多様性をとらえる分類指標の体系化の試み, 言語処理学会第 17 回年次大会, 2011 年 3 月 9 日, 豊橋技術科学大学（愛知）.

〔図書〕（計 6 件）

① 小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕, 国立国語研究所, 特定領域研究「日本語コーパス」平成 22 年度研究成果報告書『現代日本語書き言葉均衡コーパス』形態論情報規定集 第 4 版（上・下）, 2011, 359.

② 小木曽智信・中村壮範, 国立国語研究所, 特定領域研究「日本語コーパス」平成 22 年度研究成果報告書『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 改訂版, 2011, 145.

〔その他〕
ホームページ等
http://download.unidic.org/

６．研究組織
(1)研究代表者
傳　康晴（DEN　YASUHARU）
千葉大学・文学部・教授
研究者番号：70291458

(2)研究分担者
山田　篤（Yamada Atsushi）
京都高度技術研究所・研究部・主席研究員
研究者番号：20240004
（H19→H20：連携研究者）

峯松　信明（Minematsu Nobuaki）
東京大学・大学院新領域創成科学研究科・准教授
研究者番号：90273333

内元　清貴（Uchimoto Kiyotaka）
情報通信研究機構・総合企画部・プランニングマネージャー
研究者番号：60358885
（H19→H20：連携研究者）

小木曽　智信（OGISO TOMONOBU）
国立国語研究所・言語・資源研究系・准教授
研究者番号：20337489

小磯　花絵（KOISO HANAE）
国立国語研究所・理論・構造研究系・准教授
研究者番号：30312200