科研費
KAKENHI

科学研究費助成事業　　研究成果報告書

令和　２ 年　６ 月　９ 日現在

機関番号：３４５０４

研究種目：基盤研究(C)（一般）

研究期間：2017～2019

課題番号：１７Ｋ０２７６１

研究課題名（和文）The Corpus of Kansai Vernacular Japanese

研究課題名（英文）The Corpus of Kansai Vernacular Japanese

研究代表者

ＨＥＦＦＥＲＮＡＮ　Ｋ（HEFFERNAN, KEVIN）

関西学院大学・総合政策学部・教授

研究者番号：６０５８０５９５

交付決定額（研究期間全体）：（直接経費）　　1,500,000円

研究成果の概要（和文）：研究目的は、関西地方の日本語のコーパスを作成し、そのコーパスをインターネットで利用できるようにすることです。合計で、138の社会言語学インタビューを行いました。各インタビューを書き起こし、Mecabを使用して品詞データを付けました。全ての品詞情報データを手動的にチェックし、間違いを修正しました。最終的な正解率は約98%と推定します。
修正されたデータや文字起こし方法の説明書などをGoogleのWebサイトで公開しました。データはクリエイティブ・コモンズライセンスの下で共有されています。ユーザーは無料で使用できます。ただし、ユーザーが営利目的でデータを使用することは禁止されています。

研究成果の学術的意義や社会的意義
The corpus of Kansai Vernacular Japanese includes speakers ranging in age from 15 years old to 80 years old. This large age range allows researchers to explore how language is used differently by each generation. Such an approach is useful for doing research language change and standardization.

研究成果の概要（英文）：The two primary objectives of this research project were to create a corpus of Kansai vernacular Japanese, and to make that corpus available on the internet. In total, 138 sociolinguistic interviews were conducted. Each interview was transcribed, and checked for errors. The transcriptions were parsed and tagged with part of speech data using Mecab. The tagged data was checked by students hired for this job, and mistakes were corrected. I estimate the final accuracy rate is about 98%.
The final data, along with supporting documents such as a description of the transcription methods, are available on a google website. The data is shared under a creative commons license. Users may be downloaded and used free of charge. However, users are prohibited from using the data for profit.

研究分野：linguistics

キーワード：Japanese　dialect　corpus

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。

１．研究開始当初の背景

One of the goals of Variationist linguistics is to explore the way language changes over time. A popular methodology for doing so is the apparent-time method. This method requires that the researcher gather data from speakers who span a wide age range, from as young as teenagers to elderly people. The researcher then observes how people of different age groups speak differently. As a simple example, consider the use of the word なんか. This word has increased in usage among younger people. Furthermore, carefully observing the differences in usage between younger and older speakers shows that the grammatical changes of なんか parallel the English word *like*. This parallelism suggests that there is a universal mechanism common to all languages of the world that controls the way words change into discourse markers such as なんか and *like*.

When this project was started, there was very little data natural conversation data available for research, and the available data did not cover a wide age range. Furthermore, this project collected conversations between speakers of the Kansai dialect. Previously, there was no large scale publicly available dataset of Kansai speech.

２．研究の目的

The two primary objectives of this research project were to create a corpus of Kansai vernacular Japanese, and to make that corpus available on the internet for others to use. Furthermore, a wide age range of speaker data was collected in order to allow for apparent-time studies of the way that spoken Japanese, and in particular Kansai dialect, was changing from one generation to the next.

In order to make the data easy to use by others, it is shared under a creative commons license. This license allows anyone to download the data free of charge and use it for non-commercial purposes. Also, if the user changes the data (for example, improves it), then he or she must share the data free of charge with others.

３．研究の方法

In total, 138 sociolinguistic interviews were conducted. The interviews were conducted by students attending a university in Kansai for course credit. The students were trained by researcher to conduct an interview following established sociolinguistic methods. The primary goal of such an interview is to have the interviewee relax and speak in a very casual style. All of the interviewers and interviewees are native Kansai Japanese speakers. Each interview was transcribed, and checked for errors. The transcriptions were parsed and tagged with part of speech data using Mecab. The tagged data was checked by students hired for as part-time workers, and mistakes were corrected. The final accuracy rate of the part of speech information is about 98%. Furthermore, each line of data is marked for the speaker (either interviewer or interviewee) to allow researchers to easily sort out the speech of the student interviewers from the interviewees.

４．研究成果

The primary result of this project is the interview data. The final data, along with supporting documents such as a description of the transcription methods, are available on a google website. Following is a sample of ten lines of data. The entire data collection is over two million lines long.

```
図書館	s,名詞,一般,*,*,*,*,図書館,トショカン,トショカン
の	s,助詞,格助詞,一般,*,*,*,の,ノ,ノ
人	s,名詞,一般,*,*,*,*,人,ヒト,ヒト
と	s,助詞,格助詞,一般,*,*,*,と,ト,ト
喋る	s,動詞,自立,*,*,五段・ラ行,基本形,喋る,シャベル,シャベル
なんて	s,助詞,副助詞,*,*,*,*,なんて,ナンテ,ナンテ
ありえ	s,動詞,自立,*,*,一段,未然形,ありえる,アリエ,アリエ
へんかっ	s,助動詞,*,*,*,特殊・ナイ,連用タ接続,へん,ヘンカッ,ヘンカッ
た	s,助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。	s,記号,句点,*,*,*,*,。,。,。
```

Other than that data, several research articles and a book were also produced. The research articles investigated two areas: language change and language processing. With regards to language change, the following topics were reported on: changes in the grammatical nature of the word みたい; changes in *u-onbin* in Kansai dialect; and the spread of lexical phrases of such as みたいな感じ from a large urban area to a small rural area.

With regards to language processing, the following topics were reported on. The omission of the case particles を and に in spoken Japanese; and the processing of the negative forms of verbs in spoken Japanese. Examining these language phenomena in details shows us how language is processed by our memory.

Finally, *The grammar of Kansai vernacular Japanese* was published by the K.G. University Press. This book is an introduction to Kansai Japanese written in English with Japanese examples, an aim at intermediate level learners of Japanese. The book covers a wide range of grammar of spoken Japanese, including Kansai dialect, but also grammar used throughout Japan. Some examples of the Kansai dialect grammar covered by the book are the copula *ya*, the verbal negative suffixes -へん, -ひん, -ん, and the word あかん. Some examples of general spoken Japanese covered by the book are 的 (as in 僕的には) and the emphasis of adjectives by double consonants (as in いったい！). Each grammatical point is illustrated with several examples of natural spoken Japanese taken from the corpus data. Furthermore, each grammatical point is rated for frequency of usage based on the rate of usage in the corpus data. This rating system allows beginner students to focus on the most commonly-occurring phrases only.

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　2件／うちオープンアクセス　1件）

１．著者名 Heffernan Kevin、Sato Yo	４．巻 3
２．論文標題 Relative frequency and the holistic processing of morphology	５．発行年 2017年
３．雑誌名 Asia-Pacific Language Variation	６．最初と最後の頁 67～94
掲載論文のDOI（デジタルオブジェクト識別子） 10.1075/aplv.3.1.04hef	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Heffernan Kevin、Hiratuka Yusuke	４．巻 3
２．論文標題 Morphological relative frequency impedes the use of stylistic variants	５．発行年 2018年
３．雑誌名 Asia-Pacific Language Variation	６．最初と最後の頁 200～231
掲載論文のDOI（デジタルオブジェクト識別子） 10.1075/aplv.16009.hef	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計7件（うち招待講演　0件／うち国際学会　7件）

１．発表者名 Heffernan, Kevin and Yo Sato
２．発表標題 The varying complexity of the syntactic role of nouns: Evidence from Japanese corpora
３．学会等名 4th Asia Pacific Corpus Linguistics Conference（国際学会）
４．発表年 2019年

１．発表者名 Sato, Yo, and Kevin Heffernan
２．発表標題 How distinctive is your corpus? A metric for the degree of deviation from the norm
３．学会等名 4th Asia Pacific Corpus Linguistics Conference（国際学会）
４．発表年 2019年

１．発表者名
Yo Sato and Kevin Heffernan

２．発表標題
Creating dialect sub-corpora by clustering: a case in Japanese for an adaptive method

３．学会等名
Eleventh International Conference on Language Resources and Evaluation（国際学会）

４．発表年
2019年

１．発表者名
Heffernan, Kevin, Imanishi, Yusuke, Honda, Masaru, and Sato, Yo

２．発表標題
Case particle omission correlates with object syntactic complexity: Evidence from the Corpus of Kansai Vernacular Japanese.

３．学会等名
Exploiting Parsed Corpora: Applications in Research, Pedagogy, and Processing（国際学会）

４．発表年
2017年

１．発表者名
Heffernan, Kevin

２．発表標題
The diffusion of lexical bundles from an urban center to a rural community in Japan

３．学会等名
16th International Conference on Methods in Dialectology（国際学会）

４．発表年
2017年

１．発表者名
Heffernan, Kevin, and Sato, Yo

２．発表標題
Age-related patterns in lexical bundle usage: Evidence from a corpus of vernacular Japanese

３．学会等名
Corpus Linguistics International Conference（国際学会）

４．発表年
2017年

１．発表者名 Sato, Yo, Heffernan, Kevin, Kishie, Shunsuke, and Hattori, Kota
２．発表標題 Quantifying ‘standardness’ of the language use in a locality: a study with Twitter data
３．学会等名 Corpus Linguistics International Conference（国際学会）
４．発表年 2017年

〔図書〕　計1件

１．著者名 Kevin Heffernan	４．発行年 2019年
２．出版社 関西学院大学出版会	５．総ページ数 174
３．書名 The Grammar of Kansai Vernacular Japanese	

〔産業財産権〕

〔その他〕

The Corpus of Kansai Vernacular Japanese
https://sites.google.com/view/kvjcorpus

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考