日本語語彙習得に関わる普遍性と個別性
－漢字をめぐる問題を中心に－

松下達彦（東京大学）

1. はじめに

大仰なタイトルをつけてしまって若干後悔しているが，これにはわけがある。私事で恐縮だが，私は海外の応用言語学プログラムで博士課程を修めたので，欧米を中心に発達した研究成果を日本語に応用できるかどうか，逐一考えざるを得なかった。そのため，常に日本語教育と英語教育等の共通点と相違点を考える習慣ができた。

応用言語学，第二言語教育の枠組みは共通で，世界の研究成果は大いに参照すべきである。しかし，一方で，どうしても日本語固有の問題として考えざるを得ない問題もいくつかある。特に，語彙面では漢字の問題を外して考えることはできない。漢字の造語力と（一つの書字形態が複数の音韻形態に対応する）音訓システムが世界に類を見ない複雑な表記体系を生み出し，語レベルでもさまざまな学習・教育上の問題を生んでいる。

本稿では，語彙をめぐるいくつかの点について他言語と共通の枠組みとそうでない点について述べる。特に，漢字の問題を中心に，日本語固有の語彙習得問題として考えなければならない問題をいくつか取り上げる。最後に，本シンポジウムの趣旨に則り，語彙習得と文法習得の関係について少しだけ述べたい。

なお，本稿でいう普遍性とは英語などの他言語と共通であるという程度の意味で，必ずしも心理言語学的な意味で言っているものではないことを初めにお断りしておく。

2. 語彙力の重要性

４技能に占める語彙力の割合は相当高く，各技能の運用力の分散を最大で 0.5 以上説明できる(Milton, 2013)。Jeon & Yamashita (2014)は 4000 以上の論文から基準を満たした 59 の論文を選んで再分析したメタ分析の論文であるが，それによれば，文法知識，語彙知識がそれぞれ読解力と r =.85, r =.79 程度の相関をもち，これを二乗しても r^2 =.72, r^2 =.62 となるので，いずれも読解力の分散の半分以上を説明できると言えよう。

日本語では調査が少ないが，小森・三國・近藤(2004)は語彙力が読解力の分散をどの程度説明できるか計算し r^2 =.47 としている。Koda (1989)の語彙と読解の相関係数から計算してみると r^2=.55 である。野口(2008) は 2005 年実施の旧日本語能力試験を分析したもので，文字・語彙と読解の間で最も相関の低い２級レベルでは r^2 ＝.44 だが，初級に相当する 3 級と 4 級では相関係数 r は.8 を超え，r^2 はそれぞれ.64 と.67 になる（p.62 の表の数値より計算）。調査対象や調査方法が異なるので単純な比較は難しいが，いずれにしても語彙力が運用力に占める割合が相当に高いことは日本語も他の言語も同様である。

3. 語彙力と漢字力の関係

3.1 一般的な関係

語彙力と漢字力には，当然のことながら高い相関がある。松下・佐藤・笹尾・田島・橋本(2016)では，「日本語を読むための語彙量テスト（5 万語レベル）」(VSTRJ-50K)と漢字変換力テスト(KCT)の総得点の相関係数は r =.82 で，ラッシュ分析の能力値で見ると r =.85 である。一般的に語彙量と漢字変換力にはかなり高い相関があるといえる。いずれも多肢選択式で，前者は非定義文に埋め込まれたターゲット語の定義を選ぶ形式で漢字かな混じり表記と意味を結びつける能力を要求し，後者はカナ表記されたターゲットの漢字を選ぶ形式で（疑似的に）音韻から意味に合う漢字を結びつける知識を要求するもので，要求される知識は一致しないが，それでもこれだけ高い相関がある。ただし，KCT ではカナ呈示に変えたことで，疑似的にターゲット語の音韻知識と書字知識を同時に測る形になっており，それが後述の第一言語の影響を低減し，語彙知識との相関を高めている可能性もある。VSTRJ-50K と KCT では後者の方が第一言語の影響が少ないことがわかっている。

基本的に漢字力は語彙力と類似した能力であり，いずれも 4 技能と高い相関がある。おそらく英語では，語構成要素の理解の能力と語彙力の間にこれほど高い相関は出ないであろう。

3.2 「読む・書く」の語彙力と「聞く･話す」の語彙力

一般にこの二つの違いは応用言語学でそれほど意識されていないように思われる。表音文字を使用し，発音とスペリングの関係が透明な（＝書かれたとおりに発音すればいいような）言語で，識字力のある成人であれば，この二つに大きな差は出ないと考えられる。例えば英語の Vocabulary Size Test (Nation & Beglar, 2007)は文字で実施されるが，それを読み書き限定の語彙力だと注記されることはほとんどない。しかし，文字の認知に負荷の多い漢字を大量に使用する日本語では，非漢字圏学習者の場合，「聞く･話す」の語彙力を「読む・書く」の語彙力よりも先に発達させることが考えられ，漢字圏学習者の場合はその逆になる可能性が高く，いずれにしてもこの 2 種類の語彙力は大きく異なるので，松下 (2012a) では「日本語を読むための語彙量テスト」と称している。現在，音声呈示による語彙テストの開発を計画している。

3.3 語彙力と漢字力への第一言語の影響

言うまでもなく，上述の語彙力の差は，第一言語（L1）とも大きく関わる。L1 が中国語であれば，「読む・書く」の語彙力が「聞く･話す」の語彙力よりも高く出ることが予想されるが，逆に非漢字圏学習者では後者の方が高くなるのが一般的だと考えられる。

例えば松下ほか (2016) は VSTRJ-50K について，日，中，韓，その他の四つの言語タイプ別に DIF (Differential Item Functioning)分析を行い，L1 別，語種別に VSTRJ-50K の項目難度を比較すると，L1 中国語学習者にとっては相対的に外来語が難しく，漢語は易し

いが，英語やその他の L1 の学習者は全くその逆であることを報告している。これは明らかに漢字を媒介とする L1 知識の転移の有無，および非漢字圏学習者にとっての漢字学習の困難が原因だと思われる。

松下・陳・王・陳(2017)によれば，上位 2 万語のうち，30%にあたる約 6 千語がいわゆる日中同形同義語で，5%に相当する約千語が同形類義あるいは同形異義の要注意語である。中国語母語学習者の多くは学習しなくても日中同形同義語を目で見て理解できるが，学習しなければ音で聞いても理解できない語が多い。そのような語が 30%も存在するのである。これはかなり巨大な built-in lexicon (Dalton, 2004)だと言えよう。

4. 学習単位への負荷に見る日本語と英語の違い

4.1 テキストカバー率のデータから見る漢字の重要性[1]

一定のテキストカバー率に到達するのに，日本語では英語よりも多くの語数が必要である。例えば英語では 95%のカバー率に到達するのに word family で 4,000 語程度必要である(Nation, 2006)。word family とは，屈折変化による語形のほか，接辞付加による派生形まで含むもので，Bauer & Nation (1993)が Level 6 と呼んでいる単位である。例えば develop という語の word family には develops, developed, developing という屈折変化のほか，developable, undeveloped, developmentally, predevelopment といった派生形まで含む。日本語では lemma に近い単位を「語」と呼ぶことが多い[2]。lemma は屈折変化による語形は含むが，接辞付加による派生形は含まない。概ね辞書の見出し語に近い単位とされる。あるテキストにおける lemma 数は，おおよそ word family の 1.6 倍と言われるので(Milton, 2009, p.12)[3]，英語では 95%のカバー率に到達するのに lemma で 6,400 語程度だということになる。98%のテキストカバー率で見ると，書き言葉の場合，word family で 8,000～9,000 語である(Nation, 2006)から，lemma に換算すればおおよそ 12,800～14,400 語程度である。ところが日本語では，95%のカバー率を達成するのに 9,000～12,000 語，98%を達成するのに 20,000～24,000 語も必要である(Matsushita, 2012, p.155)。この数字だけを見ると，日本語の語彙学習負担は英語よりもかなり大きいということになる。

ところが，日本語では，ひらがな＋基本漢字 300 字の組み合わせで語のテキストカバー率は 80%に到達し，漢字上位 1000 字（ほぼ教育漢字に相当）で 95%に到達する（表 1）[4]。テキストカバー率 95%というのは，英語で言えば 6,400 語（lemma）程度のカバー率と同レベルである。漢字の学習は非漢字圏学習者にとって簡単ではないが，それでも語数に比べればうんと少ない「ひらがな，カタカナ，アルファベット＋漢字 1000 字」程度で 95%のテキストカバー率が達成可能になるのである。

---

[1] 本節の内容は，本稿筆者の PhD 論文(Matsushita, 2012)の第 6 章の内容に基づく。

[2] 国立国語研究所の語彙資料などで伝統的に用いられ、形態素解析辞書 UniDic の短単位の「語彙素」として受け継がれている単位は lemma に近い。

[3] Milton はこの 1.6 という数値の根拠を明確に示しておらず、非常に大雑把な推計だとしている。

[4] 表 1 は松下(2013, p.240)の表 6-1 を日本語に訳したものである。

表１　日本語の文字種別・漢字順位別の書字形異なり語数とその割合・テキストカバー率

文字種 (+漢字レベル)(*) A: アルファベット ひ: ひらがな カ: カタカナ	指定文字種によってカバーされる書字形異なり語数	指定文字種によってカバーされる累積書字形異なり語数	語（書字形）によるテキストカバー率（%）	語（書字形）による累積テキストカバー率（%）	文字によるテキストカバー率（%）	文字による累積テキストカバー率（%）
アルファベットのみ	17,712	17,712	0.7	0.7	1.1	1.1
ひらがなのみ (*)	20,272	37,984	59.7	60.4	51.9	52.9
アルファベットとひらがなの混合	1	37,985	0.0	60.4	0.0	52.9
カタカナのみ (*)	49,349	87,334	3.3	63.6	7.3	60.2
アルファベット・ひらがな・カタカナの混合	625	87,959	0.0	63.6	0.0	60.2
頻度順 1- 100 位の漢字 +A&ひ&カ	7,187	95,146	10.1	73.8	9.7	70.0
頻度順 101-200 位の漢字 +A&ひ&カ	7,360	102,506	5.2	79.0	5.8	75.8
頻度順 201-300 位の漢字 +A&ひ&カ	7,318	109,894	3.6	**82.6**	4.1	79.9
頻度順 301-400 位の漢字 +A&ひ&カ	6,636	116,530	2.8	85.4	3.3	83.1
頻度順 401-500 位の漢字 +A&ひ&カ	6,830	123,360	2.6	88.0	2.9	86.0
頻度順 501-600 位の漢字 +A&ひ&カ	6,820	130,180	2.0	90.0	2.4	88.4
頻度順 601-700 位の漢字 +A&ひ&カ	6,585	136,765	1.6	91.6	1.8	90.2
頻度順 701-800 位の漢字 +A&ひ&カ	6,393	143,158	1.4	93.0	1.6	91.8
頻度順 801-900 位の漢字 +A&ひ&カ	6,186	149,344	1.1	94.1	1.4	93.2
頻度順 901-1,000 位の漢字 +A&ひ&カ	5,427	154,771	1.0	**95.1**	1.2	94.4
頻度順 1,001-1,100 位の漢字 +A&ひ&カ	4,703	159,474	0.8	96.0	1.0	95.3
頻度順 1,101-1,200 位の漢字 +A&ひ&カ	4,262	163,736	0.7	96.6	0.8	96.1
頻度順 1,201-1,300 位の漢字 +A&ひ&カ	4,222	167,958	0.6	97.2	0.7	96.8
頻度順 1,301-1,400 位の漢字 +A&ひ&カ	3,691	171,649	0.5	97.7	0.5	97.4
頻度順 1,401-1,500 位の漢字 +A&ひ&カ	3,541	175,190	0.4	98.1	0.4	97.8
頻度順 1,501-1,600 位の漢字 +A&ひ&カ	2,909	178,099	0.3	98.4	0.4	98.2
頻度順 1,601-1,700 位の漢字 +A&ひ&カ	2,793	180,892	0.3	98.6	0.3	98.5
頻度順 1,701-1,800 位の漢字 +A&ひ&カ	2,554	183,446	0.2	98.9	0.3	98.7
頻度順 1,801-1,900 位の漢字 +A&ひ&カ	2,164	185,610	0.2	99.0	0.2	98.9
頻度順 1,901-2,000 位の漢字 +A&ひ&カ	1,993	187,603	0.2	99.2	0.2	99.1
頻度順 2,001-2,100 位の漢字 +A&ひ&カ	1,933	189,536	0.1	99.3	0.1	99.3
頻度順 2,101-2,200 位の漢字 +A&ひ&カ	1,495	191,031	0.1	99.4	0.1	99.4
頻度順 2,201-2,300 位の漢字 +A&ひ&カ	1,427	192,458	0.1	99.5	0.1	99.5
頻度順 2,301-6,323 位の漢字 +A&ひ&カ	15,373	207,831	0.5	100.0	0.5	100.0
頻度順 1-6,323 位の漢字 +A&ひ&カ	207,831	207,831	100.0	100.0	100.0	100.0

*延べ語数 33,153,137、延べ文字数 54,858,368。

*ひらがな・カタカナには長音記号「ー」を含み、漢字にはその他の記号（「々」など）を含む。

*漢字の順位は「現代日本語文字データベース」（CDJ）（松下2013）に基づく。

もちろん漢字を知っていれば自動的に漢字語を知っているということを意味するわけではない。つまり，日本語学習においては，語構成要素である漢字の字形，基本義，語構成ルールやメタファー，音と訓の関係などへの理解が重要であることを意味している。日本語教育のための日本語研究のためには，漢字語の意味的透明度（単漢字の意味からどの程度語の意味を推測できるか）を調査してデータベース化することは，大いに意味がある。意味的透明度を数値化できれば，テキストの語彙的負荷の測定にも役立つはずである。

多くの語が漢字の組み合わせでできており，初見の語であっても漢字を見れば意味も読

みもわかることが多い。それができれば語彙学習の負担は大幅に軽減するはずである。もし，一つの漢字をめぐる複数の漢字語（音と訓を含む）が，脳内の漢字の字形のイメージを媒介にリンクされていないと，おそらく日本語語彙学習の負担は相当に大きなものになるであろう。「変える」「変わる」「変化」といった高頻度語に使われている「変」の基本義がわかれば，「変更」「変換」「豹変」「一変」「改変」といった語が「変」の字形イメージを媒介にリンクされ，理解が早くなるということである。

これはおそらく，話しことばの語彙習得においても重要である。漢字音には同音のものが多く，例えば「変換」の「変」には「返」「辺」「編」「偏」「片」など多数の同音字がある。「換」も同様で「間」「館」「漢」「感」「観」「缶「官」「管」「韓」など相当な数になる。このときに「変わる」「換える」などが字形イメージを媒介に /ヘン/，/カン/ という読みとリンクしていれば，文脈の助けによって漢字の特定と意味理解が促進されると思われる。

英語教育においても語構成要素の理解は語彙学習を助けるとされている。しかし，その重要性のレベルは日本語学習においてはるかに高いように思われる。日本語においては，意味理解において，部首などの漢字構成要素の意味，単漢字の意味，語の意味，というように，複数レベルの意味が重なり合ってできている。複数の異なるレベルの学習単位があり，その組み合わせへの理解が重要で，しかも，それは書きことばの学習に留まらず，話しことばの語彙学習にも深く関わっていると考えられる。

4.2 漢語系接辞の重要性と１字漢語の問題

上節では word family という分析単位をいちいち lemma という単位に換算してきた。英語ではしばしば使われる word family を直接に日本語の分析に応用できればよいが，それはおそらく難しい。日本語には語と同様の具体的意味を持ち，語と同レベルの学習負荷を要する接辞が大量に存在するためである。例えば，「会議室」の「-室」は接辞だが，部屋という具象物を示す。「富士山」の「-山」も /-サン/ と読めば接辞だが，意味的には語と変わらない。「会議室」や「富士山」を「会議」や「富士」の word family として扱うと、英語などで言う word family の概念を大きくはみ出すように思われる。しかも，「山」は/ヤマ/ と読めば語になる。同一表記語で異なる音声形態を持ちえるので、読み方によって word family になるかどうかが変わってしまう。「創造主」の「-主」は接辞だが，「主」を/ヌシ/，/アルジ/，/オモ/と読めば語である。

このようなケースがあるため，日本語の接辞の数を正確に数えるのはかなり難しいが，接辞として使用できる成分が少なくとも 753 個も存在する（Matsushita, 2012, p.86)。接辞にも語にもなる「山」のようなケースを含めると相当な数に上るはずである。これは 100 にも満たない（Bauer & Nation 1993 では 91 個しかない）英語の接辞に比べてかなり多いし，日本語辞書の見出し語に接辞が立てられ，多くの語彙研究で語として数えられることが多いのもうなずける。

しかも，漢語系接辞は学術テキストにおいて多用される。文章の硬さを品詞の割合で見

る場合，漢語系接辞は最もよい指標だと考えられる(Matsushita 2012 の Chapter 4)。

5　学術系語彙の習得と Lexical Bar について

日本語の学術系語彙（学術共通語彙と専門分野別の語彙）の 7 割以上は漢語である。これは文芸語彙（文芸テキストに特徴的な語彙）の 7 割以上が和語であるのと対照的である。学術系語彙は学術テキストでは頻出するが，日常会話ではあまり用いられないため，学術テキストに触れて意図的な学習を行わないと習得が難しい。例えば中級レベルの学術共通語彙(松下 2011, 2016; Matsushita, 2012)のレベル I の 559 語は，学術テキストでは 11.1%のテキストカバー率が出るが，日常会話では 0.8%しかない(Matsushita 2012, p.301)。一方，文芸テキストによく用いられる語彙（日本語文芸語彙，松下 2012b）は上位 1 万語程度までは日常語との重なりが大きく、1 万語を超えるレベルになると日常語とは異なる純粋に文芸的な語彙が増える(松下 2016)。この関係を図式的に示したのが図 1 (石澤・岩下・伊志嶺・桜木・松下 2018)である。

問題は，学術系語彙が日常の語彙とは離れていて，そこに語種の違いがあり，漢字の学習負担も伴うことである。実は，英語でも類似の問題が指摘されている。英語の日常語はゲルマン系の言語をベースにしているが，学術系の語彙にはギリシャ語･ラテン語を起源とする語が多い。Corson (1985)はこれらの語彙の習得が社会文化的な背景と関係していることを指摘し，Lexical Bar と呼んだ。日本語も英語も語種が学術系語彙と関わっている点が共通だが，これは普遍的とも個別的とも言えない。しかし，学術系語彙の使用域が日常語彙と大きく異なり，その習得に困難が伴う点は共通である。

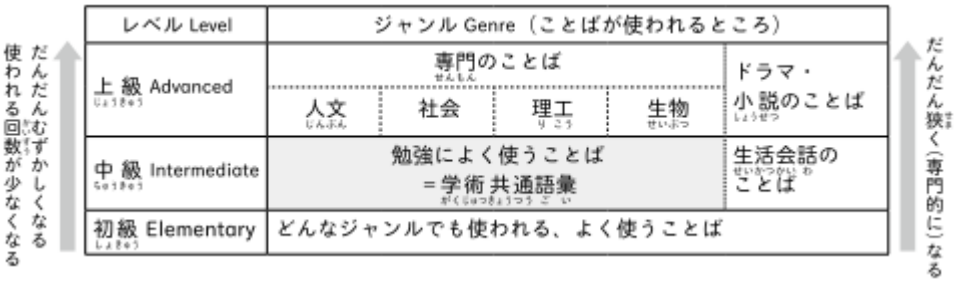

図 1　語彙のジャンルとレベルの関係を簡略化した図（石澤ほか 2018, p.iv より）

6. 語彙習得と文法習得の関係

文法習得は狭義には語順の習得であるが，広義には機能語の用法の習得だと考えられることが多い。いわゆる助詞・助動詞だけではなく，副詞や接続表現の習得まで含めることもある。要は，語彙習得のうち，意味の抽象度の高いケースやふさわしい文脈の習得などは，広義の文法習得だと考えられやすいということであろう。日本語能力試験の文法の問題を見ても，上級レベルの問題ほど語彙の問題だと思われるものが少なくない。

語彙習得が進めば文法習得に必要な文脈を豊富にすることができる。筆者はかつて「中国語母語学習者のための語彙学習先行モジュール」を提唱した(松下 2002; 2005)。

また，語の連続をかたまりとして処理するチャンキングが進むことで読解や聴解の流暢さ（単位時間当たりの言語処理量）は上がっていくであろう。一方，機能語の習得は抽象度が高いため，おそらくボトムアップ処理とトップダウン処理の両方を適用する中で進んでいくものである。例えば，ある子どもは体のどこかが「痛い」ことをすべて「おなかが痛い」と表現する時期があった。おそらくその後に「おなか」や「痛い」を習得し，「が」の意味・用法も習得されていくものと考えられる。

語彙と文法の習得は相互に助け合ってスパイラルに進むものであろう。

7. まとめ

本稿では，初めに語彙力の重要性について述べた後，語彙力と漢字力の関係について，「読む・書く」の語彙力と「聞く･話す」の語彙力の違いにつながっていること，それが L1 の異なりによる相違と関係していることを指摘した。さらに，テキストカバー率と語彙数，漢字数の関係のデータから，日本語学習においては，語構成要素である漢字の字形，基本義，語構成ルールやメタファー，音と訓の関係などへの理解が重要であることを指摘した。さらには，日本語には統語的には接辞だが意味的には語とも言える語構成成分が多いことや語にも接辞にもなる 1 字漢語の存在を指摘した。また，学術系語彙が日常では接触頻度が少ないこと，語種の違いと関わること，それが日本語においては漢字学習負担とも重なっていること，英語にも類似の問題があることを指摘した。最後に，語彙習得と文法習得の関係について，思うところを略述した。

＊本稿は 2018 年 12 月にシンポジウム会場で発表された予稿集原稿のウェブサイト上への掲載にあたり，一部加筆修正を施して 2019 年 1 月にサイト上で発表されたものである。

参考文献

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！　大学で学ぶためのことば』vol.1.，くろしお出版

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』125, pp.83-92

野口裕之（2008）「試験結果の分析」国際交流基金・日本国際教育支援協会編『平成 17 年度日本語能力試験　分析評価に関する報告書』凡人社，pp.45-111

松下達彦（2002）「中国語を母語とする日本語学習者のための語彙学習先行モジュールの提案　～第二言語習得理論、言語認知、対照分析、語彙論の成果を踏まえて～」《日语学习与研究》, 2002 年第 1 期, pp.50–54

松下達彦（2005）「語彙学習先行モジュールの日中バイリンガル児童・生徒への応用　―母

語の漢字知識を活かす―」『母語・継承語・バイリンガル教育研究』創刊号，pp.84-95
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
松下達彦（2012a)「日本語を読むための語彙量テスト」の開発」『2012 年日本語教育国際研究大会予稿集第一分冊』pp.310
松下達彦（2012b）「日本語文芸語彙リスト」Ver.1.0.
URL:http://www17408ui.sakura.ne.jp/tatsum/list.html#jlw
松下達彦（2013）「現代日本語文字データベース」
URL:http://www17408ui.sakura.ne.jp/tatsum/CDJ-top.html
松下達彦（2016）「コーパス出現頻度から見た語彙シラバス．ニーズを踏まえた語彙シラバス」くろしお出版，pp. 53-77
松下達彦・佐藤尚子・笹尾洋介・田島ますみ・橋本美香（2016）「第一言語・第二言語としての日本語語彙量と漢字変換力の測定」『2016 年日本語教育国際研究大会予稿集』（USB メモリによる会場配布のみ）
松下達彦・陳夢夏・王雪竹・陳林柯（2017）「日中対照漢字語データベースの開発と応用」『2017 年度日本語教育学会秋季大会予稿集』pp.366-371
Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253–279.
Corson, D. J. (1985). *The Lexical Bar*. Oxford: Pergamon Press.
Dalton, F. E. (2004). *Gairaigo* -- The built-in lexicon? The common loanwords in Japanese based on high-frequency English vocabulary and their effect on language acquisition (Unpublished doctoral thesis). Victoria University of Wellington, New Zealand.
Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, 64(1), 160–212.
Matsushita, T. (2012). In what order should learners learn Japanese vocabulary? A corpus-based approach (PhD thesis). Victoria University of Wellington.
Milton, J. (2009). *Measuring Second Language Vocabulary Acquisition*. Bristol: Multilingual Matters.
Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.) *L2 Vocabulary Acquisition, Knowledge and Use* (pp. 57–78).
Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*(1), 59–82.
Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.