教科書コーパスを利用した難易度別コロケーション辞書の提案

李在鎬（筑波大学）†
佐々木馨（国際交流基金）

Proposal of Collocation Dictionary Based on the Textbook Corpus Analysis

Jae-ho Lee (University of Tsukuba)
Kaori Sasaki (Japan Foundation)

要旨

近年，コミュニケーション能力を重視した言語教育の必要性が指摘されているが，形態素解析などで使用する言語単位（短単位）は，言語教育における単位としては不十分と言わざるを得ない。コミュニケーション能力の育成をはかるためには，実質的な意味機能が担える単位が必要であり，また，学習者の習熟度に応じた網羅的な表現のリストが必要であるが，こうしたリストは存在しない。そこで，本研究では，日本語リーダビリティシステムの構築のために利用した「レベル別コーパス」（文章の難易度がアノテーションされたコーパス，60 万語規模）をもとに，N-gram データを作成したあと，コロケーション表現を抽出した。抽出の結果として，8,121 項目のリストが完成した。各項目は，「レベル別コーパス」での出現頻度を差異係数で処理し，初級レベルとして 3,903 項目，初中級レベルとして 1,472 項目，中級レベルとして 2,746 項目を抽出した，現在，人手で確認作業をすすめており，来年度の春に公開する予定である。本発表はその中間報告である。

１．研究背景と目的

日本語教育研究においてコーパスを利用する意味は，次のように要約できる。コーパスは，個人単位の言語直感では得られない一般的レベルの言語の使用実態を明らかにできる。そのため，コーパスを利用することで，汎用性のある言語教育コンテンツが作成できる。

コーパスの利用範囲は非常に広く，日頃の教育活動での利用はもちろんのこと，教材開発や辞書開発などの汎用的な教育コンテンツの作成において，重要な資料になり得る（具体的な利用例は李・石川・砂川 2012，中俣 2014，本田(他)編 2014，庵・山内 2015 参照）。

しかし，コーパスは，生の言語使用データであるため，そのままの形では言語教育の場に持ち込めない。とりわけ，語彙や文法表現などの言語的素材が持つ潜在的な難易度に対する配慮が必要である。学習者の理解度や習熟度に応じた難易度の調整がなされてこそ，十分な教育効果が期待できる(李 2011)。こうしたことから，学習者に提示する学習コンテンツに関しては難易度に関する調整が常に必要になる。例えば，「日本語教育語彙表」（http://jhlee.sakura.ne.jp/JEV.html，Sunakawa et al.(2012)）では，均衡コーパスと日本語教材コーパスをもとに 17,920 語の語彙表を作成しているが，それには，日本語教師の主観判定に基づく難易度情報が入っており，すべての単語が初級前半，初級後半，中級前半，中級後半，上級前半，上級後半のいずれかにカテゴリー化されている。

さて，本研究は，「日本語教育語彙表」の拡張として，日本語のコロケーション辞書構築

† jhlee.n@gmail.com

を目的とする。具体的な課題としては，1）日本語教科書コーパスをもとに共起語（機能語，内容語問わず）に関する網羅的調査を行うこと，2）語形に関する網羅的調査を行うことを目的とする。

2．データと方法

日本語学習における学習効果を考えた場合，難易度に関するアノテーションは不可欠と言える。しかし，コロケーション表現の難易度を決めるのは，容易ではない。その一番の理由として，コロケーション表現の難易度は単語の難易度から直接予測することができない。例えば，「歌」と「読む」は，「日本語教育語彙表」で調べるといずれも初級前半の語彙である。しかし，この２つがコロケーションを作り，「歌を読む（一般的には「詠む」と表記する）」となった場合，初級の表現としては明らかに違和感がある。同じことが，「日記」と「つける」は中級前半の単語であるが，「日記をつける」になると，さらに難易度があがる。こうした問題を考えた場合，コロケーション表現そのものに対して，何らかの難易度を付与すべきと考える。しかし，その作業には膨大な労力を要する。

これを踏まえ，本研究では，日本語教科書コーパスをもとに構築した「レベル別コーパス」（Lee et al. 2015 in press）を利用することで作業の効率化をはかった。具体的には，難易度判別に代わるものとして，「レベル別コーパス」での出現頻度をもとに，差異係数を計算し，差異係数の値をもとに難易度を決めるという方法論を使用した。なお，「レベル別コーパス」とは，リーダビリティシステムを構築するためのトレーニングデータであり，日本語の教科書データと BCCWJ を利用して構築したものである。コーパスサイズは，以下のとおりである。

表 1.　「レベル別コーパス」のコーパスサイズ

	初級前半	初級後半	中級前半	中級後半	上級前半	上級後半
異なり語	3,178	2,858	5,156	10,291	6,833	4,712
延べ語	72,691	68,746	87,433	174,953	69,268	122,269

単位：UniDic に基づく短単位

表１における６スケールのレベルイメージは，以下のとおりである。

表 2.　６スケールのレベルイメージ

レベル	レベルイメージ
初級前半	単文を中心とする基礎的日本語表現に関して理解できる．複文や連体修飾構造などの複雑な文構造は理解できない。
初級後半	基本的な語彙や文法項目について理解できる．テ形による基本的な複文なども理解できる。
中級前半	比較的平易な文章に対する理解力があり，ある程度まとまった文章でも内容が把握できる。
中級後半	やや専門的な文章でも大まかな内容理解ができ，日常生活レベルの文章理解においてはほぼ不自由がなく遂行できる。
上級前半	専門的な文章に関してもほぼ理解できる．文芸作品などに見られる複雑な構造についても理解できる。
上級後半	高度に専門的な文章に関しても不自由なく，理解できる．日本語のあらゆるテキストに対して困難を感じない。

本研究が目指すコロケーション表現の抽出も，最終的には表 2 のレベルイメージに準拠することを目指すが，現時点では，初級，初中級，中級の 3 レベルのものとして整理している。

さて，本研究では，とりわけニーズが高いと思われる初級と中級レベルのコロケーション辞書を作成する目的で，表 1 の初級前半～中級後半のデータを利用し，N-gram によるコロケーション表現の抽出を試みた。具体的には，以下の手順で作業を行った。

- ステップ 1.「レベル別コーパ」スの中から初級前半～中級後半のデータを MeCab 0.996 ＋UniDic 2.2.0.1 で解析する。
- ステップ 2. 形態素解析済みデータに対して 3gram～6gram の連結データを作成する。
- ステップ 3. 連結データを集計し，サブコーパス別および合計出現頻度を計算する。
- ステップ 4. 合計出現頻度 5 以上のものを絞り込む
- ステップ 5. サブコーパスによる差異係数を計算し，レベルを決める。

３．結果

ステップ 1 の結果，403,823 語のデータが得られた。ステップ 2 の結果，75,668 項目のデータが得られた。ステップ 3・4 の結果，8,121 項目のデータが得られた。見出し語の例と見出し語の数を表 3 に示す。

表 3. N-gram による見出し語の数と実例

	見出し語数	見出し語例
3gram	4994	ています/ありません/と思います/ても良い/た事が/になった
4gram	2117	というのは/しています/かもしれない/がありますか/ことができます
5gram	752	てしまったんです/ことが分かりました/だと思いますか
6gram	258	と言われています/といっていました/はどこにありますか
総計	**8121**	

３つの短単位で構成された 3gram の見出し語は，4994 項目が得られた。具体例としては，「～ています」などの初級の学習項目に相当するものが多い。次に，４つの短単位で構成された 4gram の見出し語は，2117 項目，5gram の見出し語は 752 項目，6gram の見出し語は 258 項目が得られた。7gram 以上のデータも作成してみたものの，コーパスサイズが小さいこともあって，頻度 5 以上のものは少ない上に，表現として不完全なものが多いため，対象から外した。

次に，得られた見出し語の特徴分析のため，品詞単位で調べてみた。表 4 に 3gram から 6gram で高頻度パターン上位５位を報告する。

表 4. 品詞の組み合わせの高頻度パターン

	品詞の組み合わせ	具体例
3gram	［助詞-格助詞/名詞-普通名詞-一般/助詞-格助詞］	の方が
3gram	［助詞-格助詞/動詞-一般/助詞-接続助詞］	によって

3gram	[動詞-一般/助詞-接続助詞/動詞-非自立可能]	思っている
3gram	[助詞-格助詞/動詞-一般/助動詞]	と思います
3gram	[名詞-普通名詞-一般/助詞-格助詞/動詞-一般]	事が分かる
4gram	[助詞-格助詞/動詞-一般/助詞-接続助詞/動詞-非自立可能]	と思っている
4gram	[動詞-一般/助詞-接続助詞/動詞-非自立可能/助動詞]	思っています
4gram	[名詞-普通名詞-一般/助詞-格助詞/動詞-一般/助詞-接続助詞]	文章を読んで
4gram	[動詞-非自立可能/助詞-接続助詞/動詞-非自立可能/助動詞]	しています
4gram	[助詞-格助詞/動詞-非自立可能/助詞-接続助詞/動詞-非自立可能]	をしている
5gram	[助詞-格助詞/動詞-一般/助詞-接続助詞/動詞-非自立可能/助動詞]	と思っています
5gram	[助詞-格助詞/動詞-非自立可能/助詞-接続助詞/動詞-非自立可能/助動詞]	をしています
5gram	[助詞-接続助詞/動詞-非自立可能/助動詞/助詞-準体助詞/助動詞]	ていたのだ
5gram	[動詞-一般/助詞-接続助詞/動詞-非自立可能/助動詞/助動詞]	言っていました
5gram	[動詞-非自立可能/助詞-接続助詞/動詞-非自立可能/助動詞/助動詞]	していました
6gram	[助詞-格助詞/動詞-一般/助詞-接続助詞/動詞-非自立可能/助動詞/助動詞]	といっていました
6gram	[名詞-普通名詞-一般/助詞-格助詞/動詞-一般/助詞-接続助詞/動詞-非自立可能/助動詞]	会社に勤めています
6gram	[助詞-格助詞/代名詞/助詞-格助詞/動詞-非自立可能/助動詞/助詞-終助詞]	に何がありますか
6gram	[助動詞/助詞-格助詞/動詞-一般/助詞-接続助詞/動詞-非自立可能/助動詞]	たいと思っています
6gram	[助動詞/助詞-格助詞/動詞-一般/助動詞/助詞-接続助詞/動詞-非自立可能]	だと言われている

次に，難易度判別のために，初級教科書での使用頻度と中級教科書での使用頻度をもとに差異係数を使用し，どちらの（レベルの）教科書でよりたくさん使用されているかを調べた。差異係数がマイナス値のものを初級，差異係数が 0～0.49 のものは初中級，0.50～1.0 のものを中級とし，集計してみた。

表 5. Ngram×レベルのクラス集計表

	初級レベル	初中級レベル	中級レベル
3gram	2156	991	1847
4gram	1084	352	681
5gram	467	99	186
6gram	196	30	32
総計	**3903**	**1472**	**2746**

以上の方法で，完成したデータは，以下の通りである。

■ 初級レベルのコロケーション

Gram	語彙素(基本形)	発音(出現形)	品詞	難易度	中級合計	初級合計	差異計数
3	て居ります	(ています)	[助詞-接続助詞/動詞-非自立可能/助動詞]	初級	373	363	0.013586957
3	有りますず	(ありません)	[動詞-非自立可能/助動詞/助動詞]	初級	135	180	-0.142857143
3	が有ります	(があります)	[助詞-格助詞/動詞-非自立可能/助動詞]	初級	118	195	-0.29600639
3	たのです	(たんです)	[助動詞/助詞-準体助詞/助動詞]	初級	80	160	-0.333333333
3	と思います	(とおもいます)	[助詞-格助詞/動詞-一般/助動詞]	初級	117	122	-0.020920502
3	成りますた	(なりました)	[動詞-非自立可能/助動詞/助動詞]	初級	105	118	-0.058295964
3	ますたか	(ましたか)	[助動詞/助動詞/助詞-終助詞]	初級	7	198	-0.931707317
3	のですか	(んですか)	[助詞-準体助詞/助動詞/助詞-終助詞]	初級	59	143	-0.415841584
4	為るて居ります	(しています)	[動詞-非自立可能/助詞-接続助詞/動詞-非自立可能/助動詞]	初級	102	98	0.02
3	ますずか	(ませんか)	[助動詞/助動詞/助詞-終助詞]	初級	31	158	-0.671957672
3	と思うて	(とおもって)	[助詞-格助詞/動詞-一般/助詞-接続助詞]	初級	92	81	0.063583815
3	済みますず	(すみません)	[動詞-一般/助動詞/助動詞]	初級	16	151	-0.808383234
3	ても良い	(てもいー)	[助詞-接続助詞/助詞-係助詞/形容詞-非自立可能]	初級	31	136	-0.628742515

■ 初中級レベルのコロケーション

Gram	語彙素(基本形)	発音(出現形)	品詞	難易度	中級合計	初級合計	差異計数
3	為るて居る	(してい)	[動詞-非自立可能/助詞-接続助詞/動詞-非自立可能]	初中級	239	138	0.267904509
3	を為るて	(おして)	[助詞-格助詞/動詞-非自立可能/助詞-接続助詞]	初中級	209	123	0.259036145
3	為るますた	(しました)	[動詞-非自立可能/助動詞/助動詞]	初中級	168	121	0.162629758
3	居るますた	(いました)	[動詞-非自立可能/助動詞/助動詞]	初中級	148	76	0.321428571
3	の中に	(のなかに)	[助詞-格助詞/名詞-普通名詞-副詞可能/助詞-格助詞]	初中級	134	56	0.416526316
3	て居ります	(ていまし)	[助詞-接続助詞/動詞-非自立可能/助動詞]	初中級	119	52	0.391812865
4	て居るますた	(ていました)	[助詞-接続助詞/動詞-非自立可能/助動詞/助動詞]	初中級	119	52	0.391812865
3	につくて	(について)	[助詞-格助詞/動詞-一般/助詞-接続助詞]	初中級	121	45	0.457831325
3	かも知れる	(かもしれ)	[助詞-副助詞/助詞-係助詞/動詞-一般]	初中級	88	55	0.230769231
3	を持つて	(おもって)	[助詞-格助詞/動詞-一般/助詞-接続助詞]	初中級	81	49	0.246153846
3	て仕舞うた	(てしまった)	[助詞-接続助詞/動詞-非自立可能/助動詞]	初中級	82	42	0.322580645
3	て居らない	(ていない)	[助詞-接続助詞/動詞-非自立可能/助動詞]	初中級	68	49	0.162393162
3	だのだ	(なので)	[助動詞/助詞-準体助詞/助動詞]	初中級	81	36	0.384615385

■ 中級レベルのコロケーション

Gram	語彙素(基本形)	発音(出現形)	品詞	難易度	中級合計	初級合計	差異計数
3	て居るた	(ていた)	[助詞-接続助詞/動詞-非自立可能/助動詞]	中級	374	55	0.743509744
3	為るて居る	(している)	[動詞-非自立可能/助詞-接続助詞/動詞-非自立可能]	中級	309	75	0.608375
3	と為るて	(として)	[助詞-格助詞/動詞-非自立可能/助詞-接続助詞]	中級	297	12	0.922330097
3	て来るた	(てきた)	[助詞-接続助詞/動詞-非自立可能/助動詞]	中級	219	43	0.671755725
3	て居るの	(ているの)	[助詞-接続助詞/動詞-非自立可能/助詞-準体助詞]	中級	215	45	0.653046154
3	たのだ	(たので)	[助動詞/助詞-準体助詞/助動詞]	中級	190	51	0.576763485
3	れるて居る	(れている)	[助動詞/助詞-接続助詞/動詞-非自立可能]	中級	212	26	0.781512605
3	と言う事	(とゆーこと)	[助詞-格助詞/動詞-一般/名詞-普通名詞-一般]	中級	206	16	0.855855856
3	に成るて	(になって)	[助詞-格助詞/動詞-非自立可能/助詞-接続助詞]	中級	164	35	0.649241206
3	に困るて	(によって)	[助詞-格助詞/動詞-一般/助詞-接続助詞]	中級	160	12	0.866666667
3	のだ有る	(のである)	[助詞-準体助詞/助動詞/動詞-非自立可能]	中級	172	2	0.977011494
3	と言うて	(といって)	[助詞-格助詞/動詞-一般/助詞-接続助詞]	中級	137	26	0.689981595
3	為るれるて	(されて)	[動詞-非自立可能/助動詞/助詞-接続助詞]	中級	138	21	0.735849057

４．まとめと今後の課題

本発表では，日本語教科書データを利用したコロケーション辞書作成について紹介した。3gram から 6gram の見出し語として 8,121 項目のリストが構築できた。全体的に機能語に対するリスト化については，ある程度成功しているが，コーパスサイズが小さい問題があり，内容語に対するリストとしてはまだまだ不十分な状態である。今後の予定として，均衡コーパスに対するリーダビリティ値を計算し，「レベル別コーパス」を大きくした上で，内容語も含めたコロケーション辞書の拡張を行いたい。また人手によるチェック作業を継続し，数などを踏まえた上で，初級前半，初級後半，中級前半，中級後半のコロケーション表現のリストとして完成させたい。

謝　辞

本研究は、文部科学省科学研究費補助金「読解教育支援を目的とする文章難易度判別システムの開発（課題番号：25370573，代表者：李在鎬）による補助を得ています。

文　献

庵 功雄 ， 山内 博之 (2015)『データに基づく文法シラバス (現場に役立つ日本語教育研究 1)』くろしお出版

中俣尚己 (2014)『日本語教育のための文法コロケーションハンドブック』くろしお出版

本田 弘之，岩田 一成，義永 美央子（2014)『日本語教育学の歩き方―初学者のための研究ガイド』大阪大学出版会

李在鎬（2011）「大規模テストの読解問題作成過程へのコーパス利用の可能性」,『日本語教育』148, pp.84-98.

Lee, Jae-ho &Yoichiro Hasebe(2015 in press)　“Readability Measurement for Japanese Text Based on Leveled Corpora”

李在鎬，石川慎一郎，砂川有里子（2012）　『日本語教育のためのコーパス調査入門』くろしお出版