様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　４　月　２４　日現在

機関番号：　１０１０１
研究種目：　挑戦的萌芽研究
研究期間：　2015～2016
課題番号：　１５Ｋ１２８８９

研究課題名（和文）英語語彙表構築法についての研究

研究課題名（英文）Compiling English vocabulary lists using machine-readable dictionaries

研究代表者
園田　勝英（Sonoda, Katsuhide）

北海道大学・メディア・コミュニケーション研究院・特任教授

研究者番号：７０１１３６９４

交付決定額（研究期間全体）：（直接経費）　　600,000円

研究成果の概要（和文）：英語語彙の基礎データベースとして Longman Dictionary of Contemporary English (3rd and 5th eds, 以下LDOCE) を採用し、Python を用いる最新のデータ分析環境である Jupyter Notebook 上で分析を行った。この結果、LDOCEの記述は言語学的に信頼性が高いばかりでなく、比較的新しい言語学研究を反映するものであることを確認した。これに基づいて、さまざまな分析を行うことにより辞書情報の機械的分析が教育用語彙表の改善に役立つことを実証した。

研究成果の概要（英文）：Computational analyses have been carried out of the XML versions of Longman Dictionary of Contemporary English (3rd and 5th eds) with Jupyter Notebook, which is a state-of-the-art Python-programming environment for general data analysis. It has been found out that lexical descriptions in LDOCE are quite reliable linguistically and are based on fairly recent linguistic research. In particular, LDOCE's verb entries include systematic and explicit descriptions of syntactic frames in which the verbs appear for each of their meanings. This has led us to classify English verbs computationally using LDOCE after the fashion of Beth Levin (1993), where she classifies English verbs using diathesis alternations as a main criterion. Together with other analyses, it is demonstrated that computational analyses of dictionary data is essential for improving English vocabulary lists for educational purposes.

研究分野：　英語学

キーワード：　英語語彙表　機械可読辞書　動詞の分類　LとR

１．研究開始当初の背景

1990 年代後半から 2000 年代初めにかけて、コーパス研究の進展に合わせて、多くの英語語彙表が 作成されてきた。この流れはわが国においては JACET8000 (2003) や COCET3300 (2007) などにお いて頂点に達し、教育の現場において広く活用されている。しかし、これらは英語教育における語彙学習をさらに深く研究するための基礎資料としては、十分なものとは言えない。特に、現在の語彙表は語を単に羅列したものであり、次のような欠点を持っている。(1) 収録された語は、どの語も同じ一語であり、語ごとに異なる学習負担量についての情報がない。 例えば、hotel は唯一つの意味を持つ語であるのに対し、time は複数の品詞にまたがり多くの 意味や熟語を持つが、現在の語彙表ではどちらも同じ一語として扱われる。(2) 語と語の間には様々な関係があるが、これらは一切語彙表には反映されない。派生、同義、反意、コロケーション内共起などの関係は、できれば語彙表内に含められるべきである。(3) イディオムの学習は、語彙学習の中で大きな位置を占めるが、現在の語彙表では全く扱われていない。特に、句動詞が排除されているのは、大きな欠点である。

現代英語の動詞についての調査研究は、過去半世紀以上にわたり活発に行われ、統語的、形態的、意味的な特性が急速に解明されつつある。その中でコーパスが大きな役割を果たしていることはよく知られていることである。しかしその一方で、現代英語の動詞全体を対象として、コーパス等を用いて計量的に観察することは、今のところ単純な頻度表作成などに限られていて、余り進んでいるとは言えない。

２．研究の目的

機械可読化された辞書を分析することにより、学習者が習得すべき語彙知識を詳細かつ包括的に分析する。これにより、学習者が習得しなければならない語彙的知識の範囲を画定しレベル別に分類する一般的方法を提案する。また、従来専らコーパスの語彙頻度分析に基づいていた語彙表構築法を超える語彙表構築法を確立する。特に、母語話者のメンタルレキシコンを現在のところ最も忠実に再現している包括的な語彙情報データベースとしての機械可読版学習者用英英辞典を、語彙表作成のための中心的な資料とする。

次世代の英語語彙表は学習者が習得しなければならない語彙的知識の範囲を画定し、それをレベル別に分類するべきである。この展望のもとに、専らコーパスの語彙頻度分析に基づいていた従来の語彙表の構築法を革新することを目指す。特に、母語話者のメンタルレキシコンを最も忠実に再現している語彙情報データベースとしての機械可読版学習者用英英辞典を、語彙表作成資料の中心に位置づける。機械可読版の辞書をコンピュータ上で分析することは、言語系研究者にとって、技術的および方法論的に困難が多く、コーパス言語学の未開拓の分野として残されている。本研究は、このような機械可読版英語辞書をコーパス言語学の観点から分析し、語彙表構築に利用しようとするものである。

３．研究の方法

最近データサイエンスにおいて活用されているプログラム言語 Python に基づくプログラム開発環境である Jupyter Notebook 上で、Longman Dictionary of Contemporary English (3rd and 5th eds)の XML 版を分析した。Python の付加的機能として、XML 分析用の lxml, データベース操作のための pandas, 英語用統語解析のための spacy を用いた。特に、紙媒体の印刷版を読むことでは確認が難しい、語彙記述の根底にある文法理論を推定や、記述の信頼性の検証などを行うことができた。また、品詞ごとに記載されている情報の種類とその記述形式を確認した。これに基づいて、多くの種類の言語情報を、辞書の記載内容から分離抽出する作業を行った。

４．研究成果

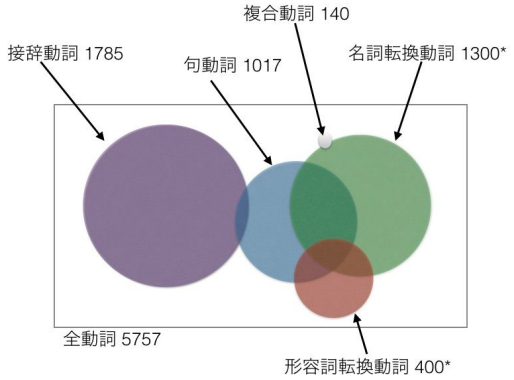
(1) BNC-XML の Written part に出現する一般動詞の数(基本形に直した異語数、以下同様)は約 35,000 であるのに対し、LDOCE 等の上級用英語学習辞書に収録されている一般動詞の数は 6,000 弱である。この差は、語彙の多様性や流動性を反映するものと考えられるが、一般的な英語母語話者が持つ動詞の語彙知識は LDOCE に記載されているものとほぼ等しいと考えられる。したがって、辞書に収録されている約 6,000 の一般動詞を一応「現代英語の動詞全体」と考えられる。

これらのうち、約 140 は bad-mouth, breast-feed, mass-produce などの複合動詞(compound)である。さらに、約 1,800 語の動詞は en-, be, de-, ...; -ify, -ize, -en などの接辞を付加することによって生み出されている。例としては、接尾辞を持つものとして awaken, actualize, amplify, abbreviate など。接頭辞をもつものとして、adjoin, befall, constrict, counterattack, depress, disagree, download, endanger, interact, forecast, outlive, precede, reform, subscribe, translate, untie, upload など。また、約 2,500 の動詞は、同綴りの名詞を持つ。これらのうち 1300 くらいが名詞からの転換によって作られる名詞転換動詞(denominal verbs)であると推定される。同様に同綴りの形容詞を持つ動詞が約 400 ある。

Give up など句動詞については、ゲルマン語由来の英語本来の動詞が作る表現であり、ラテン語由来のロマンス系の動詞は句動詞

には参加しないということが言われている。LDOCE において句動詞は全部で約 2,500 あるが、それらに参加する動詞の種類は 400 である。

以上を図で示すと以下のようになる。



(2) 一般に通説として流布しているだけでなく、Goto (1971)および Miyawaki et al. (1975)の研究が実証的に明らかにしたように、日本人は L/R 聞き分けができない。しかし、日本で英語教育に携わっている人の多くは、Goto らの研究を知らず、日本人が L/R の聞き分けができないということが実際にどのようなことであるのか、あるいは本当のことであるのか、それは練習を積めば克服できる問題であるのか、そのハンディキャップはどの程度の損害を日本人にもたらしているのかといったことについて、はっきりとは知らない。

辞書の中から L/R の最小対を網羅的に抽出することによって、L/R を聞き分けられない日本人は英語を使う上でどのようなハンディキャップを負っているのかを推定することができる。これによって中級レベルまでは大きな不利にはならないが、上級になると L/R を区別できないことは、大きな障害になることが分かった。すなわち、最小対の数が上級になるにつれ多くなり、その中には clash -- crash のようにオノマトペの最小対が含まれる。

(3) 辞書から機械的に同綴りの動詞と名詞の組を抽出し、動詞と名詞の意味的関係について観察を試みた。(LDOCE に含まれるそのような組の数は 2600 以上あり、未だすべてについて観察ができていない。) 意味的関係で最も多いのは、travel (旅行、旅をする)のように、名詞が行為そのもの (the act of V-ing) を意味するものであった。他には、author (著者、本を書く)のように特定の人間とその人間が行う行為の組合せや、hammer (金槌、金槌で打つ)のように道具とその道具を用いて行う行為の組合せなどの類型が見られることが分かった。「転換」(conversion)と呼ばれるこの派生関係は、語彙表構築の際に十分考慮されるべきであることが分かった。

(4) Beth Levin (1993) の English Verb Classes and Alternations においては、動詞をそれが参加できる diathesis alternations に基づいて分類することが試みられた。この研究は動詞の意味とそれが現れる統語的枠組みの関係を追及する試みとして注目されているが、未だこの分類は完成していない。LDOCE の動詞項目では、diathesis alternation の多くが直接的に記載されている。またその記述方針は非常に首尾一貫している。たとえば、break の意味 1 において [transitive] I had to break the window to get into the house. と [intransitive] He kept pulling at the rope until it broke. の両方が併記されている。このことは、英語動詞の多くの diathesis alternation が LDOCE から機械的に抽出できることを示唆している。このことは、動詞を意味に基づいて分類する一つの方法となる。

＜参考文献＞

Hiromu Goto. 1971. Auditory perception by normal Japanese adults of the sounds L and R. *Neuropsychologia*, vol.9, pp.317-323.

Miyawaki, Kuniko et al.1975. An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception and Psychophysics*, vol.18, pp.331-340.

Beth Levin. 1993. English Verb Classes and Alternations: A Preliminary Investigation. The University of Chicago Press.

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 0 件）

〔学会発表〕（計 1 件）

園田勝英、L/R と小学校英語教育、JACET 英語語彙研究会、2016 年 3 月 5 日、東京電気大学千住キャンパス(東京都)

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 0 件）

名称：

発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
園田　勝英（SONODA, Katsuhide）
北海道大学・メディアコミュニケーション研究院・特任教授
研究者番号：７０１１３６９４

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
（　　）