# 日本語の単語難易度や複雑度を利用した講義の難易度指標作成

木藤 善信 [1]　　木村 祐介 [2]　　椎名 広光 [2]

[1,2] 岡山理科大学大学院 総合情報研究科 情報科学専攻

[3] 岡山理科大学 総合情報学部 情報科学科

sigel4280@live.com[1], CmbdlDpg2f2@gmail.com[2], shiina@mis.ous.ac.jp[3]

## 1　はじめに

日本語を母語としない留学生にとっては，日本語がある程度上達しなければ日本人向けの講義を受講することは容易ではなく理解が困難である．一方，教員自身もスライドや発話が留学生からみてどれ程の難易度であるかを理解していない．双方にとって講義での日本語単語の使用レベルを把握することは有用であると考えられる．学生にとっては講義の難易度や理解しなければならない単語を把握することができる．教員にとっては，作成したスライドの単語難易度の視点から見た問題点の把握や改善点，発話の状況を把握でき，スライドの修正や講義の改善なると考えられる．本研究では，講義の発話や資料がディジタル化されて保存されているインターネット環境を利用する VOD 講義を使用して，日本語難易度を提供するシステムを提案する．

開発したシステムでは，VOD 講義の発話を字幕に変換したデータと提示されているスライドの PowerPoint(以下 PPT) を入力とし，VOD 講義で使用されている日本語単語の難易度評価，講師の発話内容の係り受け構造による発話分析，日本語の助詞の連接や出現数による複雑度による評価を行っている．

## 2　調査対象とする VOD 講義について

本研究では岡山理科大学を VOD による e-Learning のシステム (図 1)[1] を利用している．特に本研究では，講義名「データベース」,「インターネット入門」,「アルゴリズム入門」の PPT 情報とその発話内容を用いて分析を行った．また，講義画面のうちスライドの情報にあたる PPT と発話情報の字幕のテキストを解析して 3 つの指標を作成している (図 2)．


図 1: VOD 講義システム

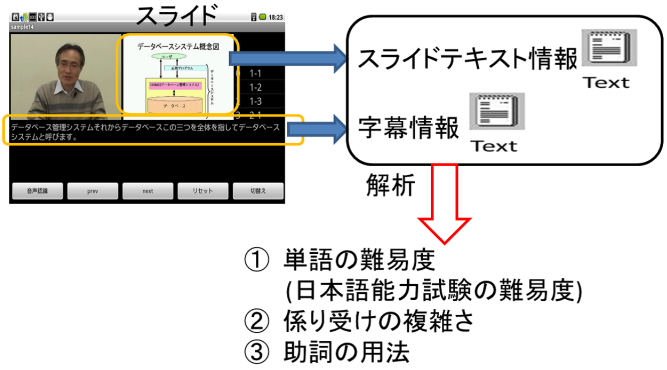


図 2: システム構成

## 3　VOD 講義に対する日本語難易度評価

日本語の難易度評価は PPT の頁単位で PPT と発話内容に対して行う．日本語の難易度の評価には PPT とその発話内容に含まれる名詞の日本語能力試験級の値を用いる．名詞に級値を割り当てるために VOD 講義中の PPT と発話内容に対して形態素解析 [2] を行う．PPT と発話内容より抽出された名詞に対して徳弘 [3] から対応する日本語能力試験 [4] 旧試験のグレード (以下級) の値を割り当てる．徳弘に該当する級が付与されていない名詞については SVM[5, 6] による日本語単語の難易度の推定 [7] より名詞に難易度の級値

表 1: 講義ごとの単語難易度評価

| 種別 | データベース | インターネット入門 | アルゴリズム入門 |
|---|---|---|---|
| PPT | 2.25 | 2.27 | 2.77 |
| 発話 | 2.63 | 2.78 | 2.53 |

表 2: 講義ごとの単語難易度評価

| PPT | データベース | | インターネット入門 | | アルゴリズム入門 | |
|---|---|---|---|---|---|---|
| の頁 | PPT | 発話 | PPT | 発話 | PPT | 発話 |
| 1 | - | 2.75 | 2.00 | 2.89 | 2.25 | - |
| 2 | 2.00 | 2.58 | 2.30 | 2.42 | 2.62 | 2.63 |
| 3 | 2.29 | 2.38 | 1.75 | 2.50 | 3.33 | 2.70 |
| 4 | 1.80 | 2.74 | 1.50 | 2.57 | 3.14 | 1.40 |
| 5 | 2.13 | 2.42 | - | 2.61 | 2.75 | 2.74 |
| 6 | 2.14 | 2.29 | 2.00 | 2.37 | 2.40 | 2.78 |
| 7 | 1.80 | 3.00 | 2.38 | 2.55 | 3.50 | 2.67 |
| 8 | 2.40 | 2.52 | E | E | - | 2.84 |
| 9 | 1.80 | 2.33 | | | 2.17 | 2.50 |
| 10 | 2.00 | 2.30 | | | E | E |
| 11 | 1.80 | 1.86 | | | | |

E:講義終了

を割り当てる．

難易度の評価は講義別に PPT と発話内容に対して行う．評価の対象の VOD 講義は講義「データベース」の第 1 回目のセクション 1，講義「インターネット入門」，講義「アルゴリズム入門」はそれぞれ第 2 回目のセクション 1 とする．表 1 には講義ごとの難易度，表 2 には講義別の PPT の頁単位での PPT と発話内容の難易度をそれぞれ示す．表の数値は 1〜4 の範囲で表され，1 に近いほど日本語能力試験の 1 級の割合が多く，4 に近いほど 4 級の割合が多くなっている．

表 1 の PPT の難易度と発話の何度を比較すると，どちらもほぼ同じかまたは PPT の値のほうが高くなっている．これは書き言葉である PPT より話し言葉である発話内容のほうが難易度の低い名詞が多いことを示している．また表 2 からは PPT の難易度が高いところでは発話内容の難易度も比較的高くなっていることから，PPT の難易度の上下と発話内容の上下は正の相関を持っているのではないかと考えられる．

## 4 講師の発話内容の係り受け構造による発話分析

講師の発話内容の特徴，傾向，複雑さを分析するために，単位時間ごとの係り受け段数の状況を調査する．講師の発話内容に対して CaboCha[8] を用いて係り受け段数を文単位で求める．ここでは一文の係り受け段数値として，一文ごとの係り受けの最大値を用いることにしている．図 3 の例文では点線部で示した「今日から / 行います。」,「ビデオオンデマンドで / 行います。」における係り受けは 1 段，実線部にした「データベースの / 講義を / 行います。」における係り受けは 2 段となっており，実線部の係り受け段数を用いて数値を計算する．一分間の文章数に対して係り受け段数値の相加平均を求め，同 3 講義について図に適用した例を示す．また，同区間中の発話の名詞について徳弘 [3] より級値を付与し，区間内の名詞数に対する相加平均を求める．この値を 3 講義に対して算出し，係り受け段数の平均値と級値の平均値を講義毎に図 4 から図 6 のグラフに示す．

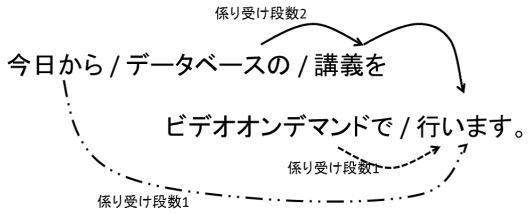


図 3: 係り受け段数値の例文

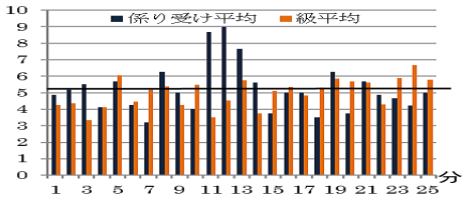


図 4: 講義「データベース」の係り受け段数と級値の平均

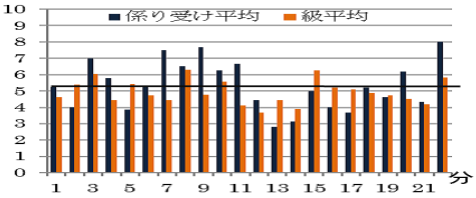


図 5: 講義「インターネット入門」の係り受け段数と級値の平均

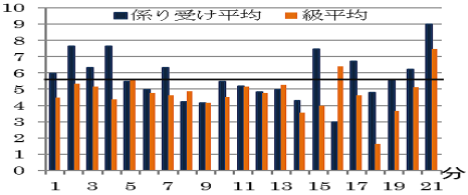


図 6: 講義「アルゴリズム入門」の係り受け段数と級値の平均 (分単位)

講義「データベース」では全体を通して 2 種類の平均値は一定の値を取っているのに対して講義「インターネット入門」，講義「アルゴリズム入門」に関しては値の上下が激しくなっている．このことから講義によって発話の複雑さが一定であるものと，全体を通して大きく変化するものがあるのではないかと考えられる．また毎分の係り受け平均の平均値を示したものが各グラフの中心付近の横線である．平均値より値が高い個所では複雑な発話がみられること，また具体例を挙げる等の簡易な発話であるため平均値より値が低くなる個所があるというようなことがあった．

## 5　助詞の用法に関する構造分析

前節では発話内容の文節間の係り受け構造により文章の構造評価を行った．それに対して， 本稿では助詞の接続や使用数に関する用法の関係を 4 種類のカテゴリーによる講師の発話内容の分析について述べる．

発話内容中の助詞の係り受け状況と種類により文単位で文章の構造的難易度の数値化を行う． $n$ 個の文章を持つ文章集合を $S_n$=$\{s_1,s_2,\ldots,s_k,\ldots,s_n\}$ とし，任意の文章 $s_k$ に対する難易度の条件集合を $W$=$\{w_1,w_2,w_3,w_4\}$ とする ．

(カテゴリ 1:格助詞の連接数) 任意の文章 $s_k$ に対して，格助詞が連接している文節集合の数を $p_k$ とし，それらの文節集合における助詞の連接数の集合を $C$= { $c_1$, $c_2,\ldots,c_{p_k}$} とし，$w_1$ を次式で定義する．

$$w_1 = \sum_{l=1}^{p_k} c_l.$$

携帯電話は / それぞれ(に) / 電話番号(を) / 持っていますよね。

図 7: 格助詞の連接

図 7 の例文では〇で囲まれた 2ヶ所で格助詞が連接しており，ここでの連接数は 1 とする．

(カテゴリ 2:一文節中の助詞の出現回数) 任意の文章 $s_k$ に対して，助詞が複数回出現する文節の数を $q_k$ とし，それらの文節の助詞の出現回数の集合を $D$ = $\{d_1,d_2,\ldots,d_{q_k}\}$ とし，$w_2$ を次式で定義する．

$$w_2 = \sum_{l=1}^{p_k} d_l.$$

携帯電話は / それぞれに / 電話番号を / 持って(て)います(よ)(ね)。

図 8: 助詞の出現回数

図 8 の例文では最後の文節の〇で囲まれた 3 か所で助詞が出現しており，ここでの一文節中の助詞の出現回数を 2 とする．

(カテゴリ 3:単独で出現する格助詞の出現数) $w_1$ での連接する格助詞数に対し，単独での格助詞の出現数を示す．任意の文章 $s_k$ に対して連接していない格助詞の出現数を $E_k$ とし，$w_3$ を次式で定義する．

$$w_3 = E_k.$$

(カテゴリ 4:接続助詞と並立助詞の出現数) 任意の文章 $s_k$ に対して，接続助詞と並立助詞の出現数を $F_k$ とし，$w_4$ を次式で定義する．

$$w_4 = F_k.$$

以上の 4 つのカテゴリーに $\alpha$,$\beta$,$\delta$,$\gamma$ の重みをつけて，任意の文章 $s_k$ における構造難易度の数値 $t_k$ を次のように定義する．

$$t_k = \alpha w_1 + \beta \cdot w_2 + \gamma \cdot w_3 + \delta \cdot w_4.$$

VOD 講義 “ データベース ” の第 1 回目のセクション 1 の発話内容を対象に， 日本人学生に対して文章の難易度アンケートを 5 段階評価で実施し，各文章における学生数に対するアンケート値の相加平均の集合を $R$ = { $r_1$,$r_2,\ldots,r_k,\ldots,r_n$ } とする．各重みの総和を 1.00，刻み幅を 0.01 として，$\sum_i^n (r_i - t_i)^2$ が最小となる時の各重みの値を最適値として難易度の数値化を行う．各重みの最適値は $\alpha = 0.29$，$\beta = 0.19$，$\gamma = 0.30$，$\delta = 0.22$ と計算された．

重みの最適値の妥当性を示すために，各重みの遷移に対するアンケートとの差を用いてグラフを作成する．図 9 のグラフは 2 つの重みを固定し，残りの 2 つの重みを変化させた際のアンケートとの差を表示している．グラフの形状は突出した値がみられず，滑らかな形状を取っている．またこの他にも固定する重みを変化させグラフの作成を行ったが，図 9 と同様の形状を

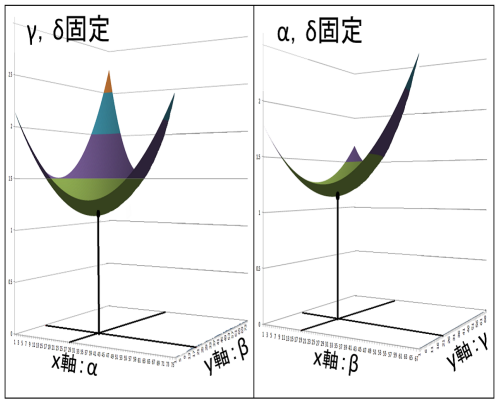


図 9: カテゴリーの重み最適値

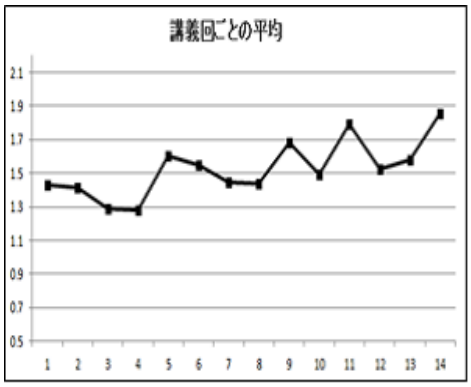


図 10: 講義「データベース」における講義回毎の文章の構造的難易度

成したため，先に算出された各重みの最適値は頑健性を持つと考えられる．

VOD 講義「データベース」について，以上から算出される各講義における文章の構造的難易度に対する平均値を図 10 に示す．

「データベース」講義において，文章の構造的難易度は講義回が後半になるにつれ徐々に難易度が高くなる傾向にある．講義回が進むに毎に講義はより専門的な内容の説明になり，講義の内容と講師の発話の構造的難易度には正の相関があるのではないかと考えられる．

## 6　おわりに

研究では VOD 講義中の講師の発話内容と講義スライドの一致度の指標の提案，講義中に使用されている名詞に対して日本語能力試験級の値を用いた日本語の難易度の評価，講師の発話に対する文節間の係り受け状況による講師の発話の構造的評価，及び講師の発話に対する助詞の係り受け状況による発話の構造的難易度評価を行った．

## 参考文献

[1] 北川，大西: “対面講義と e-learning(LMS + VOD) を併用した講義形式の実践と分析”，日本教育情報学会学会誌 Vol.22 No.3 pp.57-66 (2007)

[2] 形態素解析システム茶筌, http://chasen.naist.jp/hiki.ChaSen/

[3] 徳弘康代，日本語学習のためのよく使う順漢字 2100, 三省堂,2008.

[4] 日本語能力試験公式ウェブサイト, http://www.jlpt.jp

[5] V. Vapnik, Statistical Learning Theory, Springer, 1998.

[6] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[7] K.Nakanishi, N.Kobayashi, H.Shiina, F.Kitagawa Estimating word difficulty using semantic descriptions in dictionaries and Web data, 2012 IIAI International Conference on Advanced Applied Informatics, pp324-329,2012.

[8] CaboCha/南瓜:Yet Another Japanese Dependency Structure Analyzer, http://code.google.com/p/cabocha/