

語彙・漢字頻度表示器J-LEXによる リライト教材作成

CREATING LEXICALLY CONTROLLED MATERIAL TEXTS
BY J-LEX,
AN ONLINE LEXICAL ANALYZER OF JAPANESE TEXTS

松下達彦(東京大学)

Tatsuhiko Matsushita

The University of Tokyo

University of Hawaii, Kapiolani Community

College, August 7, 2015

本日のワークショップ

1. はじめに
2. リライトにおける語彙レベルコントロールの理論的な基礎
3. J-LEXのベースワード
4. J-LEXを利用したテキストのリライト

1. はじめに: 内容重視の活動の重要性

語彙レベルのコントロールを要する活動

- 内容重視の第二言語教育 (CBI: Content-based Instruction)

岡崎(1994,2002): 多読は内容重視の日本語教育の一つ

ハドソン遠藤(2011): 短編小説の講読コース設立を提案

- 速読
- 多読

- 文学ばかりでなく、時事的な話題や専門分野など多様な内容で
- コミュニケーションの目的を言語教育の目的とすることは自然

より大きな教育目標

- 批判的思考、創造性、自律性の獲得
- 地球的な視野や国際社会の一員としての知識や行動力の獲得

はじめに：言語の調整の重要性（1）

内容重視の活動

- 産出（話す／書く）を中心にした活動
 - 言語レベルは自然に調節
 - 相互確認も可能
- 受容的な活動（読む／聞く）の場合、対象テキストの語彙や文法のレベルが高すぎると、内容重視の活動は成り立たない
- 多くの生素材は、そのままでは語彙や文法の負荷が高すぎて中級学習者には使えない

リライトは内容重視の言語教育、特に中級学習者のためには非常に重要

はじめに：言語の調整の重要性（2）

- 行政サービスにおける「やさしい日本語」

例：横浜市「やさしい日本語での情報発信について」

<http://www.city.yokohama.lg.jp/lang/ej/kijun.html>

語彙レベルのコントロールがきわめて重要

科研費基盤研究A「やさしい日本語を用いた言語的少数者に対する言語保障の枠組み策定のための総合的研究」（代表者：庵功雄）

1. リライトにおける語彙レベルコントロール の理論的な基礎：テキストカバー率

テキスト中の既知語カバー率と読解レベルの関係
＝リライトの理論的基礎

英語では

- 55%～70%の内容理解に既知語のカバー率95%
自力で読解を楽しむには98%程度必要 (Hu & Nation, 2000;
Laufer & Ravenhorst-Kalovski, 2010など)

日本語では、

- 96%あたりに読解レベルの閾値？ (小森ほか2004)
閾値：読解レベルが急激に上がるポイント
- 既知語数12000語程度でカバー率93%程度に達し、約7割の理解
度になると推計 (Matsushita, 2014)

いずれにしても

例えば、語彙量約5000語の学生が、上位5000語のカバー率93%未
満の文章を読むのは無理がある

リライトにおける語彙レベルコントロールの理論的な基礎：辞書があればいい？(1)

辞書を引いて読めればよい？

- オンラインでポップアップ辞書や語彙リストの自動作成機能を利用することは言語の学習を大いに助ける 例えば MOLE (土竜ちゃん)

しかし

- 多読：何回も立ち止まっては読むという行為自体が成立しない (Day & Bamford, 2002 など)
 - graded readers：厳しい語彙レベルのコントロール
- 速読：既知語率は100%であるべき (Nation, 2007)
 - 途中で止まって考えること：学習にはなっても流暢さを上げること (fluency development) にはならない

リライトにおける語彙レベルコントロールの理論的な基礎：辞書があればいい？(2)

- 習得という観点：focus on form の考え方に従えば、形式と同時に意味の処理に意識が向いていることが必要
→一定レベルの既知語テキストカバー率に達していることが必要条件

結局、

- 内容重視の第二言語教育や多読、速読においては、言語形式の学習のための精読などよりも高いレベルの既知語率が求められる
 - 学習者の語彙レベルを適切な方法で測定
 - テキストの語彙レベルを適切にコントロール
- →最も効率的な内容重視の活動

リライトにおける語彙レベルコントロールの理論的な基礎: そのほか

- 読解に占める下位能力のうち語彙力が最も大きな要素であるということ (Bernhardt, 2005; Matsushita, 2012など)
- 語彙はおおよそ頻度順に習得が進む (Read, 1988など)

3. J-LEXのベースワード(1)

- ベースワード: 語彙・漢字頻度プロファイラーに仕込まれている語彙・漢字の頻度リスト
- J-LEXでは松下(2011, 2013)に基づいて作成されたベースワードが仕込まれている
- ベースワードの有効性: さまざまなテストコーパスにおけるテキストカバー率が、例えば旧日本語能力検定試験の級別語彙リストよりも高いことなどで実証済み(Matsushita, 2012)

J-LEXのベースワード(2)

- BCCWJの2009年モニター版
- 書籍約2800万語、Yahoo知恵袋約500万語
- 頻度と分散度を掛け合わせた指標でランク付け (Juilland's u)
- 初級語彙(上位1285語)だけは、旧日本語能力試験の4級3級語彙(一部、古臭い語だけは除外)を利用

4. J-LEXを利用したテキストのリライト

1) オンライン語彙・漢字頻度プロファイラー
J-LEX(菅長・松下2013)

<http://www17408ui.sakura.ne.jp/index.html> を
開く

2) 分析したいテキストをウィンドーに貼りつける

3) ドロップダウンリストから使用したいベース
ワード(語彙または漢字)を選ぶ

4) 語彙レベル／漢字レベル等を数字で設定する
(例: [3000]位以上を赤でハイライト)

5) 判定ボタンを押す

6) 右側の結果ウィンドーを見て、適切なテキストカバー率が達成されていない場合、赤でハイライトされている語を減らすように左側のウィンドーのテキストを書き直し、再び判定ボタンをクリックする。

(リライトの過程を残したい場合は、右側の結果ウィンドーのテキストをコピーし、MS-Wordなどにペーストする。色もそのまま保持される。その後、「変更履歴の記録」の機能を使用する)

7) テキストが適切な語彙レベルに調節された後に、調整済みテキストを再びウィンドーに貼り付け、今度は漢字レベルを指定して判定する
指定レベルよりも頻度の低い漢字は

- 1) ふりがなをつける
- 2) かな表記にする
- 3) 他の語に置き換える

などの対応をとる。

具体的なコツ

内容やリズムをできるだけ損なわないようにする

語彙面では、ただ単に低頻度語彙を減らすのではなく、以下のように考える。

中級後半クラス対象の素材だとしたら、

- 書き換え後に95%点が7K以下程度になるようにする。
- 書き換え後に98%点が10K程度になるようにする。
- ただし、例えば95%点を考える際に、95%を超えるレベルに以下のような語がある場合は

それを95%以下の既知語に準じるものと考えてよい。

繰り返し使われるキーワード

- 書きことばでは低頻度でも、話しことばでは高頻度でなじみがあると思われる語
- 外来語や漢語で、学習者母語などの知識から正しい意味が理解できる語
- 単漢字の組み合わせで意味が容易に分かる漢字語
- 語構成要素の組み合わせで容易に理解できる複合語（特に「突き倒す」のような複合動詞など）
- 誤解析によって低頻度語とされているもの（こういうこともあります・・・ご容赦を！）

上記諸点を考慮に入れたうえで、

- 98%点を超える語はできるだけ高頻度語に書き換えるか、削除する
- 95%点から98%点あたりの語(理想的には7Kから10Kあたり)の語を学習ターゲット語彙と考慮して、このレベルの語が1回しか使われていなかったら、できるだけ“複数回登場”するように書き換える。

(1回しか出てこない語は覚えられないため。)

- 高頻度語に入れ替える simplification だけでなく、
 - 関係節や挿入句で語釈的な表現を文章に混ぜ込む elaboration も使ってよい。
 - 難しい構文や慣用句も多ければ減らす。
 - と同時に、不自然にならないようにする。
- この辺の総合的なバランスが難しいところですが・・・

J-LEXを利用したテキストのリライト

これらのほか、

- J-LEXは頻度レベルごとの延べ語数や異なり語数、累積テキストカバー率などをジャンル別に返す機能がある(図2)
- 1000語レベル別に語数、カバー率などを表で示す機能もあり、特に累積テキストカバー率を見れば、どのレベルの語をリライトすればよいか、どのレベルの語を学習のターゲットにすればよいか分かる。
- J-LEXは、上述のような教材開発や各種サービスにおける「やさしい日本語」作成のための利用のほか、学習者等の作文の語彙・漢字レベルの判定や、試験や研究のためのテキストの語彙・漢字レベルのコントロールに使用することも可能

引用文献 References

- Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics*, 25, 133–150.
- Day, R. & Bamford, J. (2002). Top ten principles for teaching extensive reading. *Reading in a Foreign Language*, 14(2), 137-141.
- Hu, M. H., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403–430.
- Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1), 15–30.
- Matsushita, T. (2012). *In What Order Should Learners Learn Japanese Vocabulary? A Corpus-based Approach*. Unpublished PhD Thesis. Victoria University of Wellington.

- Matsushita, T. (2014). How is the relationship between vocabulary knowledge and reading comprehension? A case of Japanese. AILA World Congress 2014.
- Nation, P. (2007). The four strands. *Innovation in Language Learning and Teaching*, 1(1), 2–13.
- 岡崎 眸(1994)「内容重視の日本語教育 —大学の場合—」『東京外国語大学論集』49, 227-244.
- 岡崎 眸(2002)「内容重視の日本語教育 —多言語多文化共生社会における日本語教育の視点から—」岡崎眸(編)科学研究費補助金研究成果報告書『内省モデルに基づく日本語教育実習理論の構築』322-339【一部加筆修正】, URL:
<http://www.dc.ocha.ac.jp/comparative-cultures/jle/Okazaki/naiyoo-juushi.html>, 最終確認日:2015年2月15日
- 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, 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」

URL: <http://www17408ui.sakura.ne.jp/index.html> 最終確認日: 2015年2月15日

ハドソン遠藤(2011)「CBI コースとしての短編講読講座」18th

Princeton Japanese Pedagogy Forum, URL:

[http://www.princeton.edu/pjpf/past/18th-](http://www.princeton.edu/pjpf/past/18th-pjpf/05_EndoPJPF2011.pdf)

[pjpf/05_EndoPJPF2011.pdf](http://www.princeton.edu/pjpf/past/18th-pjpf/05_EndoPJPF2011.pdf) 最終確認日: 2015年2月15日

松下達彦(2011)「日本語を読むための語彙データベース」URL:

<http://www.geocities.jp/tatsum2003/> 最終確認日: 2015年2月15日

松下達彦(2012)「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, 310.

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/> 最終確認日: 2015年2月15日