「クックパッドデータセット」の言語学的特性

長谷川　守寿

1. 目的

料理レシピサイト「クックパッド」(cookpad.com)は、1998年3月の開設以来、登録数が増加し、2020年4月7日現在約326万件のレシピが登録されている。本研究では、国立情報学研究所のIDRデータセット提供サービスによりクックパッド株式会社から提供を受けた「クックパッドデータセット」を利用し、2014年9月30日以前に公開されたレシピと献立に関するデータを対象に、「クックパッドデータセット」とは、どのようなコーパスなのかを、言語学的観点から定量的・定性的分析により明らかにし、今後どのような研究に使用することが出来るか、使用する際の注意点は何かを考察する。

2. 先行研究

クックパッドデータを用いた研究には、料理名に出現する英語の前置詞を研究した小野ほか(2017)や小野(2019)、「レシピ」に出現するオノマトペを研究した福富ほか(2018)、「レシピ」を元に料理を作った際の感想が書かれた「つくれぽ」に現れるオノマトペを調べた楊(2020)があるが、クックパッドデータ全体を言語学的観点から概観した研究は管見の限り見当たらない[1]。

クックパッドデータは、リリースノート（release_note.txt）によると、2015/02/24から配布されているデータで、クックパッド株式会社より提供されたレシピデータおよび献立データである。データの詳細はデータ仕様書（ckpd_data_spec.pdf）に記載されているが、A4用紙7枚程度の文書であり、内容は「１．クックパッドデータの利用方法」「２．クックパッドデータの仕様」「３．補足事項」だけである。そこで、どのようなデータが含まれているのかを明らかにすることにより、料理名やオノマトペだけでなく、どのような研究目的での使用が可能か、さらに使用する際にはどのような注意点があるかも明確になると思われる。

3. 方法

3.1. クックパッドデータについて

データ仕様書の「２．クックパッドデータの仕様」によれば、「データベースには、12個のテーブルが含まれ」、「6個はレシピに関するテーブル、残り6個は献立に関するテーブル」である（以後、日本語訳も同書に従う）。

レシピに関するテーブルは、「recipes」（レシピ）「ingredients」（材料）「steps」（手順）「tsukurepos」（つくれぽ）「search_categories」（レシピのカテゴリ）他１である。献立に関するテーブルは、「base_kondates」（献立）「base_kondate_recipes」

[1] コーパスや、コーパスを用いた様々な研究については石川(2012)他を参照のこと。

（献立のレシピ）「user_kondate_infos」（献立の作者やポイントなど）他３である。

それぞれのテーブルは、さらに複数のカラム[2]と内容に分かれ、カラムは計41種である。その内で「user_id」のようにカラム名にIDがつくもの（21種）は、後述する一つの例外を除き、「ad7d585b06850f8437ff5fb97d3c7a823ff21bb1」のようにユーザー名を特定されないように全て数字・アルファベットでマスキングされているので、調査する必要はない。その例外である「dish_type_id」はレシピのタイプを示し、主菜（0）か副菜(1)の二値しか持たない。さらに「published_at」（レシピの公開日）、「position」（手順の位置）、「entered_at」（つくれぽの投稿日）、「cooking_time」（献立の調理時間）、「published_at」（献立の公開日）も数字だけである。よってどのようなデータが含まれるのか調査する対象は、残りの14種である。考察対象を表１にまとめる。

なお、quantityは100ｇや１本のような分量を表すものだけかと想像したが、実際は想像と異なるものが見られたので、調査対象として取り上げる。

表１　本調査の考察対象

テーブル名	カラム名	内容
recipes	title	レシピのタイトル
	description	レシピの概要
	serving_for	レシピの分量
	advice	レシピのコツ・ポイント
	history	レシピの生い立ち
ingredient	name	材料の名前
	quantity	材料の分量
steps	memo	手順の内容
tsukurepos	message	つくれぽの内容
	comment	レシピ作者のコメント
search_categories	title	カテゴリのタイトル
base_kondates	title	献立のタイトル
user_kondate_infos	description	献立のポイント
	setup_tips	献立のコツ

（ckpd_data_spec.pdfより）

3.2. 手順

頻度を元にした量的調査と、頻度だけでは捉えられない特徴を目視で抜き出す質的調査を行う。そのためまずクックパッドデータからカラム別にデータを抽出しファイルを作成する(3.2.1)。次にファイルの特徴を明らかにするため、参照データとなるコーパスを準備する(3.2.2)。そしてファイル間の関係を明らかにするため、クラスター分析[3]を行い(3.2.3)、参照データと各ファイルの特徴語を抽出する(3.2.4)。本調査では、書き言葉の典型的な例である新聞を参照データとする。クックパッドデータは、料理という限定された分野であり、文自体には修正が加わっておらず、また印刷されていないという特徴がある。この特徴が表現としてどのように現れるかを、本調査では極端なもの同士を比較することによって明らかにしたいと考え、対照とするものとして新

[2] 「行」と「列」で表現される表の「列」に相当する。

[3] データの各個体間の類似度を距離とみなして、その距離を基準にグループ分けしていく手法（石田 2017, p. 139）である。

聞を選んだ。新聞は多種多様な内容を含む文書であり、校正を経ていることで文法的な正しさが保障され、印刷されている点もクックパッドデータとは対照的で、比較対象としては相応しいと判断した。

3.2.1. 対象ファイルの抽出

クックパッドデータのそれぞれのカラム（表1）から、量的調査用データ（各1000件）と目視用データ（各100件）を抜き出す。量的調査用データは、各カラムの大まかな特徴を捉えるためにクラスター分析と関連語検索に使用する。目視用データは、量的調査では捉えられない細かい特徴を実際に著者が目視して明らかにするために使用する。例えば、助詞・助動詞などは、KH Coder[4]の中では「その他」として扱われ、デフォルトの設定では集計の対象にならない。また品詞別に特徴語を抽出するが、後述するように品詞には制限を加えている。KH Coderでは対象外となったものでも特徴的な語を明らかにするため、目視用データを準備する。

ここでは量的調査用データ（1000件）の抜き出し方を説明する。無作為抽出を行うために、各カラムのデータ件数を調べ、その値から999を引き、1からその数までの間で乱数を一つ発生させ、その数の場所から1000のデータを連続で取り出す。中には「なし」とだけ入力されているものなどもあるが、それらも一つとして数える。なお、目視用データもデータ数が異なるだけで、抽出方法は同一である。

3.2.2. 参照データの作成

現段階で最新のデータである「CD-毎日新聞2018年」[5]から、無作為に1000件記事を抽出する。本文のみを対象とし、見出しは含まない。具体的には層別抽出法は行わず、無作為抽出法を使用し、重複なしで発生させた数字と同じ記事IDを持つ記事を抜き出す。その結果、政治・経済・国際・スポーツ・時評・芸能も含んだデータができあがることになる。抽出した新聞記事1000記事を茶まめ[6]（ver2.0）を用いて形態素解析した結果、動詞の上位10語（1）は、『現代日本語書き言葉均衡コーパス』語彙表（https://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html,2019/12/01取得）に含まれる「出版・新聞(PN)」の動詞上位10語（2）とほぼ同じ結果となったため、比較対象としては1000記事で充分と判断する（下線は（1）（2）で共通する語）。

（1） 為る、居る、有る、言う、成る、来る、思う、見る、行く、出来る
（2） 為る、居る、有る、成る、言う、出来る、因る、来る、見る、行く

[4] 樋口耕一氏が開発したフリーソフトウェア。形態素解析器、形態素解析用辞書、統計エンジンを内包しており、１つのソフトウェア上で、データの形態素解析から各種の検索、さらには頻度データを用いた高度な統計分析までを連続的に行うことができる（石川 2012）。

[5] 「CD-毎日新聞 2018 年」は筆者が毎日新聞社と交わした利用許諾契約・覚え書きに基づき使用した。

[6] 各種の形態素解析用辞書 UniDic が使用できる形態素解析器。

3.2.3. クラスター分析

語の出現が似通っているカラムは何かを明らかにするために、クックパッドデータの14のカラムを対象にクラスター分析を行う。ファイル名は表１の内容をそのまま使用する。クラスター分析にはRMeCab[7](石田2017)を用い、ファイルから取り出す品詞は「動詞・名詞・形容詞」に限定し、距離の定義は平方ユークリッド法、クラスターの結合法はウォード法を使用する。

3.2.4. 特徴語抽出

それぞれのカラムに特徴的な語を抽出するために、KH Coder(ver3.Alpha.11b)を使用する。KH Coderの関連語検索を用いると、「特定のコードと強く関連しているのはどんな語かを容易に探索できる」(樋口2014、p.147)。本稿では「特定のコード」がカラムに該当するが、それぞれのカラムを特徴づける語の抽出に関連語検索を用いる。

前処理で行う形態素解析には茶筌を使用し、強制抽出する語には(3)のように平仮名・片仮名で表記される料理名・素材名や、「つくれぽ／れぽ」のようにクックパッドデータで多く使われる語を含めた(“／”は複数のデータを示す際に用いる。以下同)。これは、例えば「もやし」などは、デフォルトの設定では動詞「もやす」として処理されて、誤った形態素解析結果となってしまうからである[8]。

(3) あさり／いちご／うなぎ／えのき／おせち／つくれぽ／れぽ／レポ、

前処理終了後、特徴語抽出を行い、各カラムと新聞データの比較を行う。特徴語抽出では、動詞・名詞・形容詞それぞれに分けて抽出する。それぞれの品詞に含まれる下位分類は(4)(5)(6)である。形容詞には、いわゆる形容動詞も含める。【】内は、茶筌の出力による品詞名：筆者による注釈である。

(4) 名詞【名詞-一般：漢字を含む２文字以上の語】、名詞B【名詞-一般：平仮名のみの語】、名詞C【名詞-一般、漢字１文字の語】、サ変名詞【名詞-サ変接続】、固有名詞【名詞-固有名詞-一般】、組織名詞【名詞-固有名詞-組織】、人名【名詞-固有名詞-人名】、地名【名詞-固有名詞-地域】、タグ
(5) 動詞【動詞-自立：漢字を含む語】
(6) 形容動詞【名詞-形容動詞語幹】、ナイ形容【名詞-ナイ形容詞語幹：「頼りない／

[7] 石田基広氏が開発したパッケージで、Rから日本語のテキストやファイルを指定してMeCabに解析させ、その結果をRで標準的なデータ形式に変換して出力するインターフェイス（石田2017,p.20)。

[8] 他にも形態素解析システムでは「。」「！」「？」などを文の切れ目とし、「♪」や顔文字などは文末としては処理されない。そのため「いい感じですね♪」等の場合「すね」を一語として認定してしまう等の問題も発生している。より精度の高い結果を導くために、形態素解析の修正は避けて通れない。

問題ない」等】、形容詞【形容詞：漢字を含む語】、形容詞B【形容詞：平仮名のみの語】、形容詞（非自立）【形容詞-非自立：「がたい／つらい／にくい」等】

ここで動詞B【動詞-自立：平仮名のみの語】は対象外とする。例えば新聞には「貴重（きちょう）」のような表記があるが、「ちょう」を動詞「ちる」と誤解析される問題が起こる。またクックパッドデータには「ぶっかけ」のような平仮名表記が多いが、「かけ」を動詞「かく」と解析する誤りが多発する。そのため、より正しい結果が得られるように平仮名のみで表記された動詞は排除する。また「ある／いる」のような「動詞-非自立可能」は頻出する語であり、特徴語とならないため対象外とする。

なお、クラスター分析では、動詞・名詞・形容詞の全ての下位分類を対象とするが、特徴語抽出では、品詞の一部の下位分類に制限する。また特徴語抽出では、KH Coderの結果をそのまま示し、異表記の統合は行わない。

4. 結果

4.1. クラスター分析の結果

クラスター分析の結果図１となった。図上方の名前・タイトルに関するカラムが一つのクラスター（①）をなし、分量のカラムが一つになる（②）など、ファイル名に「名前」「タイトル」、「分量」をつくことからも当然と言えば当然の結果であるが、下に位置する８種のカラムについては、「献立のポイント」と「レシピの生い立ち」が近かったり、「レシピの概要」と「つくれぽの内容」が近いなど、想像には反するものが存在していると思われる。以後、各カラムについて、量的分析と質的分析を加えていく。

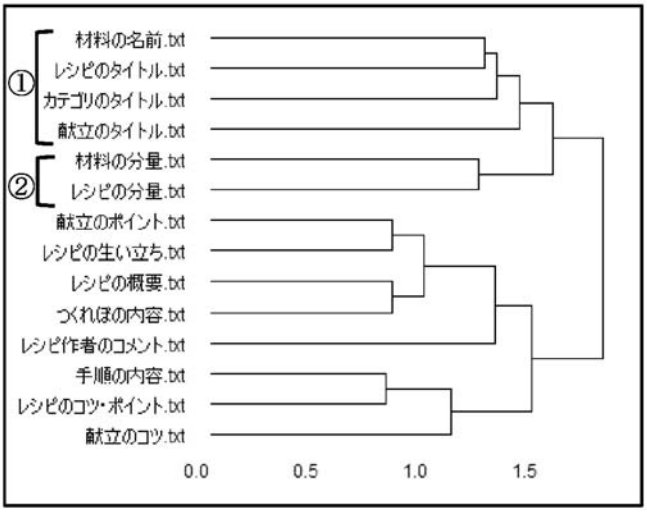

図１　クラスター分析の結果

4.2. 関連語検索の結果

全てのカラムと比較した新聞の特徴を述べ、その後図１の上から各カラムを対象に考察する。なお本稿の例文の出典は全てクックパッドデータからなので省略する。

4.2.1. 新聞の特徴

新聞と各カラムを比較した場合、新聞を特徴づける語はほぼ共通していて、表２の

ようになる（Jaccard係数[9]が.01以上のものを示す。以下同じ）が、異なる場合もある。

動詞に異同が見られるのは「レシピのコツ・ポイント」と比較した場合であり、(7)のような例が見られるため新聞を特徴づける語に「思う」がなく、「レシピの生い立ち」と比べると(8)のような例が見られるため、新聞を特徴づける語に「思う／言う／見る」がない（下線は該当箇所を示す。以下同）。名詞に違いが見られるのは、「献立のコツ」と比較した時であり、(9)のような例があるため新聞を特徴づける語に「先」がない（「米」はアメリカを示す）。

形容詞は三つ（ない（.027）／問題(.017)／多い(.010)）のみであった。形容詞に異同が見られるのは、「レシピのコツ・ポイント」「レシピの生い立ち」と比較した場合であり、(10)のような例文が見られるため、新聞を特徴づける語に「ない」がない。同様に「献立の記述」と比較した場合、(11)のような例文が見られるため、新聞を特徴づける語に「多い」がない。

表２　新聞を特徴づける語

動詞		名詞	
思う	.014	日本	.032
受ける	.013	大阪	.028
行う	.012	東京	.023
言う	.010	米	.021
求める	.010	人	.018
見る	.010	野球	.014
出る	.010	大会	.013
話す	.010	先	.013
		世界	.013
		首相	.013

(7) 粉チーズは多めがいいと思います
(8) 料理コーナーを見て。（略）酢が（略）使いやすいと言っていたので（略）
(9) はんぺんとうま煮を先に準備し、ソテーの仕上がりにあわせて配膳します
(10) ないです！量は適当なのでお好きに調整してください☆
(11) 冷蔵庫の有りもので頑張ったら、統一感ないけど、品数多く（略）

4.2.2. 材料の名前

「材料の名前」は、動詞・形容詞の出現がない（少ない）のが特徴として挙げられ、形容詞は「やさしい」(.012) だけで、これは「ミツカンやさしいお酢」という商品名の一部として出現する。材料の名前を特徴づける語として抽出された名詞は(12)で、主に「塩／砂糖／醤油／酢」などの調味料名である。

目視用ファイルを見ると、「しょうが／ニンニク／ねぎ／豚肉／にんじん」など材料名も見られるが、数値的には低い。これは材料名を表す名詞は数多いことと、また材料名の場合「にんじん／ニンジン／人参」のように表記の多様性があり、現状の処理では一つにまとめていないため、数値が低くなったと考えられる。なお、上位10語で「材料の名前」と「レシピのタイトル」で共通する名詞は「塩」のみであった。

(12) 塩(.069)、砂糖(.035)、水(.030)、醤油(.025)、卵(.023)、酢(.021)、ごま油(.019)、にんにく(.019)、酒(.018)、こしょう(.017)

[9] データ集合の類縁性を評価する類似度系の指標の１つで、情報理論の分野でも幅広く使用されている（石川 2012）。数式やそのほかの係数については石川(2012)を参照のこと。

4.2.3. レシピのタイトル

「レシピのタイトル」は「材料の名前」と近いものと考えられる。名詞で抽出されたのは(13)であり、「麹」のような一時期流行したものを含み、健康志向がうかがえる。動詞は「炒める(.063)／煮る(.025)／和える(.015)／焼く(.012)／揚げる(.011)」が抽出されたが、実際には(14)のように料理名の一部で使われる。また形容詞は「簡単(.132)」のみで(15)などで使われている。(16)のように記号・顔文字を含むタイトルや、オノマトペを含んだタイトル(17)が多いのも特徴といえる。

(13) サラダ(.057)／トマト(.048)／塩(.041)／野菜(.039)／麹(.033)／豚(.030)／カレー(.029)／肉(.029)／ソース(.028)／豆腐(.026)
(14) 簡単♡♥鮭&野菜のほっこり煮／長芋のカレーチーズ焼き
(15) 簡単焼き肉／ほったらかしで簡単！さつま芋レモン煮
(16) ✿ねぎ塩入り♡ミニミニハンバーグ✿／カレイ…薄味のウマウマ…(o˘◡˘o)♡
(17) しっとりホクホク栗きんとん

4.2.4．カテゴリのタイトル

「カテゴリのタイトル」はレシピのカテゴリに関するもので、カテゴリは1098種類挙がっている。名詞の中で特徴づける語として抽出されたのは(18)である。この中で他のカラムで見ることのない「再現」や「運動会」を取り上げる。これらは(19)のように何かを「再現」したメニューや、(20)のように「運動会」に適したメニューを表す際に使われている。また、目視用ファイルを見ると、(21)のように材料名そのままのものが入力されているものもあり、「材料の名前」と近い位置に存在する理由と考えられる。動詞は「使う(.037)」一つであったが、(22)のように使われる。

(18) 再現(.040)／パン(.038)／おかず(.034)／サラダ(.026)／お菓子(.023)／ケーキ(.020)／アレンジ(.020)／鍋(.018)／運動会(.017)
(19) 再現バーガー／再現牛丼／ラーメン店の再現メニュー
(20) 運動会用主食／運動会用いなり寿司／運動会用のり巻き
(21) 豆腐／トマト／じゃがいも／チーズ／鮭／ビーフン／フォー
(22) 挽肉を使ったカレー／火を使わない夏料理

4.2.5. 献立のタイトル

「献立のタイトル」を特徴づける語として、名詞で抽出されたものは(23)である。「ごはん／ご飯／御飯」が別語として挙がっているが、多様な表記を一つに集計した場合さらに大きな値になる。動詞では「食べる(.021)／炒める(.016)／煮る(0.12)」が抽出されたが、「幼児食」「野菜炒め」「じゃが煮」のように料理名の一部を構成している。形容詞は「簡単(.034)／美味しい(.020)」の二語で(24)が使用例である。

また「弁当」が多く含まれ、目視用ファイルを確認すると(25)のような「弁当」を

省略した表現も多数見られ、これらも集計すると「弁当」の値はさらに高くなる。

(23) 弁当(.144)／晩(.111)／ごはん(.085)／夕食(.079)／ランチ(.071)／ご飯(.069)／朝食(.068)／献立(.054)／御飯(.046)／夕飯(.044)
(24) 簡単美味しい夕食
(25) 娘弁／キャラ弁／マイ弁

4.2.6. 材料の分量

「材料の分量」を特徴付ける語の中に、動詞・形容詞で.01以上のものはなく、名詞は(26)の６語のみであった。(27)のような場合「大さじ」は一語、「大匙」は「大」と「匙」の２語として解析しているため、「大さじ」の値は小さくなっているが、統合した場合もう少し大きな値となる。分量では「1枚／1個／1袋／1リットル」(茶筌の品詞では「名詞-数」「名詞-接尾-助数詞」) といったデータが多いのであるが、これらはKH Coderのデフォルトの設定では「その他」の品詞として扱われるため、特徴語としては現れてこない。また目視用ファイルを見ると、「少々」といった入力が多数ある。さらに(28)のように分量といえるのかどうか怪しい入力もある。

(26) 大さじ(.129)、適量(.084)、小さじ(.070)、カップ(.022)、好み(.020)、匙(.013)
(27) 大さじ１／大匙１
(28) 前日おかずの残り／丼茶碗分／好きなだけ

4.2.7. レシピの分量

「レシピの分量」を特徴づける語は非常に少なく名詞３語のみ抽出されている（直径(.014)／丸(.013)／パウンド(.011)）が、これは「材料の分量」同様、KH Coderのデフォルト設定では、数詞や助数詞が抽出されないためである。「材料の分量」が料理の材料の分量が示されていたのに対して、「レシピの分量」では「2人分」のように、できあがる料理の量を示している。人数が記述されているものや、「約24個分」等どのくらいの個数が出来るかという記述等、様々な記述が見られるが、空欄である場合も多い。またデータからはスイーツに関するレシピが多いことが窺える。

クラスター分析では、数詞や助数詞のような名詞の下位区分も使っているため、料理の分量とレシピの分量が近くに位置づけられたものと考えられる。

4.2.8. 献立のポイント

「献立のポイント」を特徴付ける語を示した表3の網掛け部分は、表4と共通している部分であるが、動詞・名詞・形容詞に共通点が多い。また、「晩ご飯」のような語の区切り方に関しては「献立のタイトル」と同じような問題がある。さらに特徴語の名詞の中に「冷蔵庫」が含まれるが、これは(29)のように冷蔵庫の中のもので料理を作

りたいというニーズに応えるものと思われる。

文の内容として、献立の何をメインに据えそれはなぜかという文が多く、(30)のように接続助詞の「ので」の使用も特徴としてあげられる。後述するが、(31)のように特定の季節のデータが選ばれてしまった問題点も見られる。

表3　「献立のポイント」を特徴づける語

動詞		名詞		形容詞	
作る	.112	野菜	.117	美味しい	.106
食べる	.109	献立	.112	簡単	.074
使う	.059	レシピ	.092	暑い	.049
出来る	.029	冷蔵庫	.053	大好き	.026
入れる	.023	ご飯	.049	ヘルシー	.025
頂く	.023	弁当	.044	好き	.023
買う	.018	ごはん	.043	メイン	.017
置く	.017	晩	.032	良い	.014
合う	.016	菜	.031	おいしい	.013
行く	.015	肉	.031	美味	.011

(29) 冷蔵庫の余り物で／冷蔵庫にあるもので／冷蔵庫の中のもので

(30) じゃが芋が沢山あったので、じゃが芋消費に肉じゃがを作りました。

(31) 暑い夏なので…ガスコンロ不使用で作りました＾＾；／暑いので、出来るだけ時短で晩御飯＾＾／夏らしくざる蕎麦をメインにしました。

4.2.9．レシピの生い立ち

前述の通り「献立のポイント」と「レシピの生い立ち」を特徴づける語はかなり重複している。「レシピの生い立ち」では、(32)のようにどのような理由でこのレシピが作られたのかを表す文が多く、その際(33)のように「ない」を使った表現も多く見られる。また、構文的には「ので」を用いて理由を表す従属節(34)をもつ複文も多い。

表4　「レシピの生い立ち」を特徴づける語

動詞		名詞		形容詞	
作る	.232	レシピ	.083	美味しい	.107
食べる	.157	料理	.044	簡単	.061
思う	.077	味	.040	大好き	.043
使う	.046	アレンジ	.037	好き	.035
考える	.040	弁当	.028	ない	.028
入れる	.036	子供	.028	おいしい	.026
余る	.026	笑	.027	いい	.018
作れる	.024	野菜	.026	良い	.016
出来る	.018	冷蔵庫	.026	よい	.015
焼く	.018	サラダ	.025	手軽	.015

(32) 息子の離乳食を手軽に、かつ栄養を取れるようにしたいと思い考えました。

(33) 洗い物を増やしたくなかったから笑

(34) （略）肉がなく、唐揚げが残っていたので、

4.2.10. レシピの概要

「レシピの概要」は、後述の「つくれぽの内容」とかなり重複する部分が多い（網掛けは表5と表6で共通する部分）。また先行研究で扱われているように、オノマトペの使用(35)も見られる。自分のレシピで作ってもらいたいという意識の表れと見て

いいのか、「作ってみてね」(36)や「いかがでしょうか」等の表現を文末に持つものが多く見られる。

また(37)のようにどういう人向けのレシピなのかを明確にする表現も見られる。

(35) 煮るととろとろになり／でもフワフワだよ～♪
(36) 簡単なので作ってみてね＾＾
(37) 鶏肉好きの人にはたまらない料理です／うに好きにはたまらない一品です♪

表5　「レシピの概要」を特徴づける語

動詞		名詞		形容詞	
食べる	.084	味	.076	簡単	.173
作る	.070	野菜	.043	美味しい	.091
使う	.052	ご飯	.036	おいしい	.054
出来る	.033	弁当	.036	ヘルシー	.035
入れる	.027	スープ	.035	甘い	.031
作れる	.025	レシピ	.035	好き	.018
合う	.023	味噌	.032	いい	.017
焼く	.021	ソース	.027	大好き	.013
混ぜる	.020	チーズ	.026	よい	.013
煮る	.019	一品	.026	手軽	.012

4.2.11．つくれぽの内容

「つくれぽの内容」はレシピを元に料理を作った人の感想で、「デス・マス」形の使用が多く、作った後に書き込むことが多いので、タ形がほとんどである。また前出の(36)と(38)のように、「レシピの概要」と「つくれぽの内容」は同じような表現を用いた文も多いが、評価に関する部分が「つくれぽの内容」では独特で、(39)のように食べた人の反応を表した文が見られる。「レシピの概要」に近いということで、当然(40)のようなオノマトペの使用も多い。

表6　「つくれぽの内容」を特徴づける語

動詞		名詞		形容詞	
作る	.114	味	.051	美味しい	.373
食べる	.062	レシピ	.032	簡単	.110
入れる	.038	好評	.031	おいしい	.084
出来る	.034	子供	.024	美味	.043
頂く	.022	笑	.023	甘い	.028
喜ぶ	.018	娘	.023	大好き	.021
焼く	.014	感謝	.019	いい	.016
作れる	.012	弁当	.018	嬉しい	.016
助かる	.010	息子	.018	良い	.013
合う	.010	ご馳走	.018	手軽	.013

(38) 簡単に作れました！
(39) 家族に大好評でした／おいしいって喜んでいました。
(40) サクサク♡最高っす

4.2.12．レシピ作者のコメント

レシピを参考に料理を作ったユーザーの「つくれぽの内容」に対して、レシピ作者がコメントをつけたものが「レシピ作者のコメント」である。(41)のような「嬉しい」という形容詞を含む表現や名詞「嬉」を含む表現が多い。名詞の「れぽ／レポ」は「つ

くれぽ」の、「コメ」は「コメント」の略語であり、これらも特徴といえる。さらに「つくれぽ」には料理の写真が含まれることから0のような「素敵／綺麗」を含んだ表現が多く、美しさを表す形容詞が見られる。動詞では、いわゆる可能動詞に含まれる「貰える／頂ける」(41)が入っているのが特徴である。

表7　「レシピ作者のコメント」を特徴づける語

動詞		名詞		形容詞	
作る	.132	感謝	.220	嬉しい	.265
気に入る	.048	れぽ	.189	美味しい	.167
食べる	.042	レポ	.175	素敵	.111
頂ける	.036	つくれぽ	.132	美味	.076
合う	.035	感激	.059	良い	.029
喜ぶ	.026	嬉	.036	綺麗	.025
出来る	.014	コメ	.034	いい	.024
貰える	.014	口	.032	よい	.024
盛り付ける	.011	試し	.028	良い	.024
		是非	.027	おいしい	.023

(41) 活用してもらえて嬉しいです／そう言って頂けて嬉しいです／とても美味しそう～嬉♪
(42) 素敵な盛り付けレポありがとう。

4.2.13. 手順の内容

「手順の内容」と「レシピのコツ・ポイント」の網掛け部分はそれぞれに共通して出現する語であり、動詞と名詞に共通する箇所が多いことが分かる。手順の内容を示しているもの(43)も多いのであるが、(44)のように後述の「レシピのコツ・ポイント」の(47)に近いものも含まれている。これが図1でそれぞれが近くに位置する理由と考えられる。なお、目視用ファイルを確認すると、情報が入力されていないものも多い。また「よい」が二つ挙げられているのは、形容詞B「よく混ぜる」と形容詞（非自立）「ほどよく炒めて」を別々の集計しているためである。

表8　「手順の内容」を特徴づける語

動詞		名詞		形容詞	
入れる	.129	水	.051	軽い	.024
混ぜる	.074	塩	.044	薄い	.016
加える	.074	フライパン	.042	大きい	.015
切る	.072	火	.035	やすい	.013
炒める	.051	玉ねぎ	.031	厚い	.012
焼く	.049	油	.030	よい	.012
出来上がる	.037	鍋	.030	よい	.011
熱す	.025	オーブン	.027		
入る	.023	材料	.025		
作る	.018	卵	.024		

(43) 手羽元に塩コショウし、小麦粉をまぶす。圧力鍋にサラダ油を熱し、手羽元の両面にこんがりと焼き目をつけ取り出す。
(44) 折り曲げた先を、しっかり奥まで挟み込むのがコツです。

4.2.14. レシピのコツ・ポイント

「レシピのコツ・ポイント」には、(45)のように「いい／OK」など判断を表す文

が多い。

また目視用ファイルを見ると、(46)のように「楽ちん／楽チン／楽ちーん」など表記に揺れはあるが、楽チンという言葉の多用が目立つ。これは形態素解析の問題で抽出できなかったものと思われる。「めんどい」(48)という「面倒くさい」の略語など様々な表現が見られる。理由文もあるが、多様な条件文が多用されているのも特徴である。

表 9　「レシピのコツ・ポイント」を特徴づける語

動詞		名詞		形容詞	
入れる	.111	好み	.066	美味しい	.072
焼く	.054	味	.051	ない	.031
使う	.049	ＯＫ	.037	やすい	.028
作る	.043	火	.030	おいしい	.028
食べる	.037	量	.028	簡単	.026
思う	.035	水	.027	いい	.024
混ぜる	.034	調整	.025	好き	.019
加える	.028	野菜	.023	いい	.017
炒める	.023	塩	.023	甘い	.017
切る	.022	卵	.023	よい	.017

(45) 目玉焼きとか乗せてもいいかもです♪／（略）ラップでもＯＫ

(46) キャベツはちぎって入れれば切る手間も省けて楽チンです。

(47) 鶏肉は意外と火が通りにくいので（略）弱火でじっくり焼くのがコツです。

(48) 自分はしょうがやニンニクはめんどいのでチューブのやつを使用（笑）

4.2.15. 献立のコツ

「献立のコツ」を特徴づける語は動詞・名詞のみで、形容詞はなかった。名詞「完成／準備」もこのカラムでのみ現れる語で、(49)のように使われる。また「冷蔵庫／レンジ」という電気製品は表11に出ているが、(50)のように「レンチン」という「レンジでチンする」の略語も加えると「レンジ」はさらに高い値になる。

また目視用ファイルを見ると、「～しながら(51)」「～てから」「～する間」を用いた複文や、「あらかじめ」「先に」等の副詞を用いた文が多いことに気づく。また「時短・時短する(52)」なども多用されている。

表 11　「献立のコツ」を特徴づける語

動詞		名詞	
作る	.369	サラダ	.116
焼く	.123	スープ	.068
炒める	.084	冷蔵庫	.062
切る	.079	味噌汁	.054
茹でる	.078	野菜	.053
煮る	.063	完成	.044
仕上げる	.052	ご飯	.043
和える	.049	トマト	.042
盛り付ける	.042	準備	.038
冷やす	.042	レンジ	.038

(49)（略）ツナゴマ和えを完成させます。（略）準備ができたら巻き寿しを作っていきます（略）味噌汁を作れば完成

(50)（略）若しくは海苔を切っている間に、ミートボールをレンチン♪

(51) メインの肉じゃがボールを焼きながら、他の副菜をつくっていきます。

(52) 野菜はレンチン加熱、または軽くレンチンしてから調理で時短します

5. まとめと考察

ここまで各カラムを考察してきたが、いくつかのカラムには類似した内容が含まれていることが特徴として挙げられる。例えば、「レシピの概要」に(53)のようなデータがあるが、これは「レシピの生い立ち」の(54)に近い記述と考えられる。また「手順の内容」と「レシピのコツ・ポイント」にも類似した表現が見られる。このようなことが起こるのは、各カラムの解釈はユーザーに任され、実際に記入するのがユーザー自身であるためであるが、ある特定の調査の際にほしいと思う表現や内容が、一つのカラムの中に含まれているとは限らないことになる。そのためにも図１で示したような、近くに位置するカラムも考慮に入れる必要がある。

(53) 89歳の元気なババから教わった味ですよ♪(o^―^o)
(54) 主人のお義姉さんがオムレツ上手で、結婚当初に教わりました。

ユーザーから投稿されたレシピに関しては以前からクックパッドが内容を確認した上で掲載していたようであるが[10]、細かいカラムの記述についてまでは確認が届いていない可能性もある。調査をする際には、この点にも注意して使用する必要がある。

また、あるカラムでは特定の構文がよく使われるなどの特徴も明らかになった。「レシピのコツ・ポイント」「アドバイス」など、実際のアドバイスの構文を考察するのに適切であるし、「レシピ作者のコメント」などは、褒めに対する返答という今までの(会話) コーパスにはないものとして、使用することが出来るのではないかと思われる。

このように使用する際の注意点はあるが、クックパッドデータは、先行研究で述べたオノマトペや料理名に関する研究だけではなく、上述した褒めにおける言語行動のような研究での使用も可能なのではないかと考えられる。

6. 今後の課題

クックパッドデータは、レシピや献立を含む今までにないデータで、「現代日本語書き言葉コーパス(BCCWJ)」における「Yahoo!ブログ」や「Yahoo!知恵袋」と同様の特定分野のコーパスといえる。今回は、クックパッドデータと新聞のデータの対照を行ったが、Web上のテキストという共通点を持つこれらのデータとの対照も行ってみたい。

また、今後の課題として、クックパッドデータにおける表記の多様性の処理が挙げられる。これは形態素解析の結果に影響を及ぼし、例えば「あさり（アサリ浅蜊)」など平仮名表記されたものは、正しく形態素解析されず、これらの語は動詞（漁り）として処理されてしまう。そこで、今回は平仮名表記される料理名や材料名が正しく解析されるように、「前処理」の段階で「語の取捨選択」の「強制抽出する語の指定」にこれらの語を加え対応した。しかし網羅的に対応できているわけではないため、今後は形態素解析の結果を検討し、「強制抽出する語の指定」を増やし、さらなる精度の向

[10] https://nlab.itmedia.co.jp/nl/articles/1704/13/news054.html、2020年4月7日確認

上が求められる。これには、料理分野の専門用語を準備し、強制抽出する語を増やすことで対応できそうであるが、「たら（鱈）」などはそのままでは接続助詞として解析されるため、形態素解析の結果に影響してしまい、単純に増やせばよいというわけではない。表記の揺れを吸収するために形態素解析用辞書にUniDicを使用することも考える必要がある。

さらに今回は、ある不特定の場所から連続する1000のデータと新聞記事を比較したが、季節による変動要因への配慮が欠けていた。具体的には「献立のポイント」の際に見られたのであるが、夏から秋にかけてのデータが抜き出されたようで、「暑い」「運動会」等がデータを特徴づける語として抽出されている。今後はデータの抽出方法を工夫して、季節によって変動するものと変動しないものを分ける必要がある。

本稿では、紙幅の都合で各カラムの詳細な特徴については記述することが出来ず、また調査方法・集計方法には様々な問題も含まれているのであるが、クックパッドにはどのようなデータが含まれているのか、処理する際にはどのような問題が発生する可能性があるのか、この調査結果が理解の一助となれば幸いである。

参考文献

石川慎一郎(2012)『ベーシックコーパス言語学』、ひつじ書房

石田基広(2017)『Rによるテキストマイニング入門（第２版)』、森北出版株式会社

小野雄一(2019)「日本語の料理名に現れる de と with：CookPad データから見えるもの」『言語処理学会 第25回年次大会 発表論文集』、pp.819-822

小野雄一・呼思楽・森野綾香・若松弘子・砂川詩織(2017)「日本語の料理名に出現する英語前置詞の借用について:Cookpad データと実証実験から見えるもの」『言語処理学会第21回年次大会発表論文集』、pp.1020-1023

樋口耕一(2014)『社会調査のための計量テキスト分析―内容分析の継承と発展を目指して』、ナカニシヤ出版

福留奈美・伊尾木将之(2018)「レシピサービス「クックパッド」におけるオノマトペの使用－ABAB型を中心に－」『計量国語学会第62回大会予稿集』、pp.31-36

楊天羽(2020)「日本語学習者のための「食」のオノマトペリストの提案 ―大規模データベース調査から―」、首都大学東京人文科学研究科、修士論文

謝辞

本研究では、国立情報学研究所のIDRデータセット提供サービスによりクックパッド株式会社から提供を受けた「クックパッドデータセット」を利用した。

またKH Coderで形態素解析を行う際の問題点の抽出には、東京都立大学大学院生・林鈺唯さんの協力を得た。ともに記して感謝する。

（はせがわ　もりひさ・東京都立大学）