■2 群（画像・音・言語）- 10 編（自然言語処理）

1 章 言語資源

【本章の構成】

本章では，辞書とコーパス（1–1 節）について述べる．

■2 群 - 10 編 - 1 章

1-1 辞書とコーパス

（執筆者：徳永健伸）［2009 年 8 月 受領］

辞書とコーパスは自然言語処理における中核的な言語資源である．辞書が語に関する様々な情報を整理して記述した情報源であるのに対し，コーパスは実際の言語使用の「事例」の集積であり，人間の言語使用の実態を反映した情報源である．計算機による言語処理という観点から見ると，計算機処理用の辞書の重要性が自然言語処理の黎明期（1960 年代）から認識されていたのに対し，コーパスの重要性が認識されたのは 1980 年代に入ってからである．計算機で大量のコーパスを処理する研究が現実的になるためには，電子化テキストの普及と記憶媒体や計算機の処理性能の低価格化を待たなければならなかった．以下では，辞書，コーパスそれぞれについて，その歴史と類型，そして標準化について述べる．

1-1-1 計算機用辞書

(1) 計算機用辞書の歴史

初期の自然言語処理システムでは用途に応じて必要な情報を各語彙項目に対して手作業で記述し，辞書を構築していた．電算写植の普及に伴い，1980 年代から人が利用するために編纂された辞書もその内容が電子化され，計算機で扱えるようになった．このような辞書は機械可読辞書（MRD：Machine Readable Dictionary）と呼ばれる．これらの辞書の内容は人間のための辞書と同一なので，これを言語処理に利用するためには，まず，辞書の内容を解析して必要な情報を抽出しなければならない．機械可読辞書が利用可能になると同時に，そこから言語処理用の様々な情報を抽出し，計算機処理用の辞書を構築する研究が行われた．初期に電子化され，実際に計算機処理の対象として研究された辞書として，英語の LDOCE（Longman Dictionary of Contemporary English），日本語の三省堂新明解国語辞典などがある．

これとほぼ同時期に，情報処理振興事業協会（IPA）技術センターのプロジェクトとして，主に言語学者，国語学者を中心とする研究グループにより IPAL 動詞辞書 (1987 年) が作成された．この辞書は規模こそ和語動詞 861 語と少ないが，形態，統語，意味，テンス・アスペクト情報など動詞の広範な情報を網羅し，言語処理に用いることを前提として編集されている．同グループは続いて形容詞辞書 (1990 年)，名詞辞書 (1996 年) も作成している．また，1986 年には国家プロジェクトとして日本電子化辞書研究所（EDR：Japan Electronic Dictionary Resarch Institute）が設立され，日英の単語辞書，対訳辞書，概念辞書，共起辞書などが 10 万から 40 万項目の規模で作成されている (1993 年)．このプロジェクトは主に機械翻訳システムのための辞書を想定し，工学的な観点から大規模な辞書構築を目標としていた．海外では EU を中心に多言語の辞書構築のプロジェクトが盛んに行われている．特に，多言語を扱う必要性から，EAGLES プロジェクト [1)] に代表されるように辞書の標準化に重点がおかれている．この流れは後述する国際標準機構（ISO）の辞書の国際標準化に発展している．

(2) 計算機用辞書の類型

計算機用の辞書を分類する際には幾つかの観点が考えられるが，表出形である語を基準にしたものと語が表す意味を基準にしたものに大別できる．前述の IPAL や EDR の辞書（概念辞書を除く）は前者に分類できる．語の表出形に基づいて語を分類する際，各語彙項目に記述される情報は、更に品詞や表記などの形態情報，述語がどのような項をとるかを記述し

た下位範疇化情報などの統語情報，語の意味を表す意味クラスなどの意味情報に分類できる．語彙項目にどのような情報を記載するかは，辞書の利用目的によって異なるため，初期の応用システムでは，目的に応じた情報のみを辞書に記述していたが，上述の IPAL や EDR の辞書は汎用性を目的としているため，独自の基準で語の情報を網羅的に記述している．

語が表す意味を基準に分類する代表的な辞書として，EDR の概念辞書，国立国語研究所の分類語彙表（初版：1964 年，増補改訂版：2004 年)，NTT の日本語語彙体系の意味体系 (1997 年)，Princeton 大学の WordNet (3.0 版：2006 年) [2)]などがある．これらはシソーラス（Thesaurus）あるいは語彙オントロジー（Lexical Ontology）と呼ばれることもある．

1-1-2 コーパス

(1) コーパスの歴史

コーパスは，現実に使用された言語表現の事例を収集したデータである．どのような情報源から言語表現を収集するかによって様々な種類のコーパスが考えられる．工学的な観点からは応用システムが対象とする分野の事例を収集することが有効であるが，言語学的な観点からは言語の多様性を捉えるために様々な分野から，実際の使用実態をできるだけ反映するように事例を収集することも重要な要素である．このように分野のバランスを考慮して作られたコーパスは均衡コーパスと呼ばれている．

電子化されたコーパスという意味では 1960 年代に米国 Brown 大学で構築されたブラウン・コーパスが先駆的である．Brown コーパスは明確な基準に基づいて分野のバランスを考慮した 100 万語からなる米語の均衡コーパスである．1990 年代には規模を拡大し，1 億語からなる英語の均衡コーパス，BNC（British National Corpus）[3)]が構築される．最近では，この規模は更に拡大し，3 億 8 千 5 百万語以上からなる米語のコーパス COCA（Corpus of Contemporary American English）[4, 5)]が構築されている．COCA は複数のジャンルから事例を収集しているが，毎年 2000 万語を追加する計画であり，均衡コーパスとは言えない．

これらはいずれも言語学の観点から作成されたコーパスであるが，1990 年代から自然言語処理の観点からもコーパスが構築されるようになった．その理由として，1980 年代の終わりから自然言語処理の分野で主流となったコーパスに基づく言語処理の基礎データとしてコーパスが研究の必要不可欠な要素となったことが大きい．大量のコーパスと機械学習技術の進展によって，形態素解析や統語解析などの自然言語処理の基礎技術の性能は飛躍的に向上した．

自然言語処理のためのコーパスの先駆けは米国 LDC（Linguistic Data Consortium）から配布された Penn Treebank である [6)]．Penn Treebank は Brown コーパスを包含し，更に各語の品詞情報，各文の構文情報が付与されている．Penn Treebank の流れは，その規模を増やすと同時に，述語－項構造（PropBank [7)]）や談話構造などをより複雑な情報を付与したコーパス（Penn Discourse TreeBank [8)]）へと発展している．

日本では，EDR の辞書と同時に EDR コーパスが配布された．EDR コーパスは新聞記事，雑誌から選択された日本語約 20 万文，英語約 12 万文を含み，これらの文に EDR 辞書に準拠した形態・統語・意味情報が付与されている．1996 年には RWC(Real World Computing) プロジェクトの成果の一部として RWC コーパスの配布が始まった [9)]．RWC コーパスでは，新聞記事，白書，岩波国語辞典などの内容を年次進行で公開し，形態・統語情報，図書分類コード（UDC），岩波国語辞典の語義などを付与している．1997 年には毎日新聞の記事約 4 万文

に形態・統語情報を付与した京都大学テキストコーパスが公開されている [10]．Penn TreeBank と同様に，このコーパスは，その後，述語－項構造，照応関係などの情報も付与されたコーパスへ発展している．

最近では，言語学・国語学と自然言語処理の研究者が協力する形でコーパス構築を行うプロジェクトも始まっている．2004 年には，国立国語研究所，情報通信研究機構，東京工業大学により共同開発された日本語話し言葉コーパス（CSJ）が公開された [11]．CSJ は 600 時間を越える音声とその書き起しテキストからなる．書き起しテキストには形態情報，節単位情報，統語情報などが付与されている．また，2006 年には国立国語研究所を中心として，現代日本語書き言葉均衡コーパス（BCCWJ）の構築が始まっている [12]．BCCWJ は日本語の大規模な均衡コーパスとしては最初のコーパスである．

Web の発展に伴い，Web 上のテキストを検索エンジンなどを通じて収集し，コーパス代りに用いようという試みも盛んに行われている [13]．その研究成果として，Web 上の大量のテキストを解析した結果得られた格フレーム情報 [14]や従来より長い n-gram [15]などが公開されている．しかしながら，検索エンジンを用いて Web をコーパス代りに使う手法については注意が必要であるとの指摘もある [16, 17]．

(2) コーパスの類型

言語学の観点から見たときに重要となるコーパスの特徴としては，言葉の種類（書き言葉/話し言葉），設計方針（サンプル/モニター），時代区分（共時/通時），対象言語数，付与情報の有無などが挙げられる．サンプルコーパスとは一定量のテキストをある方針でサンプリングして作成したものであり，モニターコーパスは収集を続けることによってたえず更新されるコーパスである．これは共時/通時の区分にも関係する．共時コーパスはある時代区分に限定してテキストを収集したものであり，通時コーパスは長期間にわたってテキストを収集したものである．後者は言語変化の調査など，通時的な研究に用いられる．

一方，自然言語処理の観点からは，収集する対象を電子化テキストが入手しやすいという現実的な理由や，想定する応用システムの対象領域であるという理由で選ぶことが多い．最も重要なのはコーパスに付与する情報の種類である．付与情報はタグの形式でメタテキストとして本文中に埋め込まれることがあることから，このようなコーパスはタグ付きコーパスあるいはアノテーション付きコーパスと呼ばれる．付与する情報としては，一般的には形態・統語・意味・談話情報などがある．これらの情報は階層的な関係を持つことが多く，同一のテキストに段階的に付与されることが多い．例えば，統語情報は形態情報を前提に付与されるし，語の間の意味的な関係は統語情報を前提に付与される．前述の Penn TreeBank や京都大学テキストコーパスもこのような順に発展している．

一般的に，コーパスの中心を占めるのはテキストであるが，テキストに代表される言語情報のほかにもそのテキストが使用された状況における非言語情報もあわせて記録したコーパスも構築されている．非言語情報としては，音韻情報，ジェスチャー，視線，表情などが用いられている．このようなコーパスをマルチモーダルコーパスと呼ぶ．非言語情報は特に対話などのインタラクションにおいて重要な情報であり，対話コーパスはマルチモーダルコーパスとして構築されることも多い．

1-1-3 言語資源の流通と標準化

言語資源の整備は欧米の LDC [18]や ELRA（European Language Resource Association）[19]といった言語資源の管理・流通を目的とした運営母体によって行われてきた．日本でも言語資源協会（GSK）[20]が同様の活動を行っている．これらの組織の Web ページには入手可能な言語資源のカタログ情報が公開されている．更に世界中の言語資源のカタログを作成することを目標とした OLAC（Open Language Archive Community）[21]という試みもある．

OLAC のように複数の言語の言語資源を扱うためには，情報の記述項目や記述方法に関する標準化が重要な課題となる．OLAC では Dublin Core を言語資源用に拡張したメタデータを使っている．メタデータではなく，言語資源そのものの記述に関する標準化では言語ごとの特徴を考慮する必要があり，更に標準化は困難となる．現在，言語資源の記述内容・形式に関しては ISO の部会 TC37/SC4（Language Resource Mangament）[22]で議論されており，既に幾つかの国際標準が策定されている．例えば，LMF（Lexical Markup Framework：ISO 24613）[23]は，UML（Unified Modeling Language）を用いて辞書の情報の入れ物としての構造を定義している．一般に，辞書は使用目的によって記述すべき情報が異なるので，LMF では，辞書の中核となる部分（Lexicon, Lexical Entry, Form, Form Representation, Sense, Definition などのクラス）をコアパッケージとして用意し，その他の情報は拡張パッケージとして用意している．拡張パッケージとしては，形態情報のパッケージ，言語処理のための統語情報や意味情報のパッケージなどが提案されている．このほかにも形態・統語の情報を記述するための MAF（Morphosyntactic Annotation Framework: 24611）や SynAF（Syntactic Annotation Framework: 24615），コーパスへの情報付与の枠組として LAF（Linguistic Annotation Framework: 24612）などが検討されている．このような言語資源の標準化は言語を越えたシステムの相互互換性を高めるために貢献するばかりでなく，アジアの多くの言語のようにまだ十分に言語資源を構築していない言語にとっても，資源構築のための良い指針となる．

■参考文献

1) EAGLES：“Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora,” Technical Report EAG-CLWG-MORPHSYN/R, EAGLES, 1996.
2) C. Fellbaum：“*WordNet: An Electronic Lexical Database*,” The MIT Press, 1998.
3) BNC: `http://www.natcorp.ox.ac.uk/`.
4) COCA: `http://www.americancorpus.org`.
5) M. Davies：“The 385+ million word corpus of contemporary American English (1990-2008+),” International Journal of Corpus Linguistics, vol.14, no.2, pp.159–190, 2009.
6) M.P. Marcus, M.A. Marcinkiewicz, and B. Santorini：“Building a large annotated corpus of English: The Penn Treebank,” *Computational Linguistics*, vol.19, no.2, pp.313–330, 1993.
7) M. Palmer, D. Gildea, and P. Kingsbury：“The Proposition Bank: A corpus annotated with semantic roles,” *Computational Linguistics*, vol.31, no.1, pp.71–105, 2005.
8) R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber：“The Penn Discourse Treebank 2.0,” Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), pp.28–30, 2008.
9) K. Hasida, H. Isahara, T. Tokunaga, M. Hashimoto, S. Ogino, W. Kashino, J. Toyoura, and H. Takahashi：“The RWC Text Databases,” Proceedings of The First International Conference on Language Resource and Evaluation (LREC 1998), pp.457–461, 1998.

10) 黒橋禎夫, 長尾　眞：“京都大学テキストコーパス・プロジェクト,” 言語処理学会第 3 回年次大会予稿集, pp.115–118, 1997.
11) 国立国語研究所：“日本語話し言葉コーパスの構築,” 国立国語研究所報告書 124, 2004.
12) 前川喜久雄：“KOTONOHA『現代日本語書き言葉均衡コーパス』の開発,” 日本語の研究, vol.4, no.1, pp.82–95, 2008.
13) A. Kilgarriff and G. Grefenstette：“Introduction to the special issue on the Web as corpus,” *Computational Linguistics*, vol.29, no.3, pp.333–347, 2003.
14) D. Kawahara and S. Kurohashi：“Case frame compilation from the Web using high-performance computing,” Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pp.1344–1347, 2006.
15) “Web 日本語 N グラム第 1 版,” `http://www.gsk.or.jp/catalog/GSK2007-C/catalog.html`
16) A. Kilgarriff：“Googleology is bad science,” *Computational Linguistics*, vol.33, no.1, pp.147–151, 2007.
17) 荻野綱男：“コーパスとしてのWWW検索の活用,” 月刊言語, vol.36, no.7, pp.26–33, 2007.
18) LDC: `http://www.ldc.upenn.edu/`.
19) ELRA: `http://www.elra.info`.
20) GSK: `http://www.gsk.or.jp`.
21) OLAC: `http://www.language-archives.org/`.
22) ISO/TC37/SC4: `http://www.tc37sc4.org/index.php`.
23) G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria：“Multilingual resources for NLP in the lexical markup framework (LMF),” *Language Resources and Evaluation*. vol.43, no.1, pp.57–70, 2009.