小説やアニメを日本語で理解するために、どの程度の語彙力が必要になるのか
How Large a Vocabulary is Needed for Understanding *anime* and Novels in Japanese? -A Pilot Study

畠山衛, ビクトリア大学
Mamoru Hatakeyama, University of Victoria

1. はじめに

読解の研究において、文章を概ね理解するためには文章内の語彙の 96%（小森ほか、2004）から 98%（Schmitt, Jiang, & Grabe, 2011) 以上が既知語彙でなければいけないとされている。Nation（2006）によると、一般向け英語小説の 98%以上の語彙をカバーするためには、9 千語以上が必要であり、英語の子供向け映画の聴解には、7 千語程度の語彙力が必要であるとされた。では、日本語の小説、アニメを理解するためには、どの程度の語彙サイズが必要になるのであろうか。これに答えるため、予備的調査として、小説「容疑者 X の献身」の使用語彙を、KH Coder（樋口、2017）により分析したところ、延べ 97,094 語が使用されていた。さらに、それらの語彙を語彙データベース（松下、2011）上で、頻度別に千語ごとに分類し、使用されている語彙がどの程度カバーされるかを調べた。最も頻度の高い千語によって 84.87%がカバーされ、さらに千語、つまり合計二千語を知っているとカバー率は 90.02%に上がった。しかし、既知語彙が 98%以上になるための語彙サイズを推定すると、11,000 語が必要だった。また、アニメにおいても、「となりのトトロ」（延べ 3,730 語）では、もっとも頻度の高い千語で 90.72%がカバーされるが、98%を越えるためには既知語彙 10,000 語が必要であることがわかった。「千と千尋の神隠し」においては、延べ 7,413 語が使用されており、語彙カバー率が 98%に達するためには、既知語彙として 1 万 2 千語が必要だった。これらのことから、日本語においては、小説、アニメとも英語よりもより多くの語彙数が求められることが示唆された。

2. 先行研究、研究課題

語彙習得のためには、何度も対象語彙に遭遇しないと覚えない（8-15 回以上）（Horst et al, 1998; Nation & Anthony, 2013;Waring & Takaki, 2003)と言われている。「する」「言う」「何」などの最も頻度の高い最初の 1000 語に含まれるような初級語彙は、様々なジャンル・場面において高頻度で使われているため、どんなものを読んだり聞いたりしても簡単に複数回遭遇することができる。ところが、中級以降の語彙では、頻度が低くなるにつれて、より大量に読まなければいけなくなる。例えば、松下（2017）によると、日本語能力試験 N1 レベルの語彙の出題範囲とされる最頻語 15,000 語レベルの場合、100 万語読んでも 3 回出現するかどうかで、本を 3、4 冊読んで 1 回出てくる程度である（p.6）。それでも頻度の低い語彙は、一般に音声言語よりも文章中に出やすいため、本を大量に読むこと

が求められる。また、実際に小説を読むことで付随的語彙習得が進むことが確認されている（Pellicer-Sánchez, & Schmitt, 2010）。

一方で、語彙の遭遇率を上げるためだけに、ただ語彙の難易度の高い文章を読んでも文章理解にはつながらないであろう。これまでの第二言語学習者を対象とした研究によると、文章をある程度理解するためには、その文章中に使用されている語彙のほとんどを知っていなければいけないと言われている（小森、三國、近藤、2004；Hu & Nation, 2000; Schmitt, Jiang, & Grabe, 2011）。例えば、Schmitt, Jiang & Grabe（2011）の研究では、世界 8 か国（トルコ、中国、エジプト、スペイン、イスラエル、イギリス、日本、スウェーデン）在住の 661 人の中級から上級レベルの 16 歳から 33 歳の英語学習者が参加し、ある程度背景知識があるトピック（地球温暖化と気候変動）と背景知識がないトピック（運動と思考能力に関する動物実験）を読み、それぞれについて内容理解を測る問題を解いた。さらに文章中に使われたものを含む語彙テストを受けた。その結果、ある程度背景知識のある文章であっても、文章中の語彙の 98%程度を知っていなければ十分な理解が得らないことが示された。

また、日本語学習者においても同様の結果が得られている。小森・三國・近藤（2004）による韓国・中国・台湾出身の 61 名の日本在住の留学生を対象にした研究では、古代エジプト人物画の特徴に関する説明文を読んだ被験者の内容理解度を多肢選択式問題、指示語の指示対象を特定する問題、文章全体の情報を統合し判断する問題等により測った。それによると、一部の問題において高得点が多く出すぎたことによる天井効果によって、被験者の差が十分に測定できなかったが、文章理解を促進する既知語率の閾値は 96%%程度であるとされた。

文章理解のために文章中の語彙の 98%が必要だとして、一般向けの小説を読むためにはどの程度の語彙力が必要になるのであろうか。Nation（2006）は、英語の小説（例えばフィッツジェラルド（1925）「 華麗なるギャツビー」）を対象に調査したところ、最頻 1000 語で 80%程度、2000 語で 87.7-91.7%をカバーするが、登場人物等の固有名詞も既知語と想定しても、98％をカバーするためには 8000 から 9000 語が必要であると結論づけている。Nation は子供用アニメ（シュレック）の視聴に必要な語彙数についても調べた。アニメでは、視覚・音声情報があるため、読むよりも文脈が取りやすいが、聞き逃すと戻って聞くことはできないなど単純には比べられないが、最頻 1000 語で 81.5%をカバーし、固有名詞も既知語とした場合、98%をカバーするためには、7000 語が必要であるとされた。

成人日本語母語話者の理解語彙は約 4 万語（林 1974）と言われているものの、日本語学習者にとっては、日本語で書かれた文章の理解には、約 1 万から 1 万 8 千語で十分だろうと言われている（押尾ほか 2008）が、それでは実際に、日本語学習者が一般向けの小説を読むためにはどの程度の語彙知識が必要になるのであろうか。

この問いに答えるため、本研究では以下の通り研究課題を設定した。

研究課題 1 日本語の一般的な小説を読むのに何語知らなければいけないのか。

研究課題 2 日本語のアニメを見るのに何語知らなければいけないのか。

3. 研究方法

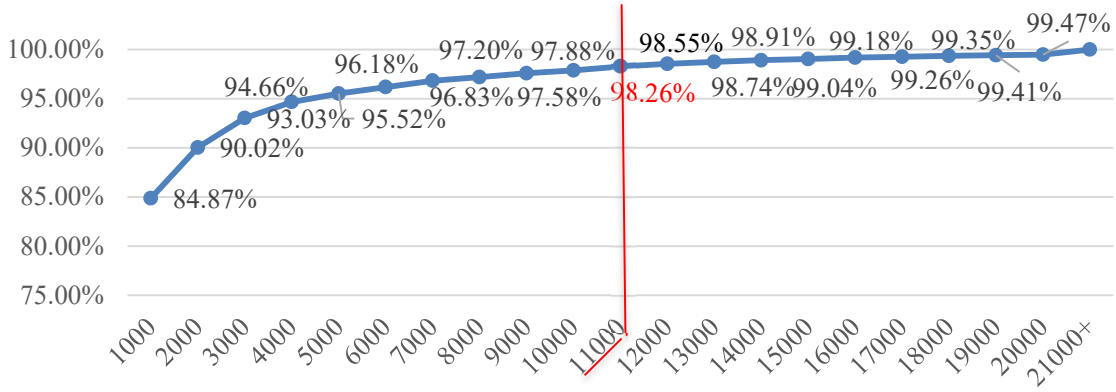
Nation（2006）の研究では著作権の切れた明治から昭和初期の英語の小説を対象としていたが、一般的な小説として映画化等によりある程度知られているものを選択することにして、予備調査として、東野圭吾（2005）の「容疑者 X の献身」を取り上げた。小説の本文中のふりがなを除去した上で、テキストマイニングのためのソフトウェア KH Coder（樋口 2022）により文章に形態素分析をかけて、小説の中で使用されている単語とその頻度数をリスト化した。さらに、単語を助詞や助動詞等の機能語、地名や登場人物の名前等の固有名詞、名詞や動詞等の内容語に分けた。機能語と固有名詞は既知語として想定し、内容語の異なり語一語ずつについて、「日本語を読むためのおよび語彙データベース」（松下 2011）および「J－LEX」（菅長、松下 2013）を参照し、一般的な文章における頻度を特定した。さらに、それらの述べ語数を頻度レベル 1000 語毎に合計し、内容語の頻度別 1000 語毎に全文章の語彙の何%をカバーするか推定した。同様にアニメの「となりのトトロ」と「千と千尋の神隠し」についても分析した。

4. 結果

4.1. 小説

「容疑者 X の献身」の形態素分析では、述べ語数で 97,094 語が検出された。このうち既知語彙として想定される助詞、助動詞などの機能語が 59,315 語と多くを占め、同様に会社名や登場人物の名前などの固有名詞が 3,643 語であった。一方、内容語は述べ語数 34,136 語、異なり語数 4,909 語であった。これらの語彙による文章のカバー率は以下のとおりである（図 1）。最頻 1,000 語によって 84.87%がカバーされ、2,000 語によって 90%に達するが、その後は 1,000 語毎にカバーされる割合が減っていき、6,000 語で 96%を超えるが、98%以上になるためには、11,000 語が必要だということが分かった。

図 1　「容疑者 X の献身」の 1000 語毎の累計カバー率



これらの異なる頻度レベルごとの語彙とは実際にどんなものであろうか。その例を表 1 に示す。それぞれの頻度レベルにおいて、この小説内で多く見られた順に並べ、語彙の後ろの数字はこの小説内の頻度数である。殺人事件の解決を描く内容のため、「警察」「事件」「殺す」「刑事」「犯人」「捜査」「容疑」「犯行」「指紋」などの語彙が多く現れていることが特徴的である。この表にある語彙では、何度も出現しているものが多く、小説を読むことで付随的語彙習得が進む可能性があると言える。

一方、この表に現れていない語彙では、各千語毎のカバー率が漸次低くなっていくことを見ても分かる通り、各語の出現回数は低くなる傾向にあり、その多くが全文中に 1 回しか現れないいわゆる one-timer（松下 2017）であった。頻度レベル 6 千語レベル以上では、平均すると約 6 割が one-timer であった。つまり、もともと頻度の高くない語彙が、小説を一冊読んでも 1 回しか現れないわけで、それらの語彙が学習者向けに加工していない生教材の難しさを示していると言える。たとえ、これらの語が文脈やその単語の部分（意味のはっきりした部首や、熟語の漢字の一つなど）から意味が推測できる場合でも、1 回しか登場しないなら、その単語の意味が付随的に学習される可能性は低いだろう。

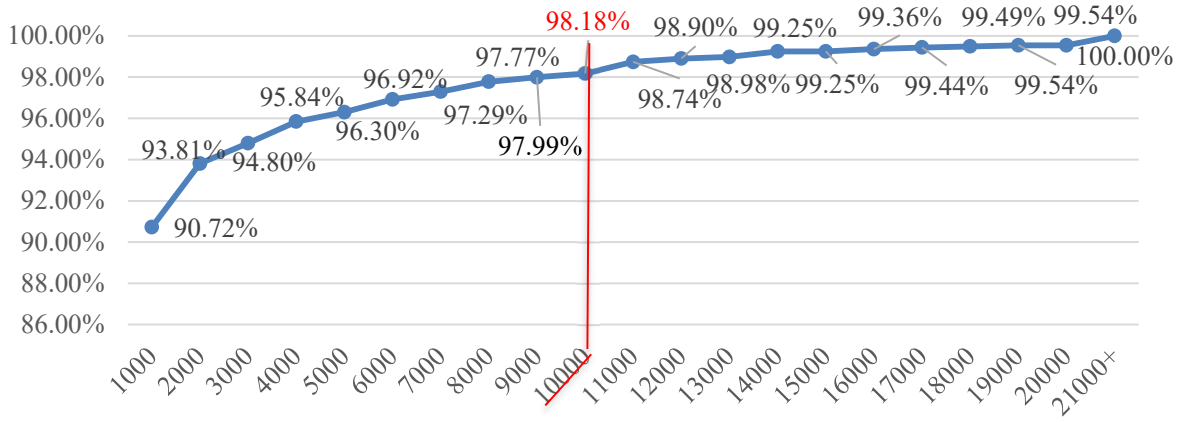
表 1　頻度レベル別単語の例（容疑者 X の献身）

1000	する 1891 言う 664 ある 482 思う 395 話 158 自分 143 電話 141 部屋 113 警察 95
2000	事件 116 数学 102 気づく 73 人間 75 殺す 67 振る 54 様子 47 確認 45 表情 43
3000	刑事 18 自転車 114 うなづく 64 犯人 53 捜査 48 疑う 46 見つめる 39 教師 37
4000	死体 102 コード 34 物理 34 口調 26 学者 25 浮かべる 23 つぶす 20 呟く 17
5000	見張る 21 チェーン 19 手帳 18 聞かす 11 立ち止まる 10 ビニール 10 天才 9
6000	容疑 24 見回す 14 ひねる 11 ひそめる 11 緩める 10 ためらう 10 衣類 9 真相 9
7000	犯行 31 口元 17 技師 14 亭主 13 カラオケ 12 傾げる 9 立ち去る 8 苦しめる 7
8000	学科 11 イラスト 9 突き止める 7 後ろ姿 5 関数 4 見据える 4 取り合う 4 すれ違う 3
9000	指紋 43 堤防 13 しかめる 13 用件 10 惚れる 8 かばう 8 ダルマ 7 思い込み 7
10000	うつむく 13 トリック 12 ホステス 12 用紙 11 柔道 10 新品 10 見せかける 7
11000	こたつ 36 ホームレス 16 インスタント 8 羽織る 6 幾何 5 レンタカー4 見破る 4
12000	アリバイ 64 ストーカー13 凶器 11 物音 7 身代わり 6 怪しむ 8 見返す 8 すくめる 6
13000	共犯 18 人殺し 7 盗難 6 呼び出し 6 絞める 8 細める 4 手間取る 2 償う 2
14000	歯車 16 ジャンパー11 空き缶 8 湯飲み茶碗 4 たじろぐ 4 見越す 3 安らぐ 2
15000	マグカップ 9 話し声 5 ウーロン茶 4 ウエイト 4 チェス 4 隠し 2 燃え尽きる 2
16000	貴女 16 班長 10 真顔 5 待ち合わせ 4 ココア 3 向き直る 4 うなだれる 4 散らかる 3
17000	テレホンカード 6 お手上げ 3 常人 3 抑揚 3 取り繕う 3 黙り込む 2 奥まる 1
18000	釈然 7 長髪 6 腕組み 5 ハンマー3 笑いかける 6 持ち去る 3 のけぞる 3
19000	数式 8 答案 6 期末 4 鑑識 3 張り込む 2 付け足す 2
20000	盲点 5 ボディガード 3 スカイライン 2 軍手 2 売り切れる 2 浮き立つ 2
21000+	バドミントン 11 勤怠 10 急用 8 ラケット 7 聞き込む 9 脱がす 8 提げる 4

4.2. アニメ

「となりのトトロ」を形態素分析したところ、述べ語数 3,730 語が検出された。このうち機能語が 1,880 語で固有名詞が 177 語であった。一方、内容語は述べ語数 1,673 語、異なり語数 572 語であった。最頻 1,000 語によって 90.72%に達するものの、カバー率が 98%以上になるためには、10,000 語が必要であった（図 2）。

図 2　「となりのトトロ」の 1000 語毎の累計カバー率



同様に「千と千尋の神隠し」のセリフを形態素分析したところ、述べ語数 7,413 語が検出された。このうち機能語が 3,918 語で固有名詞が 234 語であった。一方、内容語は述べ語数 3,261 語、異なり語数 949 語であった。最頻 1,000 語によって 89.11%がカバーされ、5000 語あれば、96%を超えるが、98%以上になるには、12,000 語が必要であることが分かった（図 3）。

図 3　「千と千尋の神隠し」の 1000 語毎の累計カバー率

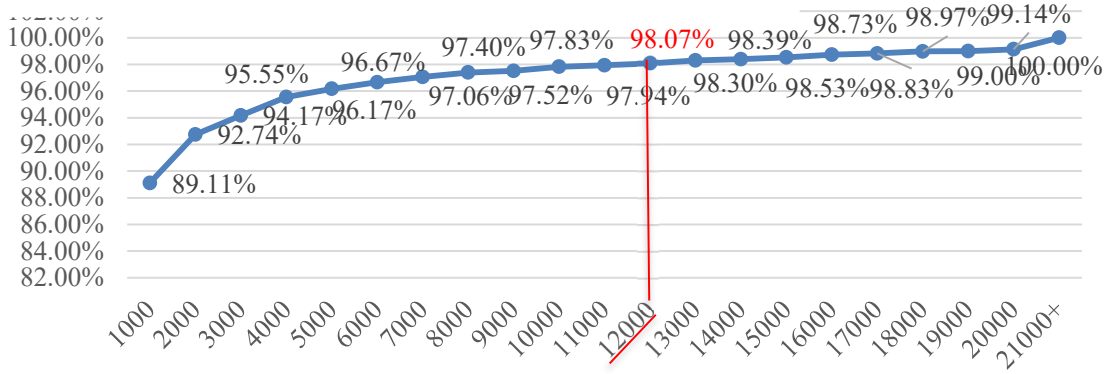


表 2 は、「となりのトトロ」の中での異なる頻度レベルごとの語彙の例である。それぞれの頻度レベルにおいて、多く見られた順に並べ、語彙の後ろの数字はこのアニメ内の頻度数である。頻度レベル 2,000 語からも一度しか現れない単語が多いとともに、頻度レベル 21,000 語以上のものも少なくなかった。

表 2　頻度レベル別単語の例（となりのトトロ）

1000　お母さん 29 お父さん 28 行く 22 出る 16 待つ 14 病院 13 来る 10 お家 5 カサ 5
2000　バカ 6 落ち着く 3 引っ越す 2 間違える 2 気がつく 2 壊す 1 サンダル 1 ヒゲ 1
3000　屋敷 5 クラブ 1 サル 1 ママ 1 天井 1 秘密 1 隠れる 1 迷う 1
4000　ネズミ 2 基地 1 手伝い 1 順番 1 助かる 1 縛る 1 湧く 1
5000　バイバイ 3 潰れる 2 泥 2 トンネル 1 髪の毛 1　延びる 1 汲む 1 腐る 1
6000　お化け 10 カニ 2 絵本 2 振り回す 1
7000　上がり 1 ボロッボロ 1 請う 1
8000　リス 4 電報 4 便所 2 田んぼ 1 本家 1 目玉 1 延ばす 2 漕ぐ 1 駐在 1
9000　トウモロコシ 2 見かけ 1 考古学 1 地蔵 1 妖怪 1
10000　クスノキ 1 バス停 1 仲良し 1
11000　ウソつき 1 眩む 1 田植え 1 くたびれる 1
12000　発車 2　木の実 3　雨戸 1
13000　結う 1 焦げる 1 ゴキブリ 1
14000　ドングリ 5 ひと休み 2 花屋 1
15000　（該当なし）
16000　迷子 3
17000　ダブる 2 天道 1
18000　キャラメル 1 市外 1
19000　南無阿弥陀仏（ナンマンダブ）2
20000　（該当なし）
21000+　手分け 1 早とちり 1 底抜け 1 寝過ごす 1 雨宿り 1 お巡りさん 1 笹 1 食べ頃 1

表 3　頻度レベル別単語の例（千と千尋の神隠し）

1000　する 116 いる 85 なる 55 行く 45 やる 30 ある 27　来る 26　戻る 19　言う 18
2000　人間 12　助ける 7 つく 6　ケガ 5 神様 5 食う 4 隠す 3 触る 3 盗む 3 お湯 3 当たる 2
3000　戻す 3 奪う 2 利く 2 断る 2 チャンス 2　平気 2　埋める 1　抑える 1　変わる 1
4000　弟子 4 助かる 3 離す 3 トラック 2　足下 2　別れ 2　くぐる 2, こする 2 焚く 2
5000　魔法 6　さめる 4 くっつく 2 呆れる 1 いやす 1　座敷 1　トンネル 1　バラ 1
6000　解ける 2 振り向く 2 汚す 2 操る 1 収まる 1 テーマパーク 1　バブル 1
7000　フン 4　団子 3　油断 3　グッド 2　石炭 2　暴れる 1 生き延びる 1 うなる 1
8000　匂う 5　魔女 5　守り 2 鎮める 2 取り消す 2 濁る 2 見捨てる 1
9000　おごる 1 混む 1 咎める 1 化ける 1　はじける 1　無礼 1　ノック 1
10000　トゲ 3　花束 3　透ける 2 極まる 1 まぎれる 1 見習う 1 替わる 1
11000　まじなう 3 こびりつく 1 紛れ込む 1 報う 1　やり遂げる 1　帰り道 1
12000　静まる 2 手先 2 裂ける 1 忍び込む 1 解き放つ 1
13000　坊や 2 仕向ける 1 つむぐ 1 粘る 1 乗り継ぐ 1 もがく 1
14000　ボーイフレンド 1 ベーコン 1 新米 1 硫黄 1　きょろきょろ 1　ほこら 1
15000　ねだる 3 煤 2　刺さる 1 客人 1 渡し 1 カンカン 1
16000　ハンコ 5　琥珀 3　引き入れる 2 連れ戻す 1 干物 1　手下 1　浅瀬 1　朝飯 1
17000　気前 2　おだてる 1 食わす 1 絞りだす 1　ボイラー1
18000　踏みつぶす 3 小娘 3　勘づく 1 泣かす 1 行き止まり 1　なにぶん 1　大儲け 1
19000　しおれる 1　新入り 1
20000　どのみち 2　こき使う 1 盗み出す 1 所用 1 上役 1　ありったけ 1　ハイカラ 1
21000+　薬湯 5 砂金 4 黒焼き 2 湯屋 2 番台 2 腹ペコ 2 甘ったれる 1 編み込む 1 いびる 1

「千と千尋の神隠し」の異なる頻度レベルごとの語彙の例を表 3 に示す。それぞれの頻度レベルにおいて、多く見られた順に並べ、語彙の後ろの数字はこのアニメ内の頻度数である。一度しか現れない単語も多く、頻度レベル 21,000 語以上のものも少なくなかった。アニメとはいえ、内容は必ずしも年少者のみを対象とした単純なものではなく、内容がより複雑な要素を伴っていることを反映していると考えられる。

以上のことから、研究課題に対する答えとしては、以下の通りである。

研究課題 1 日本語の一般的な小説を読むのに何語知らなければいけないのか。

テキスト中に使用されている言葉の 98%以上が既知語となるためには、11,000 語を知っていなければいけないと推定された。

研究課題 2 日本語のアニメを見るのに何語知らなければいけないのか。

10,000 語（となりのトトロ）から 12,000 語（千と千尋の神隠し）を知っていれば、セリフの中で使われている言葉の 98%以上が既知語となり、一定の理解が可能になるものと思われる。

5. 結論

本研究では、小説およびアニメの一定の理解を得るために必要な語彙知識量を推定すべく、予備的調査として、小説「容疑者 X の献身」、アニメ「となりのトトロ」及び「千と千尋の神隠し」の分析を行った。その結果、小説の既知語のカバー率が 98%を超えるためには 11,000 語、アニメにおいては、10,000 語から 12,000 語と推定された。これは、英語の小説の 8000－9000 語、アニメ 7000 語（Nation, 2006）よりも高く、日本語においては文章の内容理解のために必要とされる語彙知識量がより多い可能性が示唆された。この点については、今後サンプル数を増やし、アニメでは対象視聴者を考慮に入れて、さらに検討する必要がある。

また、語彙の分布の性質上、文章中に一度しか現れない単語、いわゆる「一発屋（one-tmer）」が多くあり、学習者のためには、それらを学習目標レベルのより頻度の高い言葉で書き換えて、遭遇回数の向上を図るべき重要性（松下 2017）が再認識された。

アニメのセリフは、「となりのトトロ」は、述べ語数で小説の「容疑者 X の献身」（約 97,000 語）の 3.8%（約 3,700 語）、「千と千尋の神隠し」は 7.6%（約 7,400 語）と圧倒的に少なく、複数回使用されている語彙も少ないことが示された。語彙習得のためには、対象語彙が複数回出現しないといけないことから、アニメ映画を 1 本 1 回視聴するだけではなく、シリーズのものを続けて見ることで絶対量を増やすとともに、内容の関連性からより多くの語彙に繰り返し遭遇する可能性を高める必要があると考えらえる。

参考文献

押尾和美・ 秋元美晴・武田明子・阿部洋子・高梨美穂・柳澤好昭・岩元隆一・石毛順子 (2008)「新しい日本語能力試験のための語彙表作成に向けて」『国際交流基金日本語教育紀要』4, 71-86.

小森和子・三國純子・近藤安月子 (2004)「文章理解を促進する語彙知識の量的側面－既知語率の閾値探索の試み－」『日本語教育』120, 83-92.

菅長陽一・松下達彦 (2013) 「日本語テキスト語彙分析器 J-LEX」 http://www17408ui.sakura.ne.jp/execut.php

林四郎 (1974)『言語表現の構造』明治書院

樋口耕一 (2022) KH コーダー https://khcoder.net

松下達彦 (2011) 日本語を読むための語彙データベース (VDRJ) Ver. 1.1 (研究用)

松下達彦 (2017)「日本語読解テキストのリライトの重要性とアプローチ－語彙的要素を中心に－」『日本言語文化研究会論集』13, 1-18.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language, 11*(2), 207-223.

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430.

Matsushita, T. (2012). *In What Order Should Learners Learn Japanese Vocabulary? A Corpus-based Approach.* Victoria University of Wellington. PhD Dissertation.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review-Revue Canadienne Des Langues Vivantes, 63*(1), 59-82.

Nation, I. S. P., & Anthony, L. (2013). Mid-frequency readers. *Journal of Extensive Reading, 1*, 5-16.

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do *Things Fall Apart*? *Reading in a Foreign Language, 22*(1), 31-55.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26-43.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language, 15*(2), 130-163.