The University of Osaka
Institutional Knowledge Archive

Title	「話題別コーパス」から「話題別単語帳」を作る
Author(s)	中俣, 尚己
Citation	多文化社会と留学生交流 : 大阪大学国際教育交流センター研究論集. 2023, 27, p. 59-68
Version Type	VoR
URL	https://doi.org/10.18910/90845
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

「話題別コーパス」から「話題別単語帳」を作る

中俣　尚己*

要　旨

筆者らは、話題別言語資源から話題別特徴語の一覧表を作り、それを基に話題別単語帳を作成・出版した。本稿では、その方法論について述べる。

まず既存の教材から2,000を超える単語をリスト化した。次に、『名大会話コーパス』ならびに『日本語話題別会話コーパス：J-TOCC』から作成した特徴度をLLRで示した語彙表を用い、Excelの関数を利用して全ての語に対して特徴度が最も高い話題を自動付与した。自動付与された語の割合はN3語彙で80％程度、N1語彙で40％程度であり、さらに人手による修正も不可欠であることから本稿の方法論は「半自動付与」と言うべき水準のものである。しかしながら、教材作成時の労力削減という観点からは十分な効果を発揮したと評価できる。

また、単語を組み合わせて例文を作る際に作成者が注意しなければならない点として、「有り得るシチュエーションでなければならない」「文脈から文体を考える」「ジェンダーロールの固定化に加担しない」「当たり前のことを言われても面白くない」「データを調べればわかるようなことを空想で述べない」の5点を指摘した。

【キーワード】話題別語彙表、J-TOCC、対数尤度比、特徴話題、半自動処理、例文作り、日本語教材

1　はじめに

本稿では筆者が構築した話題別コーパスの情報を活用し、話題別単語帳を作成、出版したプロセスを解説する。本稿で解説する話題別単語帳は中俣（編）（2021）『ミニストーリーで覚える JLPT 日本語能力試験ベスト単語N3 合格2100』（以下、「N3」と略記）と話題別コーパス研究会（2022）『ミニストーリーで覚える JLPT日本語能力試験ベスト単語N2合格2400』（以下、「N2」と略記）である[1)]。

これらの単語帳は短いストーリーに単語を紐づけて覚えるタイプのものであるが、2,000を超える単語をコーパスからの情報に基づき、「半自動的」に17～23の話題に分類するという、他に類をみないプロセスを経て作成されたという特徴を持つ。また、ミニストーリー、すなわちやや長めの例文についても学習者に親しみやすいものになるように様々な工夫を行った。これらの試行錯誤は今後の教材作成にも活かせるものであるため、本稿において記録として留めたい。

以下、2で話題別単語帳のコンセプトについて述べ、3でその元となった話題別コーパスを構築した背景について述べる。4では話題別語彙表から、ある単語がよく使われる話題を半自動で付与する方法とその結果および評価について述べる。5ではやや視点を変え、単語から例文を作る時の注意点について述べる。6はまとめである。

2　単語帳の企画

この単語帳の企画は元々は中俣（2018）で提案した「コロケーション・クイズ」の書籍化であった。しかし、一度答えを知ってしまうと陳腐化するクイズの書籍化は難しく、紆余曲折を経てオーソドック

*　大阪大学国際教育交流センター准教授

スな単語帳を作ることになった。そこで、当時筆者が進めていた「話題別コーパス」の成果をフルに活かした単語帳にしたいと企画した。

現実的には、単語帳で1,000を超える単語を配列する時にランダムに配列することは有り得ず、関連のある語をトピックやテーマという形でまとめて提示することがほとんどである。しかしながら、どの単語をどのトピックに割り当てるかというのは制作者の勘によるところが大きい。単語帳ではないが、旧日本語能力試験出題基準に登場する8,110語を100の話題に分類した山内（編）（2013）も、基本的には主観によって作業が行われている（p.8）。

また、コーパス研究の成果を活かした単語帳としては石澤ほか（2018, 2021）があるが、これは収録する単語の選定にコーパスから作られた「日本語学術共通語彙」（松下2011）を利用しており、必ずしも単語帳内のトピックについてはコーパスデータに基づいているわけではない。

これは「語」と「話題」を紐づけるためのコーパスデータが存在していなかったためである。しかし、筆者のプロジェクトの成果物を使えば、例えば副詞の「全部」のように一見特定の話題とのつながりが見えなさそうな語であっても、実際には「家事」の話題に有意に多く、次いで「スマートフォン」と「マンガ・ゲーム」の話題に多いということを客観的に示すことができる。それはよりその単語が使われることが多い自然な例を示すことにつながる。

このことから、話題別コーパスから得られる情報をフル活用した真正な「話題別単語帳」を作成することを決めた。

3　話題別コーパスについて

3-1　話題別コーパスの構築について

ここでは、筆者がなぜ話題別コーパスを構築するに至ったか説明する。

近年の日本語教育では文法積み上げシラバスからの脱却が進んでおり、CBI（Content Based Instruction: 内容中心の教授法）やCLIL（Content-Language Integrated Learning: 内容と言語の統合的学習）、TBLT（Task-Based Language Teaching）といった、石川（2017）が内容志向型教授法と呼ぶ教育法が主流となってきている（佐藤ほか（編）2015, 奥野（編）2018; 2021, 小口2018）。

また、初級からの日本語教科書も、大阪大学国際教育交流センターで用いられている西口（2012）『NEJ』と西口（2018）『NIJ』をはじめ、国際交流基金（編）（2013）『まるごと　日本のことばと文化』など、各課でテーマやトピックを定めて関連した活動を行うものが増えてきている。

さらに、スリーエーネットワーク（2012）の『みんなのにほんご』シリーズのような従来型の文型積み上げシラバスを採用するとしても、そこには必ず会話活動があり、会話活動は必ず何らかの話題を伴う。

これらいずれのアプローチをとるにしても、「話題」と「言語形式」の関係が分かっていれば、ある話題で注目すべき言語形式を抽出できるし、ある言語形式の導入にふさわしい話題を特定できることになる。

さて、ここで話題についての先行研究に目を向けると、談話分析や会話分析の分野では「話題転換」（花村2015）あるいは初対面場面での「話題選択」（三牧1999）などはよく研究の対象として取り上げられてきた。一方で、話題と言語形式の関係に着目した研究はほとんど見られない。

その少ない例外が山内（編）（2013）である。これは「食」「酒」「衣」「旅行」など具体的な100の話題を設定し、構文ごとによく用いられる名詞などを分類したものである。ただし、収録後の大部分は旧日本語能力試験出題基準に記載された語彙であり、分類は主観による（山内（編）2013: 8）。また、機能語は話題に従属しない語としている。

これに対して、Nakamata（2019）は小規模な接触場面会話コーパスである『日中Skype会話コーパス』を用いた調査を行い、例えばテンス・アスペクト形式の「た」や「てある」、時間副詞の「昨日」や「来年」は「ポップ・カルチャー」に多く使用されているということを指摘している。機能語が話題の影響を強く受けるという事実は、日本語教材開発において大きな意味を持つ。しかしながら、この研究で利用されたデータは話題の統制がされていたわけではない。また、接触場面のデータであることから、母語話者の会話でも話題によって同様の偏りが見られるかは未検証であった。

はたして、機能語＝文法項目は話題によって偏りが見られるのか。もしその偏りが明らかになれば、先述の通り、話題シラバス・文法積み上げシラバスを問わず教材作成にとって有益な情報となる。この

ことを検証するため、筆者はJSPS科研費基盤研究（B）18H00676「話題が語彙・文法・談話ストラテジーに与える影響の解明」というプロジェクトを発足させ、話題ならびに時間や性別などを統制した新規会話コーパス『日本語話題別会話コーパス: J-TOCC』（以下、J-TOCC）の構築と、すでに存在していた自然会話コーパス『名大会話コーパス』（藤村ほか2011; 以下、NUCC）の話題による分割という２つの言語資源の整備を行った。この２つの言語資源から明らかになった話題と語の関係をさらに教材に反映させたのが「N3」と「N2」の２つの単語帳である。

3-2　利用した語彙表について

ここでは、科研プロジェクトで制作した２種の言語資源とそれを元にした語彙表について説明する。

まず、最初に完成したのはNUCCを話題ごとに分割したデータである。これは、NUCCの全文を各ファイル３名の作業者が目視し、手動で話題に分割作業を行ったものである。話題は先述の山内（編）（2013）を参考に104種類用意した。実際にはNUCCでは話されていない話題（「算数・数学」など）もあり、97の話題が出現した。つまり、NUCCを97のサブコーパスに分割することに成功した[2)]。

続いて、その１つ１つのサブコーパスを「comainu」（小澤ほか2014）というソフトを用いて、国立国語研究所作成の『現代日本語書き言葉均衡コーパス』などで採用されている「長単位」と呼ばれる単位に形態素解析した。その後、１つ１つのコーパスを該当コーパス、他の96コーパスを結合したものを参照コーパスと見立て、NUCCに出現したすべての語に対して特徴度の指標である対数尤度比（LLR）を計算した。最後に、NUCC全体での総出現頻度が10以上の語に絞り込んだ表が『日本語話題別語彙表』（以下、「NUCC語彙表」と略称）であり、筆者のウェブサイトで無料で公開している。列方向に97の話題、行方向に3,324の語が並び、各セルには各語の各話題における特徴度をLLRで表示したものである。

「NUCC語彙表」は話題の数が97と多く、非常に広い範囲の話題を網羅していることが特徴である。また、特に話す内容について指示を与えていない自然会話コーパスを元にしているため、真正性の高いデータと言える。なお、この語彙表の特性についての詳細な情報は中俣ほか（2021b）を参照されたい。

もう一つの言語資源は、本科研で新たに構築したコーパスである『日本語話題別会話コーパス：J-TOCC』である[3)]。こちらも詳細は中俣ほか（2021a）に譲るが、親しい大学生のペア120組に15種類の話題をボードで提示し、それぞれ５分ずつ話してもらったものを文字化したコーパスである。総時間数は150時間、語数はおよそ160万語と話し言葉のコーパスとしては大規模なものである。また、録音地（東日本と西日本）、性別の組み合わせ（男男と男女と女女）でデータ数のバランスを取っているのも特徴である。コーパスは全部で1,800のテキストファイルと、フェイスシートのデータから構成され、筆者のウェブサイトで公開している。

このJ-TOCCからも話題別の語彙表を作成した（以下、「J-TOCC語彙表」と呼ぶ）。こちらもNUCC語彙表の時と同様に、comainuで長単位に分割した。そして、話題ごとのLLRを計算した。ただし、今回は総語数による絞り込みを行っていない。そのため、長単位で実に43,000種類以上の語に対して話題情報を提供できる。

なお、単語帳作成時に利用した表を後にバージョンアップし、『日本語話題別会話コーパス：J-TOCC語彙表』としてこれも筆者のウェブサイトで公開している。単語帳作成時の表とは若干のズレがあるが、ほぼ同じものとみなして差し支えない。この語彙表についての詳細は中俣・麻（2022）を参照してほしい。

J-TOCC語彙表の特徴としては、話題が15種類と少なく、うち11種類は「食べること」「スポーツ」など初級でも扱える身の回りの話題が多いことである。そして、難易度の高い話題についてはほとんど扱えていない。一方、１つの話題についての語数は非常に豊富である。そのため、例えば「食」の話題を例にとると0.1%水準で特徴語と認定できるLLRが10.83以上の語は、NUCC語彙表では236種類抽出されるのに対し、J-TOCC語彙表では753種類抽出される。

２つの語彙表の性格をまとめると、表１のようになる。

4　語の話題認定のプロセス

4-1　話題の半自動的付与の方法

ここでは、話題別単語帳から、どのように収録語リストにある話題によく使われる話題を半自動的に付与したかについて述べる。

まず、収録後リストは今回の言語資源とは独立に

表１　２つの話題別単語帳の比較

論文中での名称	NUCC 語彙表	J-TOCC 語彙表
正式名称	『話題別日本語語彙表』	『日本語話題別コーパス：J-TOCC 語彙表』
元となったコーパス	『名大会話コーパス』	『日本語話題別コーパス：J-TOCC』
話題数	97	15
語数（異なり）	3,324	43,110
元データの1話題あたりの語数	話題によって異なる。数百語～十万語	どの話題も十万語以上
会話への話題の指示	なし	あり

用意した。これは、既存の JLPT 対策として作られた単語帳の語から集めたもので、単語・読み・JLPT のレベルが付与されたリストであり、N5 から N1 まで7,389語あった。ただし、「女房」のように2020年代ではふさわしくない語については削除を行っているし、反対に現代的な文脈に合わせるため、例文作成時には積極的に外来語・新語を取り入れるようにもした。ともあれ、まずはこの語リストに半自動的に話題を割り振ることを試みた。

まず、リスト上の語を形態素解析機「web 茶まめ」にかけ、語彙素の情報を習得した[4)]。web 茶まめでは「UniDic」というコンピュータ用の辞書を用いて単語を分割・認定する。語彙素はその辞書の見出しの形であり、例えば「食べる」「たべた」「タベ」などの様々な語形に対し、「食べる」という１つの「語彙素」が表記・活用の代表として設定されている。先述の NUCC 語彙表、J-TOCC 語彙表もベースは UniDic を用いて分析されており、また各語はこの語彙素を見出しにしているため、検索キーとして利用する。「web 茶まめ」では「作成する」などは「作成」と「する」の二語に分割されるため、適宜補正を行った。

なお、この時に品詞と活用型（五段活用・下一段活用など）の情報も取得、付与した。UniDic の用語（形状詞・五段活用）と日本語教育で一般的な用語（ナ形容詞・グループ1）は異なるが、ある用語を別の用語に置換するほうが、ゼロから入力を行うよりも手間はかからない。実際には品詞情報や活用情報については学生アルバイトを雇い、最終的に手作業で確認・修正を行っている。

検索キーとしての語彙素が入力できたら、次は同じファイルに別のシートを追加し、NUCC 語彙表を読み込む。そして、各語において一番 LLR の高い話題を「特徴話題」として自動で表示させる。具体的には LLR の最大値を MAX 関数で取り出し、MATCH 関数でその値を持つ列が何列目かを割り出し、INDEX 関数で話題が書かれている１行目から該当の列の話題を呼び出す[5)]。

ここまでできれば、元のシートに戻り、先程設定した語彙素をキーとして、VLOOKUP 関数で特徴話題を呼び出せばよい。これで、各語が最もよく使われている話題は何かを半自動的に提案することができる。しかし、実際には NUCC 語彙表では語数が少なすぎ、十分な提案とはならなかった（表２参照）。そこで、15話題と少ないながらも、J-TOCC 語彙表についても同様の手続きを行い、特徴話題の提案を行わせた。この提案された話題を元に、最終的には人手で話題の決定を行った。

4-2　半自動話題付与の結果

ここでは、リストの語に特徴話題を半自動的に付与したことの結果について述べる。まず、NUCC 語彙表を用いて話題付与を行った結果は表２の通りであった。

表２　NUCC 語彙表を用いた際の話題付与率（%）

	N1	N2	N3	N4	N5
付与率合計	2.1	12.9	31.4	50.5	21.8

N3 でも３割程度、N1 にいたっては２%と、ほぼ情報が得られない結果であった。次に、J-TOCC 語彙表を使って話題を認定した結果を表３に示す。

J-TOCC を用いた場合、自動で話題を付与できた割合は N3 では８割を超えた。これは大幅な自動化ということができる。N3 の作業対象語数は1,740語であったが、ヒントなしで話題を考えなければならない語はその５分の１以下、326語のみで済んだからである。N2 でもカバー率は７割近く、作業量の低減には十分に貢献している。N1 ではカバー率は４割強と激減している。これは、身近な話題が多いという J-TOCC の特性が、専門的な語も多く登場する N1 語彙とマッチしなかったためであろう。それでも、４割程度の語について話題を自動で付与できるのは、ゼロからすべてを手入力するよりは心強い[6)]。

また、表３の数字を見ればわかる通り、どの話題

表３　Ｊ－TOCC 語彙表を用いた際の話題付与率（%）

	N1	N2	N3	N4	N5
01.食べること	1.8	2.0	4.8	4.4	4.4
02.ファッション	1.4	3.0	4.7	3.8	4.4
03.旅行	2.1	3.0	4.5	5.3	3.3
04.スポーツ	2.1	3.9	4.9	4.1	3.1
05.マンガ・ゲーム	1.8	3.2	4.3	2.3	3.6
06.家事	2.1	3.0	6.9	5.2	4.6
07.学校	3.4	4.4	6.4	4.5	5.6
08.スマートフォン	3.2	4.5	6.3	5.1	6.0
09.アルバイト	2.5	5.4	5.2	4.8	4.4
10.動物	2.6	4.0	4.2	3.5	4.4
11.天気	2.5	5.9	5.3	6.6	6.7
12.夢・将来設計	4.6	6.9	5.6	4.1	3.4
13.マナー	2.8	4.9	5.1	5.8	5.1
14.住環境	3.1	4.9	5.2	4.4	5.7
15.日本の未来	6.2	8.2	7.9	6.2	4.4
付与率合計	42.2	68.1	81.2	69.9	63.7

も満遍なく付与されたというのは、嬉しい誤算であった。話題によって含まれる語の数に多寡が生じることは覚悟していたが、意外にもどの話題にも一定数の語が含まれていた。つまり、章による語数のバラツキが大きくないということで、これは単語帳という性質からは望ましいことである。「15.日本の未来」の割合が若干高くなっているが、これは「割合」（N3）や「対象」（N2）など、学術的・抽象的な語の多くがこの話題に出現したためである。そもそもこの15種類はJ-TOCCのために設定されたものであり、単語帳の構成をコーパスに合わせる必然性はない。実際には語数の多い話題は２つの話題に分けるなどして調整し、表４のようになった。なお、この話題の再割り振りは完全に筆者の主観で行っている。

ちなみに、表４のN3の話題の配列順にもコーパス研究の成果を応用している。NUCCを話題ごとに分割した際に、どの話題からどの話題に移ることが多いのかということをネットワーク分析によって可視化し、できるだけ遷移が多い話題どうしが連続するように配列を決定した。

表４　実際に採用された話題

N3	N2
1.食事、2.家事、3.買い物、4.ファッション、5.テクノロジー、6.流行、7.人づきあい、8.スポーツ、9.動物、10.町、11.天気、12.旅行、13.学校、14.仕事、15.人生、16.健康、17.マナー、18.社会	1.食事、2.家事、3.買い物、4.ファッション、5.テクノロジー、6.流行、7.趣味、8.人づきあい、9.年中行事、10.スポーツ、11.動物、12.住、13.町、14.天気、15.旅行、16.学校、17仕事、18.人生、19.健康、20.マナー、21.社会、22,政治、23.環境・科学

4-3　半自動話題付与の評価と人手による調整

今回の手法は、半自動付与と述べているように、コーパスから計算された特徴話題をそのまま採用するのではなく、NUCCの情報とJ-TOCCの情報を筆者が確認し、最終的な話題を１つ１つ決定していった。これは、近年の自然言語処理の水準からすると稚拙な手法とも言えるが、形態素解析には誤解析がつきものであり、多義語の問題も存在する。売り物にする以上は、人の目によるチェックは避けられないと考えた。また、本手法はLLRの計算のみ外部の専門家に依頼したが、それ以外はエクセルを用いた非常に簡単な方法である。高度なプログラムに頼らずとも、話題決定の労力を1/5にすることができたことに、この取り組みの価値があると考える。

以下、半自動付与の結果を受けて、具体的にどのように筆者が話題認定を行ったかについて２，３例を示す。

まず、半自動で付与されることで助かったのは、副詞などの一見話題に依存しているとは思えない語である。例えば、「まさか」（N2）は一見、どのような話題でも使えそうに感じる。しかし、NUCCでは「結婚」に、J-TOCCでは「マンガ・ゲーム」に多いという結果が出た。このように話題をキーワードとして提示されると、例文も作りやすい。今回は、「あの２人がまさか結婚するなんて…」という例文は作りやすいがありきたりで新鮮さがないと判断し、あえて「流行」の話題に割り振った。最終的にできあがった例文は（1）である。下線を付した語は他のキーワードである。

（1）　A「ライバルってすごく重要な要素だと思う」
　　B「わかる。名作ってだいたいライバルが主人公より人気があるよね。最初強かった敵が、味方になる。まさかあのキャラクターが仲間になってくれるなんて！ってワクワクする。」
　　A「でも、主人公にふさわしい敵として最後まで戦うのもいいよ。」

多くの場合は、付与された話題に従ったが、一方

で、付与された話題を無視したこともある。例えば「学期」（N3）はJ-TOCCでは「スマートフォン」に多いという結果になったが、これは「学校」の中に収めるべきであると判断した。「食卓」（N2）も「動物」の話題に多いという結果になり、そのような例文も不可能ではないが、よりストレートに「食事」の話題に設定した。このように、よりふさわしい話題を即座に思いついた場合にはそちらに設定しなおした。一方で、「人工的」（N2）が「住環境」など直感的には結びつかなくても、よりふさわしい話題が即座に思いつかない場合には、コーパス研究の成果を活かすという理念から、付与された話題を採用した。

最後に、コーパスから話題付与を自動で行えなかった語であるが、これらの語は「バーゲン」（N3）といえば「買い物」、「心身」（N2）といえば「健康」のようにすぐに話題が想起されるものが非常に多かった。コーパスから話題付与ができなったということは、コーパスに出現しなかった語であり、それは使用頻度の低い語であると結論することができる。そのような語はえてして使用範囲が狭い一種の専門語であって、かえって話題を決定しやすいのである。むしろコーパス情報を利用したことの最大のメリットは、複数の話題で使われ、直感では一つの話題に絞りきれない使用頻度が高い語に対して、客観的なデータを元に特徴話題を付与し、教材作成者がどの話題がふさわしいか迷う時間を削減できたことにあると言えよう。

5　例文作成のプロセス

5-1　例文の仕様と制作体制

本シリーズの単語帳はタイトルにミニストーリーとあるように、話題にそった少し長めの例文（1文とは限らない。2名による短い会話であることもある）を示し、そこに含まれる単語を文脈とともに覚える仕掛けになっている。

このやや長めの例文については、まず前節で解説した語の話題付与の結果をもとに、話題ごとに単語リストを再作成した。次に例文作成者がそのリストを見ながら、キーワード（覚えるべき語）を3～8個含んだ例文を考えていき、できる限りリストにある語を使いきるという方法で行っていった。一文に含まれるキーワードが3～8個というのは紙面レイアウトの都合以上の意味はない。

また、そもそものリストが網羅的であるというわけではないため、リストにない語であっても、自然な例文作りのために必要な語であればキーワードを追加することを奨励した。特に例文作成者の1人の小西円氏から「既存の教材ではカタカナ語が手薄になっている」という指摘を受け、これを積極的に取り入れるように指示をした。

さらに、話題の最終決定も結局のところ筆者1名で行っているため、例文を作ってみてどうしても使えない語、あるいは自分の例文で使いたい語が他の話題にある場合は他の話題に「トレード」することも許可した。この作業のため、例文作成者には全単語の話題付与結果を渡した。

ここまで説明した例文作成のプロセスについては「N3」と「N2」でほぼ変更はないが、実際の作業体制は「N3」と「N2」で大きな違いがある。

「N3」制作時は科研分担者から筆者を含めて5名の体制で例文作成に取り組んだ。この時、1人が3つから4つの話題を担当し、1つの話題には多ければ200近い語が含まれていたため、作業期間が短いことも相まって負担が非常に大きかった。

そこで、「N2」制作時は科研のメンバーにこだわらず、科研メンバーの周囲の若手研究者・院生にも声をかけ、13名の体制で臨んだ。これにより、1人が担当する話題は1つから3つになり、負担は大幅に軽減された。また、「N3」の時には最終的に作られた例文は「実際に使われるシチュエーションが想定できない。」「科学的に正しくない。」「ジェンダー的なバイアスが見られる。」「実際に使われるような活き活きとした言葉になっていない。」などの問題が見られるものもあった。そこで、N2については、最初に全執筆者を集めてオンラインミーティングを行い、「例文を作成するときに留意すべきこと」について30分ほどの時間をかけて説明した。また、例文完成後、全ての例文に筆者が目を通し、「完全監修」という形にした。

5-2　例文を作成するときに留意すべきこと

以下、筆者が実際に執筆者に伝えた「例文を作成するときに留意すべきこと」について説明する。執筆者には5-2-1から5-2-5の節タイトルに該当する内容を伝えた。この5つの方針を総括すると、「リアリティのある文を作って欲しい」ということになる。以下、主として「N3」作成時に直面した問題に

触れながら、各項目について詳しく解説する。

5-2-1　ありえるシチュエーションでなければならない

これは言い換えると、自分が聞いたことがある、あるいは使ってもおかしくない例文を作ってほしいということである。ありえるシチュエーションでなければならないというのは当然のことのようであるが、話題ごとの単語リストを眺めながら例文を作っていると、ありえない例文を作るというミスを犯しがちになる。例えば、(2) のような例文はよく考えるとなぜ「性別」を書く必要があるのがわからない。完成版ではこの部分はカットされた。

(2)　田中先生の研究室をノックしたが、留守のようだ。今日は大学をお休みになっているのかもしれない。伝言メモに、「お目にかかって、先生の本を拝見したいと思っています。明日また尋ねます」と書いた。氏名と性別も書くのも忘れなかった。

また、例文を作成したが事実と異なることがわかったため、ボツになったものもある。

(3)　和紙は日本の伝統的な紙である。草を原料にして作られていた。日本のあちこちに、昔の紙作りを体験できる場所がある。ぜひとも一度体験してほしい。

これは筆者が和紙の原料であるコウゾを「草」だと思い込んでいたものであるが、現実にはコウゾは低木の皮である。

5-2-2　文脈から文体を考える

まず、例文を会話形式にするかモノローグないし論説文形式にするかという問題がある。これは単語の硬さにとって相応しい文体が異なると結論付けた。会話で使いやすい語もあれば、硬い書き言葉で使われる語もあるのは当然である。

その上で、文末は普通体か丁寧体かという選択がある。これについて普通体を基本とし、何か理由がある時には丁寧体を用いるという方針をとった。例えば、(4) のような文は、無論初級レベルの教科書では丁寧体を使うことが多いだろうが、現実を考えた場合、なぜ丁寧体を使っているのかがはっきりしない。そもそも誰が誰に対して話しているのかがわからない。そのため、これは普通体に改めた。対して、(5) は店における店員と客の会話という設定があっきりしており、丁寧体を使う必然性があるため、このような例文はそのままにした。

(4)　私の趣味は洗濯です。洗濯物を干すとき、濃い色の服は裏返して干すと色が落ちません。乾いた洗濯物は、日が暮れるまでに取り込みましょう。太陽のにおいのする洗濯物を畳むのは幸せです。

(5)　客：このシャツって、男性用ですか？
　　店員：はい、もともとは男性用なんですが、最近大きめのサイズのシャツを着るのが流行っているので、女性のお客様でもおしゃれに着ていただけますよ。
　　客：うーん、私が着るとパジャマっぽくなりそうで……。

5-2-3　ジェンダーロールの固定化に加担しない

これは例文の中で、例えば家事や仕事などを特定の性と結びつけ、固定化するようなことは避けるべきだということである。学校教育において、本来意図していないメッセージが教育を通して伝わることを「隠れたカリキュラム」「潜在的カリキュラム」と呼ぶ。これには様々なものがあるが、教材に書かれた内容もその一例であり、永田（2012）は小学校国語科教科書を分析し、登場人物が男性に偏っていることを指摘している。単語帳も一種の教材であり、しかも何度も学習者の目に触れるものであるからこそ、隠れたカリキュラムを含むことがあってはならない。必要がなければ「男」「女」「父親」「母親」といった情報を不用意に例文に盛り込まないように求めた。

また、学習者には男性も女性もいるのだから、特に理由がなければことさらに男性性・女性性を強調するような言葉遣いは不要である。例えば、(6) には「ものよ」「わよ」が使われているが、この例文で女性性を強調する意義は見いだせない[7)]。

(6)　A　この袋に入っている緑色の粉、何？
　　B　ケールっていう野菜を粉にしたものよ
　　A　ああ、健康にいいやつだね。
　　B　そうそう。お湯に溶かして飲むの。粉を固めて作った錠剤もあるわよ。
　　A　ちょっと飲んでみたいな
　　B　今ちょうどわかしたお湯がやかんに入っているから、作ってあげるよ。

対して、(7) のような「の」「わ」は若者であれば男性・女性問わずに使用する可能性があるので、採

用した。

(7) A 昨日スマホ落としちゃって、画面割れたの。修理も無理だって。めっちゃショックなんだけど。

B え、カバーつけてなかったの？

A 今度からつけるようにするわ。

5-2-4 当たり前のことを言われても面白くない

これはN2にふさわしい知的好奇心を満たすような事実を取材を基に示すという方針である。例えば「小麦」というキーワードに対して、「小麦は世界中で食べられている」という文は平凡で何も情報を持たない。しかし、(8) のようにすれば世界第3位の作物であることを知らない学習者もいるだろうし、また、他の語とセットで覚えることができる。

(8) 小麦は、トウモロコシ、米の次に世界中で作られている。小麦は、まず粉にする。それから、パンを作ったり、麺を作ったりする。イタリアのパスタは世界中で食べられている。

また、学習者の興味・関心に合わせたのは学術的な分野に限った話ではない。「流行」という話題は主に「マンガ・アニメ・ドラマ・ゲーム」を含むものであるが、筆者が執筆を担当した。この話題では具体的な作品名は例文に含まれていないが、その作品を知っている者が読めばすぐにどの作品かがわかるような例文にすることで、学習者が身近に感じられるような工夫をしている。例えば、N2の (9) の例文は、人気アニメの『名探偵コナン』を見たことがある者ならば誰でもすぐに「怪盗キッド」という登場人物を連想するような内容になっている。その中で「考えてみれば不思議なことだが」というフレーズも導入しているが、これも同作のファンなら共感してもらえるであろう。

(9) 娯楽作品では、探偵は人気のある職業だが、泥棒も人気があったりする。泥棒のキャラクターはダイヤなどの宝を盗むが、お金には興味がなく、盗んだものを返却することも多い。また、考えてみれば不思議なことだが、しばしば銃からトランプを発射して戦ったりする。

5-2-5 データを調べればわかるようなことを空想で述べない

これは「出生数が50年前と比べておよそ2分の1になった」のようなもっともらしい例文を、一次資料に当たらずに作ることのないように戒めた。出生数については、確かに2021年は1950年と比較して40%に落ち込んでいる（厚生労働省2022）。しかし、実は2017年と比較しても15%も減少しており、出生数の減少は急激に加速している。こういったことまで考慮して例文を作るべきであると執筆者に伝えた。その結果、事実に基づくリアリティのある例文を作ることができた。(10) はN2の「仕事」の例である。

(10) ウイルス対策に協力した飲食店には、協力金が平等に支給されたが、大きな飲食店からは不平不満も出ている。大きな飲食店は家賃や給料を払うだけで足が出るからだ。

6 おわりに

本稿では、Excel形式の話題別語彙表を用いて、単語帳に収録すべき語に半自動で特徴話題を付与する方法とその結果について述べた。原始的な方法ながらも、多くの語について半自動で話題を提案することができ、教材作成時の大幅な労力削減になった。実際、機械的なサポートがなければ、2,000以上の語について特徴的な話題を特定するのは困難であったろう。人間が話題認定をする際に難しいのは、使用頻度が低い語ではなく、むしろ高い語であり、高頻度語の分析は機械が得意とするところである。

全てを完全に自動化するには高度な技術が必要になるが、人間の判断をサポートする程度のことであれば、Excelでも十分な効果を発揮できたと言える。「完全自動化」にこだわらずとも、「半自動化」でも教材作成に有益であるという知見が体験的に得られたことも収益であった。

また、後半では単語から例文を作るにあたって、「有り得るシチュエーションでなければならない」「文脈から文体を考える」「ジェンダーロールの固定化に加担しない」「当たり前のことを言われても面白くない」「データを調べればわかるようなことを空想で述べない」という5つのポイントに注意すべきであるということを主張した[8)]。

なお、本研究ではExcel形式の2種の話題別語彙表を基に教材作成を行ったが、この表はExcelの操作に習熟していれば、様々な情報を取り出し有益に活用できるものの、そうでない者にとってはただの数字の羅列に等しく、日々の授業の準備にさっと利用できるものとは言い難い。そこで、2022年度から

新たな科研費プロジェクトとして、JSPS 科研費基盤研究（B）22H00668「「話題から始まる日本語教育」を支援する情報サイトの構築と話題別会話コーパスの拡充」を開始し、話題から特徴語を提案したり、語から特徴話題を提案するようなサイトを構築することを目指している。語がどのような話題で使われるかという情報は、日本語教育で行われる言語活動をより真性にかつより効果的なものにすることに貢献するであろう。

注

1） 制作上の都合で編著者のクレジットがN3とN2で異なるが、どちらも筆者の指揮・監修のもと、同様のプロセスで制作された。

2） 分割したデータは公開していないが、名大会話コーパスのどの行がどの話題に対応しているかというアノテーションデータは言語資源協会（GSK）のウェブサイトで公開しており、プログラムを使って分割データを再現できる。

3） J-TOCC は Japanese Topic-Oriented Conversation Corpus の略であり、「ジェイトック」と読む。

4） https://chamame.ninjal.ac.jp/

5） 具体的には、D1:R1 に見出しとして話題が書かれており、D2:R2 にその LLR が書かれている場合、入力式は＝INDEX（$D$1:$R$1,1,MATCH（MAX（D2:R2),D2:R2,0））となる。

6） 後述するように、実際には自動で話題が付与できなかった単語は低頻度語で専門性が高く、人間が話題を思いつきやすいという特性がある。N1 では更なる試みとして大学院生 RA に話題入力をお願いし、「ぱっと思いつかなければ空白にしておくように」と指示したが、最終的には 8 割を超える単語に付与することができた。十分な成果と言える。

7） J-TOCC 本文では「ものよ」は 1 例も見られず、全て「もんよ」という形になる。また、「わよ」は親世代の言葉を引用する形でのみ使われ、大学生が自分の言葉として文末に使用した例は皆無であった。このことから「わよ」は特定の世代のマーカーとして解釈することも考えられる。なお、終助詞の「わ」「の」「よ」は性別・地域を問わず広く用いられている。

8） なお、毎日の授業準備などの中では、本稿で挙げたような細やかな配慮は、もちろんできるに越したことはないが、必ずしも常に意識すべきだとは筆者は考えていない。しかし、単語帳としてその例文が活字化され、多くの学習者の目に繰り返し触れるということを考慮した場合には、留意する必要がある。

付記

本研究の遂行には JSPS 科研費 18H00676 ならびに 22H00668 の助成を受けた。

参考文献

石川慎一郎（2018）『ベーシック応用言語学』，ひつじ書房.

石澤徹・岩下真澄・伊志嶺安博・桜木ともみ・松下達彦（2018）『語彙ドン！［vol.1］』，くろしお出版.

石澤徹・岩下真澄・桜木ともみ・松下達彦（2021）『語彙ドン！［vol.2］』，くろしお出版.

奥野由紀子（編）（2018）『日本語教師のための CLIL（内容言語統合型学習）入門』，凡人社.

奥野由紀子（編）（2021）『日本語で PEACE［POVERTY 中上級］』，凡人社.

小澤俊介・内元清貴・伝康晴（2014）「BCCWJ に基づく中・長単位解析ツール Comainu」『言語処理学会第 20 回年次大会論文集』，pp.582-585.

国際交流基金（編）（2013）『まるごと　日本のことばと文化　入門　A1　りかい』，三修社.

厚生労働省（2022）「令和 3 年（2021）人口動態統計（確定数）の概況」https://www.mhlw.go.jp/toukei/saikin/hw/jinkou/kakutei21/dl/15_all.pdf（2022 年 12 月 23 日閲覧）

小口悠紀子（2018）「スタンダードを利用したタスク・ベースの言語指導（TBLT）」岩田一成（編）『語から始まる教材作り』，pp.17-30，くろしお出版.

佐藤慎司・高見智子・神吉宇一・熊谷由理（編）（2015）『未来を創ることばの教育をめざして ―内容重視の批判的言語教育（Critical Content Based Instruction）の実践』，ココ出版.

スリーエーネットワーク（2012）『みんなの日本語　初級Ⅰ　第 2 版　本冊』，スリーエーネットワーク.

永田麻詠（2012）「小学校国語教科書に見る隠れたカリキュラムの考察：ジェンダーおよびクィアの観点から」『国語教育思想研究』，4 号，pp.37-46,

中俣尚己（2018）「コロケーションリストの教材化」岩田一成（編）『語から始まる教材作り』，pp.153-166，くろしお出版.

中俣尚己（編）（2021）『ミニストーリーで覚える JLPT 日本語能力試験ベスト単語 N3 合格2100』，ジャパンタイムズ出版.

中俣尚己・太田陽子・加藤恵梨・澤田浩子・清水由貴子・森篤嗣（2021a）「『日本語話題別会話コーパス：J-TOCC』」『計量国語学』，33 巻 1 号，pp.205-213.

中俣尚己・小口悠紀子・小西円・建石始・堀内仁（2021b）「自然会話コーパスを基にした『話題別日本語語彙表』」『計量国語学』，33 巻 3 号，pp.194-204.

中俣尚己・麻子軒（2022）「『日本語話題別会話コーパス：J-TOCC 語彙表』の公開と日本語教育むけ情報サイトにむけた指標の検討」『言語資源ワークショップ 2022 発表論文集』

西口光一（2012）『NEJ：A New Approach to Elementary Japanese』，Vol.1，くろしお出版.
西口光一（2018）『NIJ: A New Approach to Intermediate Japanese』，くろしお出版.
花村博司（2015）「日本語の会話における話題転換研究の概観──日本語教育に生かすための研究をめざして──」『言語文化学研究　言語情報編』, 10巻, pp.65-102，大阪府立大学人間社会学部言語文化学科.
藤村逸子・大曽美恵子・大島ディヴィッド義和（2011）「会話コーパスの構築によるコミュニケーション研究」藤村逸子・滝沢直宏（編）『言語研究の技法：データの収集と分析』，pp.43-72，ひつじ書房.
松下達彦（2011）「日本語の学術共通語彙（アカデミック・ワード）の抽出と妥当性の検証」『2011年度日本語教育学会春季大会予稿集』，pp.244-249.
三牧陽子（1999）「初対面会話における話題選択スキーマとストラテジー──大学生会話の分析──」『日本語教育』, 103.
号 , pp.49-58.
山内博之（編）（2013）『実践日本語教育スタンダード』，ひつじ書房.
話題別コーパス研究会（2022）『ミニストーリーで覚える JLPT 日本語能力試験ベスト単語 N2 合格2400』，ジャパンタイムズ出版.
Nakamata, Naoki,（2019）Vocabulary Depends on Topic, and So Does Grammar, *Journal of Japanese Linguistics*, 35（2），pp.213-234.