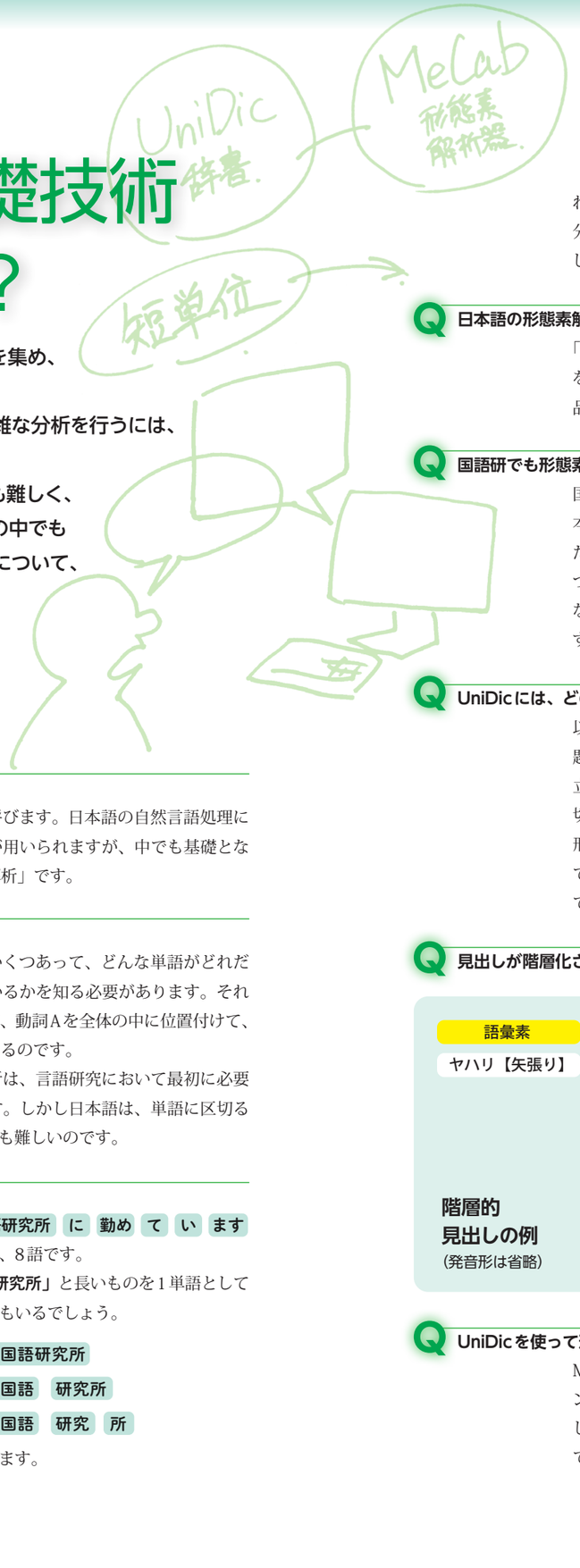
特集

コンピュータと人間の言語

自然言語処理の基礎技術「形態素解析」とは？

小木曽智信
国立国語研究所 研究系 教授

ことばの研究では、たくさんのことばを集め、それをさまざまな観点から分析します。多種多様なことばを大量に蓄積し、複雑な分析を行うには、コンピュータが不可欠です。しかし、ことばをコンピュータで扱うのはとても難しく、さまざまな技術が必要になります。その中でも基礎となる技術である「形態素解析（けいたいそ）」について、小木曽智信教授に聞きました。



Q 小木曽さんは、どのような研究をされているのですか？

専門は日本語学と自然言語処理です。

Q 自然言語処理とは？

自然言語処理とは、私たちが普段使っていることばをコンピュータで扱うための技術のことです。コンピュータで用いるプログラミング言語に対して、人が用いる言語を自然言語と呼びます。日本語の自然言語処理にはさまざまな技術が用いられますが、中でも基礎となる技術が「形態素解析」です。

Q 形態素解析とは、どういう技術ですか？

形態素解析とは、文を単語に区切り、品詞や読みなどの情報を付ける技術をいいます。言語を研究するとき、文・単語・文字・音韻など、さまざまな単位で分析しますが、それらの中で基本になる単位が、単語です。

例えば、あるテキストの中の動詞Aについて、これがどのような言葉か調べたいとしましょう。そのためには、動詞Aが出てくる回数だけでなく、対象とするテキストに単語がいくつあって、どんな単語がどれだけ一緒に使われているかを知る必要があります。それらが分かって初めて、動詞Aを全体の中に位置付けて、統計的な分析ができるのです。

従って形態素解析は、言語研究において最初に必要な基礎的な技術です。しかし日本語は、単語に区切るということが、とても難しいのです。

Q 具体的にはどういう難しさがあるのでしょうか？

「私は国立国語研究所に勤めています」という文は、何語に分けられるでしょうか？

私 は 国立国語研究所 に 勤め て い ます

このように区切ると、8語です。

でも、**「国立国語研究所」**と長いものを1単語としてよいのか？と思う人もいるでしょう。

国立 国語研究所
国立 国語 研究所
国立 国語 研究 所

と区切ることもできます。

英語では、単語がスペースで区切られ分かち書きされるので、「これが1単語だ」と分かります。日本語は分かち書きされないので、1単語が明確ではないのです。しかし正しく区切っておかないと、「京都」と検索したら「東京都」、「カレー」と検索したら「エスカレーター」が出てしまったり、係り受け解析や構文解析が正しくできなかったりして、言語研究に使えません。

Q 日本語の形態素解析をどのように行うのですか？

「形態素解析器」と呼ばれるコンピュータプログラムを使います。文を入力すると、単語に分けて、さらに品詞や活用形などを判別した結果が出力されます。形態素解析器は、形態素解析用の辞書と組み合わせて使います。形態素解析器も解析用の辞書も、さまざまなものが開発されています。

Q 国語研でも形態素解析用の辞書などを開発しているのでしょうか？

国語研では2000年ごろから、1億語からなる「現代日本語書き言葉均衡コーパス（BCCWJ）」の構築に向けた準備を始めました。コーパスとは、ことばを大量かつ体系的に収集し、研究用の情報を付与してさまざまな検索ができるようにした、ことばのデータベースです。BCCWJのために、形態素解析を精度よく言語研究に適した形で行えるように、新しい辞書を国語研が中心となって開発しました。それがUniDic（ユニディック）です。

UniDicは、MeCab（めかぶ）という形態素解析器で利用できるようになっています。MeCabは、工藤拓氏によって開発されたソフトウェアで、高速かつ高精度であることから、現在最もよく利用されている形態素解析器です。

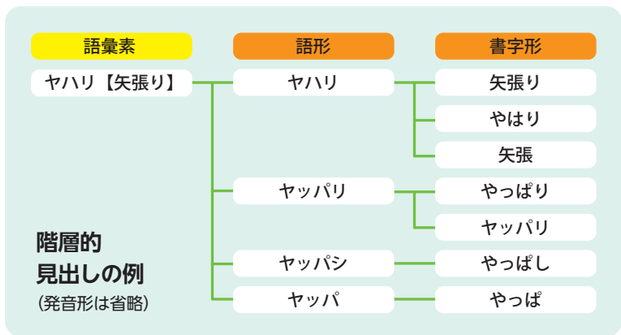
Q UniDicには、どのような特長があるのでしょうか？

以前からあった形態素解析用の辞書には、いくつか問題がありました。その一つが、単語の区切り方です。「国立国語研究所」の例で示したように、1単語を長く区切ることも、短く区切ることもできます。それまでの形態素解析用の辞書では、単語の区切り方が統一されていなかったため、統計的な分析が正確にできませんでした。

UniDicの特長の一つは、単語の区切り方を、国語研で決めた「短単位」というルールで統一していることです。短単位は、区切りの基準が分かりやすく、揺れが少なくなります。そのため、これを基盤としてさまざまな研究を行うことができるのです。

UniDicのもう一つの特長は、見出しが階層化されていることです。

Q 見出しが階層化されているとは？



UniDicでは見出しを、語彙素、語形、書字形、発音形と4つのレベルの階層構造にしています。皆さんが使う辞書の見出しに相当するのが、語彙素です。「ヤハリ」という単語は、「ヤッパリ」「ヤッパシ」「ヤッパ」と語形が揺れたり、「矢張り」「やはり」「矢張」と書字形が揺れたりします。以前の辞書は、複数ある語形、書字形を別々に扱っていました。これではそれぞれ別の見出しになるので、一括した検索や集計ができません。

UniDicでは語形や書字形の揺れにかかわらず、同一の見出しとしてまとめられているので、目的に応じて利用することが可能です。

Q UniDicを使って形態素解析を行うには、専門知識が必要ですか？

MeCabもUniDicも公開されているので、パソコンにインストールすれば誰でも形態素解析ができます。しかし、黒い画面にコマンドを打って操作するので、慣れていない人にはハードルが高いかもしれません。

そこで、もっと簡単に使えるように開発したのが、「Web茶まめ」です。Web茶まめは、インストールが不要でオンラインで使え、専門知識がなくても形態素解析ができます。

Q Web茶まめの使い方を教えてください。

正岡子規の俳句「**柿くへば鐘が鳴るなり法隆寺**」を形態素解析してみましょう。

Web茶まめの使い方

① 「Web 茶まめ」のサイトにアクセス
https://chamame.ninjal.ac.jp/

② 解析したいテキストを入力
テキストファイルをアップロードして解析することもできます。

柿くへば鐘が鳴るなり法隆寺

③ 使用する辞書を選択
この例では「近代文語」を選択します。

④ 出力したい項目を選択
一般的な項目が初期設定されています。

⑤ 出力形式を選択

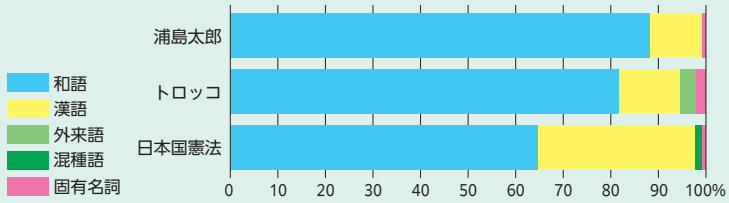
⑥ 「解析する」ボタンをクリック

⑦ 解析結果

辞書	文境界	書字形（＝表層形）	語彙素	語彙素読み	品詞	活用型	活用形	発音形出現形	仮名形出現形	語種	書字形(基本形)	語形(基本形)
近代文語B		柿	柿	カキ	名詞-普通名詞-一般			カキ	カキ	和	柿	カキ
近代文語I		くへ	食う	クウ	動詞-一般	文語四段-ハ行	已然形-一般	クエ	クヘ	和	くふ	クウ
近代文語I		ば	ば	バ	助詞-接続助詞			バ	バ	和	ば	バ
近代文語I		鐘	鐘	カネ	名詞-普通名詞-一般			カネ	カネ	和	鐘	カネ
近代文語I		が	が	ガ	助詞-格助詞			ガ	ガ	和	が	ガ
近代文語I		鳴る	鳴る	ナル	動詞-一般	文語四段-ラ行	連体形-一般	ナル	ナル	和	鳴る	ナル
近代文語I		なり	なり-断定	ナリ	助動詞	文語助動詞-ナリ-断定	終止形-一般	ナリ	ナリ	和	なり	ナリ
近代文語I		法隆	法隆	ホウリュウ	名詞-固有名詞-一般			ホーリュー	ホウリュウ	固	法隆	ホウリュウ
近代文語I		寺	寺	ジ	接尾辞-名詞的-一般			ジ	ジ	漢	寺	ジ

Q おすすめの解析を教えてください。

和語・漢語・外来語・固有名詞という語種の割合に注目すると、そのテキストの特徴が見えてきて面白いですよ。語種を自動で判別できることも、UniDicの特長です。



文部省唱歌の『浦島太郎』、芥川龍之介『トロッコ』の冒頭部分、日本国憲法 前文を、それぞれWeb茶まめで形態素解析を行い、語種の割合を比べると、左のようになりました。解析結果はExcel形式で出力できるので、グラフもつくれます。

好きなアーティストの曲の歌詞や、好きな作家の作品について、Web茶まめで形態素解析を行ってみては、いかがでしょうか。ほかのアーティストや作家のテキストと比べることで、気付いていなかった特徴が見えてくるかもしれません。

Q UniDicや形態素解析について、課題や今後の計画はありますか？

Web茶まめの「辞書選択」でお気付きかもしれませんが、UniDicは1種類ではありません。日本語といっても時代・地域によって違うので、それぞれの時代・地域に合わせた形態素解析用の辞書が必要です。古文用を少しずつ開発してきて、現在UniDicは13種類あり、奈良時代から現代の文章まで解析できるようになっています。関西方言用も間もなく公開できる予定です。ほかの方言用の辞書の開発にも着手しています。現代書き言葉、現代話し言葉の辞書についても、新しい見出しの追加などメンテナンスが欠かせません。

UniDicは、日本語研究だけでなく、文学などほかの分野や産業界でも広く使っていただいています。とてもうれしいですね。

形態素解析という技術は、日本語をコンピュータで処理し、分析するときに基礎となる技術なので、これからも重要な役割を果たしていくことでしょう。