文章中の語彙の初出課を判定するシステム

A system for identifying lesson numbers for words and phrases based on first appearance in textbooks

中野 光+・北村 達也++
NAKANO Hikaru・KITAMURA Tatsuya
甲南大学学部生+　甲南大学++
Undergraduate student, Konan University+　Konan University++
〒658-8501　神戸市東灘区岡本 8-9-1
nakano.saga@gmail.com

Abstract: We developed a web system for assisting Japanese language teachers in selecting example sentences according to the progression of lessons. This system is based on the "*Minna no Nihongo*" series of 3A Corporation and identifies the lesson numbers associated with the words and phrases as they first appear in the textbooks. When a user inputs sentences into a textbox and selects an arbitrary lesson number in the web browser, currently learned and unlearned words, based on lesson progression, are displayed in different colors. By using the proposed system, teachers can significantly save time and effort preparing educational materials.
キーワード：「みんなの日本語」、初出課、Web ブラウザ、連語

１．はじめに

教師が教材を作成する際には、対象とする学習者のレベル等に応じて文を書き換え、難易度を調節する必要がある。また、非日本語母語話者に対する広報の際には、平易な文に書き換える配慮が求められる。このような書き換え作業を支援するシステムとして、語彙チェッカー (川村, 1998)、J-LEX (松下, 2014)、これやさしいか (伊藤ら, 2014)、かぶとエディタ (北村ら, 2015) などのシステムが開発されている。これらは、入力文章中の単語のレベルを判定するシステムである。

これらのシステムは大変有用であるが、連語に対応していない点に課題が残る。また、単純に授業の進捗に対する単語や文型の既習、未習が分かれば十分という意見もある。そこで、本研究では、(株)スリーエーネットワークの「みんなの日本語初級 I」および「みんなの日本語初級 II」を用いる教師を支援することを目的として、これに準拠したレベル判定システムを開発した。このシステムはこれらの教科書に現れる語彙や文型を抽出し、利用者が指定した課よりも前に現れるか (既習語)、 後に現れるか (未習語) を判定する。このシステムはインターネットを介して利用することができる。

２．システムの機能と仕組み

2.1　概要

本システムはインターネットを経由して Web ブラウザ上で利用する。Web ブラウザの左半分と右半分がそれぞれ入力エリア、出力エリアとなっている。入力エリアにはテキストボックス、教科書内の課を指定するプルダウンメニュー、“クリア”ボタン、“チェック”ボタンがある。

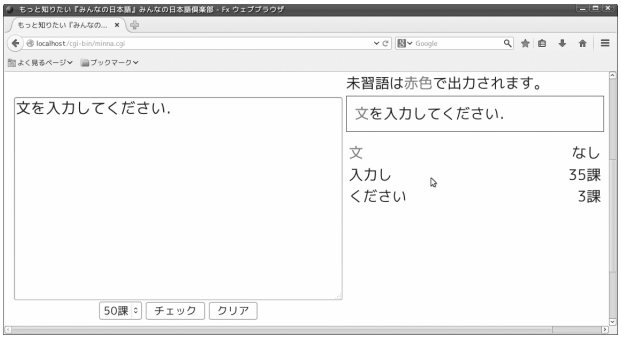

図 1 本システムのスクリーンショット

利用者がこのテキストボックスに文章を入力し、プルダウンメニューで課を指定した後、“チェック”ボタンをクリックすると、「みんなの日本語」における既習語と未習語の判定結果を出力エリアに表示する。ここで、既習語とは利用者が指定した課よりも前の課に現れた語彙を指し、未習語とは利用者が指定した課よりも後の課に現れた語彙および「みんなの日本語」に現れない語彙を指す。出力エリアでは、入力文章の既習語、記号、数字などを黒文字、未習語は赤文字にてボックス内に表示する。加えて、それぞれの語彙が教科書内で初めて現れた課 (初出課) をリスト表示する。なお、「みんなの日本語」に現れない語彙の初出課は「なし」と表示する。

2.2　語彙リスト

本システムでは、「みんなの日本語初級 I 第 2 版」および「みんなの日本語初級 II 第 2 版」の 1 課から 50 課までで取り扱う語彙 (文型を含む) とその初出課を記録した語彙リストに基づいて既習語と未習語を判定する。この語彙リストは CSV 形式のファイルとして保存されている。1 列目には初出課、2 列目には語彙、3 列目以降には品詞が記録されている。この語彙リストは、教科書のデータに基づき、時制の異なる表現や異表記を追加したものである。

この語彙リストにおいて、複数の形態素から成る連語は形態素間にスラッシュを挿入して記録されている。例えば、「どうぞ/よろしく/お願い/し/ます」のように記録されている。このように形態素分割の情報を記録しておくことによって、高速な照合を可能にしている。

なお、現段階では「こちらは〜さんです」のように途中に挿入される語彙が変わりうる文型や、「〜か〜」のように形態素解析だけでは用法が特定できない文型は判定の対象外としており、語彙リストから省いている。

2.3　既習語・未習語判定の仕組み

入力文章中の既習語と未習語を判定し出力エリアに表示するまでには以下の手順を踏む。

(1)　入力された文章の形態素解析
(2)　語彙リストの読み込み
(3)　各形態素と語彙リスト内の語彙との照合
(4)　指定された課に応じて既習語と未習語の色分け表示
(5)　単語とその初出課のリスト表示

形態素解析器は MeCab (工藤, 2006) を用いた。

語彙チェッカーを始めとする従来の単語レベル判定システムでは、形態素解析で分割されてしまう連語を取り扱うことができなかったが、本システムでは複数の形態素にまたがる照合を実装することによって、これを取り扱えるようにした。

また、語彙に複数の用法がある場合には、語彙リストの品詞も利用して照合する。例えば、「家」の場合、建物としての「家 (名詞, 一般)」と「作曲家」などの「〜家 (名詞, 接尾)」を判別することができる。

3．実行例

10 課および 25 課に設定したときの本システムの実行例をそれぞれ図 2、3 に示す。モノクロのため分かりにくいが、課の違いにより既習語と未習語に違いが見られる。

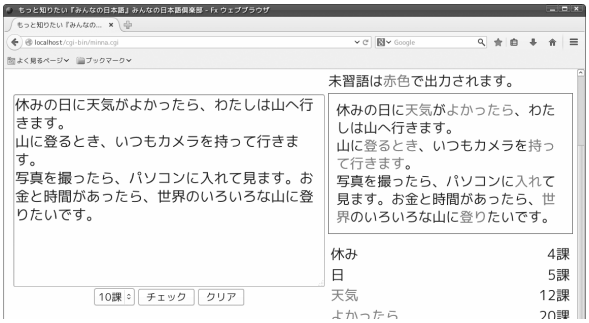

図 2 レベルを 10 課に設定したときの実行例

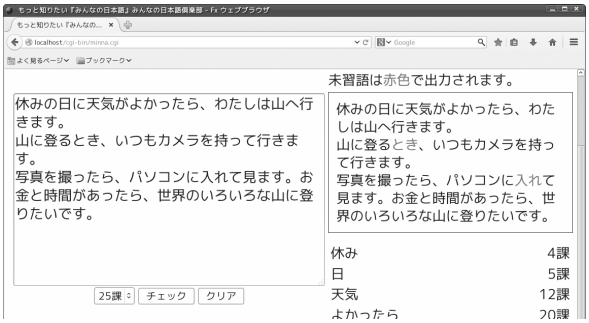

図 3 レベルを 25 課に設定したときの実行例

4．おわりに

「みんなの日本語」に基づいて入力文章中の既習語と未習語を判定するシステムを開発した。本システムは近い将来公開される予定である。本システムを利用することによって、「みんなの日本語」に準拠した副教材やテストを作成する作業が効率化できる。公開後、利用者のフィードバックを得て、改良を加えていく予定である。

謝辞　本研究は、(株)スリーエーネットワークの協力を得て行われた。本研究の一部は平成 26 年度科学研究費 (24320096) 及び私立大学等経常費補助金特別補助「大学間連携等による共同研究」の支援を得て行われた。また，甲南大学知能情報学部 田中豪君の協力を得た。

参考文献

川村よし子 (1998)「読解のためのレベル判定システムの構築：語彙チェッカーの開発と活用」『日本語教育方法研究会誌』, 5(2), 10-11.

松下達彦 (2014)「オンライン日本語テキスト語彙分析器 J-LEX」『日本語教育方法研究会誌』, 21(1), 8-9.

伊藤恵・伊藤美紀・藤田篤・木塚あゆみ・大塚裕子 (2014)「日本語支援者支援システムの構築：日本語教員養成を題材として」『情報処理学会研究報告』, 2014-CF-125(5), 1-6

北村達也・住田真一・孝橋一希 (2015)「文難易度の調整を支援するシステム「かぶとエディタ」」『日本語教育方法研究会誌』.

工藤拓 (2006), MeCab, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html