文字列類似度と意味類似度から見た漢字読み問題選択肢のパターン分類

琉球大学国際教育センター紀要
第 4 号, 2020

文字列類似度と意味類似度から見た漢字読み問題選択肢のパターン分類

新城　直樹

0. はじめに

本稿では，漢字の読みの選択肢問題を自動的に生成するシステム作りの基礎調査と分析を行う。選択肢は 4 選択肢とし，誤答選択肢には非単語は含めない。学習者像は日本語非母語話者を想定し，日本語能力試験の N1 と N2 の模擬試験問題集内の 80 問を基礎データとし，これらの問題群の意味類似度と文字列類似度（選択肢として出されるひらがな文字列）から選択肢の作り方のパターンの洗い出しを目指す。

1. 選択肢の文字列類似度

選択肢（ひらがな文字列）の文字列の類似度には，レーベンシュタイン距離(Levenshtein Distance)とジャロ・ウィンクラー距離(Jaro-Winkler Distance) の平均値を採用する。レーベンシュタイン距離は 2 つの文字列間での挿入，削除，置換の最小操作回数に基づく編集距離であり，ジャロ・ウィンクラー距離は，文字列間で共通する文字を順に抜き出したもの同士での置換回数に基づくジャロ距離（Jaro Distance）を使用して，さらに元の文字列同士で先頭から何文字共通しているかを反映させて算出する。これらは違うアルゴリズムであるが，80 問（4 選択肢）の 480 対の相関係数が 0.874 であり，規模数を大きくする前段階の調査，仮説形成という観点から，両者の平均を採用した。

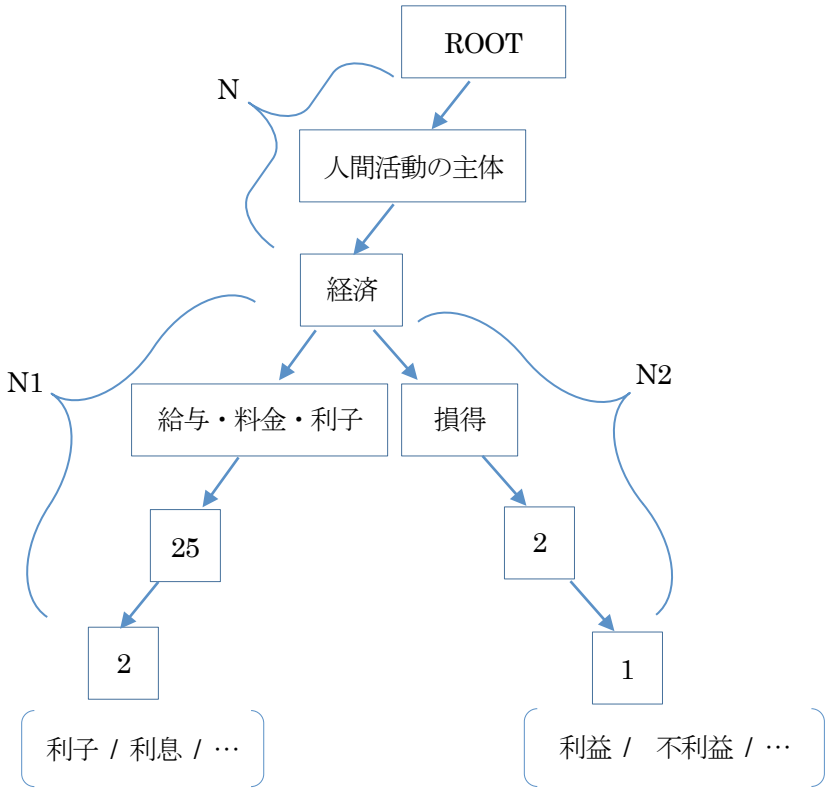
対象とする文字列には送り仮名は含めず，読み部分のみとした。

2. 単語の意味的類似性

選択肢の単語同士の意味類似度は，『分類語彙表』の 5 段階の分類枠で形成されるツリー構造内での距離とした。『分類語彙表』には「体の類」「用の類」「相の類」「その他の類」があるが，今回の基礎データでは選択肢内で体言や用言が混在しているケースは少数であったため，分類として含めなかった。

距離の計算には Wu-Palmer アルゴリズムで計算した。Wu-Palmer アルゴリズムは, 2 つの単語に共通し，かつ最も近い上位分類項目の ROOT からの深さ N と，その共通する分類項目からの各単語への深さ N1 と N2 を用いた計算である（図 1)。

図 1.　『分類語彙表』に基づいたオントロジーと Wu-Palmer アルゴリズム



Wu-Palmer アルゴリズムによる計算は以下の通りである。

$$SimWup = \frac{N * 2}{N1 + N2 + N * 2}$$

例として，「利子」と「利益」の類似度は，共通し最も近い上位分類項目の ROOT からの深さ N は 2，N1 と N2 は 3 となり，「利子 - 利益」の類似度は 0.4 となる。Wu-Palmer アルゴリズムで計算した値は 0～1 の範囲となり，0 が最も類似度が低く，1 が最も高い。

3.　分析

日本語能力試験の模擬試験問題集 8 冊から誤答選択肢に非単語がない 80 問を抽出し，4 選択肢の組み合わせ 6 対ごとに文字列類似度と意味類似度を計算した。意味類似度は，下記のように複数の分類がある場合はそれぞれの組み合わせごとに計算し，その平均を対ごとの類似度とし，最後に 4 選択肢の 6 つの組み合わせの平均を出し，これを問題の文字列類似度，意味類似度とした。

問題：省く
選択肢：はぶく / そむく / のぞく / きずく

はぶく - そむく,
【省く】関係-存在-除去-1-1 --- 【背く】関係-類-相対-2-1 =0.2
【省く】関係-存在-除去-1-1 --- 【背く】活動-交わり-約束-5-1 =0.0
【省く】関係-存在-除去-1-1 --- 【背く】活動-待遇-命令・制約・服従-1-1 =0.0
Average:0.06666666666666667

はぶく - のぞく,
【省く】関係-存在-除去-1-1 --- 【のぞく】関係-存在-出没-2-1 =0.4
【省く】関係-存在-除去-1-1 --- 【除く】関係-存在-除去-1-1 =1.0
【省く】関係-存在-除去-1-1 --- 【のぞく】活動-心-見る-1-1 =0.0
Average:0.4666666666666666
・・・・・・・・・・

また，同音異義語がある場合，同音異義語の組み合わせごとに類似度計算を行い，その平均を結果とした。下記の例では，正答選択肢「そうさく」は「捜索」と「創作」，誤答選択肢「そうさ」は「捜査」と「操作」の組み合わせごとの類似度を出し，その平均を「そうさく」と「そうさ」の意味類似度としている。

問題：捜索
選択肢：, そうさく / そうさ / たんさく / たんさ

そうさく - そうさ,
【捜索】活動-心-研究・試験・調査・検査など-1-5 --- 【捜査】活動-心-研究・試験・調査・検査など-1-1 = 0.8
【捜索】活動-心-研究・試験・調査・検査など-1-5 --- 【操作】活動-事業-扱い・操作・使用-1-1 = 0.2
【創作】活動-芸術-創作・著述-1-3 --- 【捜査】活動-心-研究・試験・調査・検査など-1-1 = 0.2
【創作】活動-芸術-創作・著述-1-3 --- 【操作】活動-事業-扱い・操作・使用-1-1 = 0.2
Average:0.35

そうさく - たんさく,
【捜索】活動-心-研究・試験・調査・検査など-1-5 --- 【探索】活動-心-研究・試験・調査・検査など-2-1 = 0.6
【捜索】活動-心-研究・試験・調査・検査など-1-5 --- 【単作】活動-事業-農業・林業-3-5 = 0.2
【創作】活動-芸術-創作・著述-1-3 --- 【探索】活動-心-研究・試験・調査・検査など-2-1 = 0.2
【創作】活動-芸術-創作・著述-1-3 --- 【単作】活動-事業-農業・林業-3-5 = 0.2
Average:0.3
・・・・・・・・・・・

同音異義語を含めることによって類似度が下がる場合があり，問題作成者は意味が類似した単語を選んだのであろうから，「そうさく（= 捜索） / そうさ（=捜査） / たんさく（= 探索） / たんさ（= 探査）」と固定して類似度を計算するべきという考えもあるであろう。しかし，解答者の視点から考えると，「捜索」の意味も読みも知らないが「創作」の意味と読みは知っている場合，問題文の文脈から「そうさく」を選ばないという判断の根拠となる情報を持つと考えることができ，そのことは「捜索」が他の 3 つの誤答選択肢より「何かを探す」という意味から類似度が下がるという考えを本稿では取る。

前提条件：「捜索」の意味も読みも知らない

	「捜査」「探索」「探査」のいずれかの意味を知っている	「捜査」「探索」「探査」の意味を知らない
「創作」の意味と読みを知っている	「そうさく」を選ばない根拠となる情報を持つ	「そうさく」を選ばない根拠となる情報を持つ
「創作」の意味も読みも知らない	「そうさく」を選ばない根拠となる情報を持つ	「そうさく」を選ぶ/選ばないに関する情報ゼロ

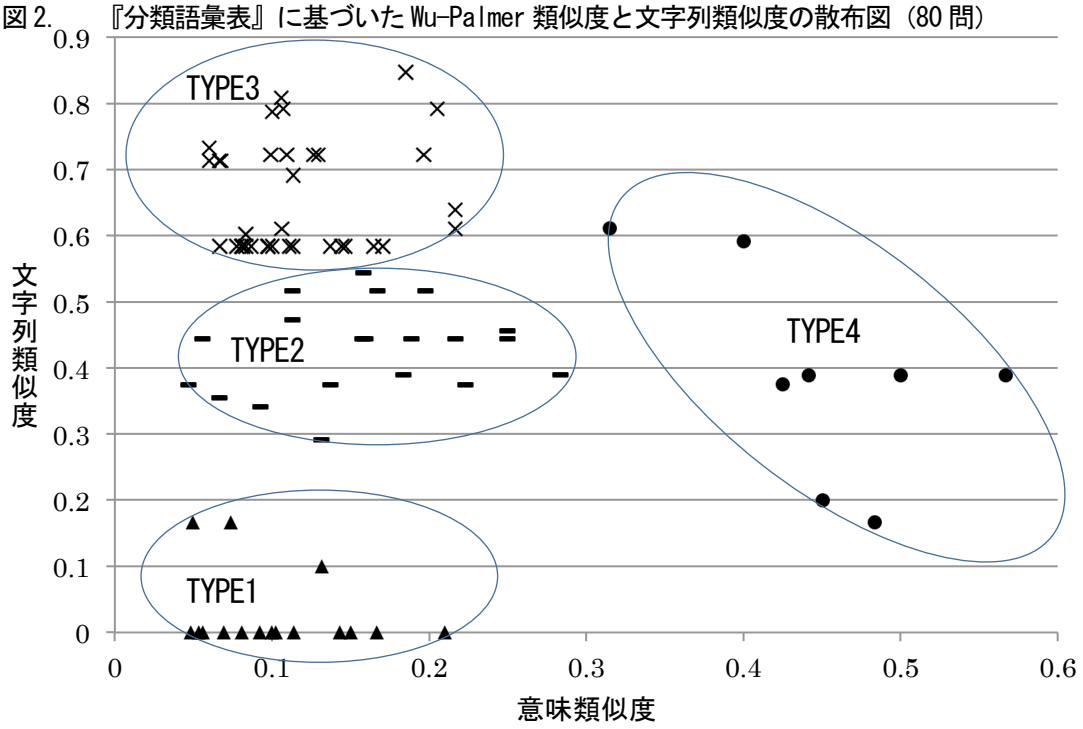
本稿では，問題作成者が類似度の高い単語を選んだかどうかではなく，解答者の知識状態によって問題の難度が上がることも類似度に関わるという立場を取るが，これについては認知心理学の分野において，白砂他（2017）で「Familiarity-matching に従えば，問題文で提示された対象が unfamiliar であった場合，人間は「同じく unfamiliar な選択肢が正答だろう」と推論する．よって，通常であればあまり有用な手がかりにならないと見なされがちな「unfamiliarity」という性質も，正答を導くうえで有用な手がかりとなりうる．(p.337)」という示唆もあり，意味類似度と familiarity/unfamiliarity の関係性，そして，解答者が選択肢問題を解く際のヒューリスティックを併せて見ていくことが重要であると考える。

3.1. 文字列類似度と意味類似度の散布図によるグルーピング

問題ごとの文字列類似度と意味類似度は，4 つの選択肢の 6 つの組み合わせごとに出した類似度の平均とした。

図 2 の散布図において，TYPE1～TYPE4 の 4 つのグルーピングを行った。TYPE1 は文字列類似度が低いもの，TYPE3 は文字列類似度が高いもの，TYPE4 は意味類似度が高く文字列類似度は TYPE1 と TYPE3 の中間であるもの，そして TYPE2 はそれらの中間に位置付けられる。

図 2.　『分類語彙表』に基づいた Wu-Palmer 類似度と文字列類似度の散布図（80 問）



	問題	正答選択肢	誤答選択肢	誤答選択肢	誤答選択肢
TYPE 1	布	ぬの	ふ	ふう	わた
	菓子	かし	はし	はこ	かす
	肩	かた	そで	けん	わき
	旧	きゅう	しん	こ	み
	机	つくえ	かばん	ひきだし	かがみ
	綿	めん	きぬ	きじ	かみ
	件	けん	ど	かい	こ
	裏	うら	おもて	わき	そば
	叫んだ	さけんだ	よんだ	とんだ	かんだ
	肌	はだ	ひふ	かわ	み
	腹	はら	おなか	むね	あたま
	汗	あせ	ひたい	うで	くび
	猫	ねこ	さる	いぬ	ぶた
	兆し	きざし	あかし	しるし	こころざし
	評価	ひょうか	へいか	ひか	かいか
	押した	おした	さした	うつした	かした
	程度	ていど	おんど	いど	かんど

TYPE1 は文字列類似度が低く，意味類似度は必ずしも低いものだけではない。このタイプは「漢字の読みを正確に覚えているかどうかより，4 つの選択肢にある単語をどれだけ多く知っているか」を問うも

のと考えられる。たとえば，意味類似度がかなり低く出ている「菓子：かし / はし / はこ / かす」の場合，問題文の文脈から「食べ物」であることが推測され，かつ，「(お) かし」が既知である場合，「菓子」という漢字の読みを知っているかどうかより，「(お) かし」という語を知っているかどうかで正答率が変わる可能性がある。また，意味類似度が比較的高く出ている「汗：あせ / ひたい / うで / くび」や「猫：ねこ / さる / いぬ / ぶた」では，文脈からの推測は難しくなるものの，その漢字の読みをあいまいながらでも覚えてさえいれば正答できる可能性がある。どちらとも必要条件は「その単語の漢字を知っているかにかかわらず，そもそもその語を知っていること」となる。

	問題	正答選択肢	誤答選択肢	誤答選択肢	誤答選択肢
TYPE 2	担う	になう	すくう	おう	うれう
	狭い	せまい	うるさい	きたない	すばらしい
	強盗	ごうとう	ごうどう	ごとう	きょうどう
	精神	せいしん	せいじん	ぜんしん	しんけい
	欺いて	あざむいて	そむいて	みちびいて	つらぬいて
	伴って	ともなって	したがって	ととのって	とどまって
	愚か	おろか	おおまか	たしか	あきらか
	祝った	いわった	いのった	うらなった	ねがった
	次第（に）	しだい	じてい	してい	じだい
	兆し	きざし	しるし	ためし	あかし
	省いた	はぶいた	そむいた	のぞいた	きずいた
	含む	ふくむ	かこむ	つつむ	たたむ
	望み	のぞみ	たのみ	なやみ	このみ
	漂って	ただよって	さまよって	めぐって	わたって
	著しく	いちじるしく	はげしく	はなはだしく	すばらしく
	袋	ふくろ	はこ	かご	かばん
	鍛えた	きたえた	さかえた	たたえた	ととのえた
	抱えて	かかえて	そなえて	たずさえて	くわえて
	奮闘	ふんとう	しゅうとう	しゅうせん	ふんせん
	福祉	ふくし	ふうしゅう	ふうし	ふくしゅう

TYPE2 は TYPE1 と TYPE3 の中間に位置するタイプである。中間にあることもあってか特徴付けが難しいが，意味類似度が低いものだけではないことから，バランスよくさまざまなバリエーションがある問題群ともいえる。

	問題	正答選択肢	誤答選択肢	誤答選択肢	誤答選択肢
TYPE 3	携わって	たずさわって	かかわって	こだわって	くわわって
	滑らか	なめらか	なだらか	ほがらか	やすらか
	済んで	すんで	やんで	とんで	かんで
	施行	しこう	せっこう	せいこう	しっこう
	悔しい	くやしい	いやしい	むなしい	さびしい
	伴う	ともなう	そこなう	うらなう	つぐなう
	観測	かんさつ	かんそう	かんそく	かんしゃ
	効果	こうか	ごうか	きか	きが
	携わる	たずさわる	くわわる	かかわる	そなわる
	設け	もうけ	しつけ	そむけ	あずけ
	似ている	にている	している	みている	えている
	異なった	ことなった	かさなった	つらなった	たかなった
	焦って	あせって	いらだって	はやまって	はかどって
	掲げて	かかげて	あげて	もたげて	ささげて
	敗れて	やぶれて	たおれて	みだれて	つぶれて
	介護	かいご	かんご	かいじょ	かいほう
	頂上	ちょうじょう	さんちょう	ちじょう	ちょうじょ
	与えられた	あたえられた	くわえられた	こえられた	ささえられた
	概略	がいりゃく	きりゃく	きかく	がいかく
	威力	いりき	いりょく	じりき	じりょく
	軌道	きどう	くどう	くうどう	きゅうどう
	考慮	こうりょ	こうろ	こうろう	こうりょう
	要請	ようせい	ようしょう	ようきゅう	ようぼう
	均衡	きんこう	きんしょう	きっしょう	きっこう
	触れない	ふれない	かれない	されない	とれない
	埋められて	うめられて	とめられて	せめられて	なめられて
	鋭い	するどい	にぶい	えらい	つらい
	滞って	とどこおって	いきづまって	たまって	とまって
	解ける	とける	よける	かける	さける
	簡単	かんたん	かんだん	かたん	かだん
	世代	せだい	せいだい	せいたい	せたい
	貧しかった	まずしかった	かなしかった	さびしかった	きびしかった
	否め	いなめ	とがめ	おさめ	いさめ
	感染	かんせん	かんでん	かんぜん	かんねん
	研修	けんしゅう	けんきゅう	げんしゅう	げんきゅう

TYPE3 は文字列類似度が高く，意味類似度も TYPE1 と同じく必ずしも低いものだけではない。TYPE1 では送り仮名がないものが多いが，1.で述べたように，対象とする文字列には送り仮名は含めず，読み部

分のみとしているため，送り仮名部分の有無やその長さと文字列類似度との関連はない。

文字列類似度が高いことから，正確な読みを覚えているかどうかを問うタイプといえる。「世代：せだい / せいだい / せいたい / せたい」や「効果：こうか / ごうか / きか / きが」では濁音の有無，「施行：しこう / せっこう / せいこう / しっこう」，「均衡：きんこう / きんしょう / きっしょう / きっこう」では促音，「考慮：こうりょ / こうろ / こうろう / こうりょう」では長音や拗音に関する正確さを問うものとなっている。また，「介護：かいご / かんご / かいじょ / かいほう」や「威力：いりき / いりょく / じりき / じりょく」のように 2 字のうち 1 字を知っていれば実質 2 択となり，漢字の読みを知っていることが有利につながることから，「単語としての漢字の読みを知っているかどうかより，漢字単独の読みをどれだけ知っているかどうか」に焦点を当てたタイプともいえる。

	問題	正答選択肢	誤答選択肢	誤答選択肢	誤答選択肢
TYPE 4	幾多	いくた	たた	きた	あまた
	逃れた	のがれた	はなれた	それた	まぬがれた
	遮る	さえぎる	さまたげる	せばめる	へだてる
	捜索	たんさく	そうさ	そうさく	たんさ
	恨む	うらむ	うらやむ	ねたむ	あやしむ
	納める	おさめる	さだめる	きめる	もとめる
	利益	りえき	りし	りそく	りじゅん
	惜しむ	おしむ	あやしむ	かなしむ	なつかしむ

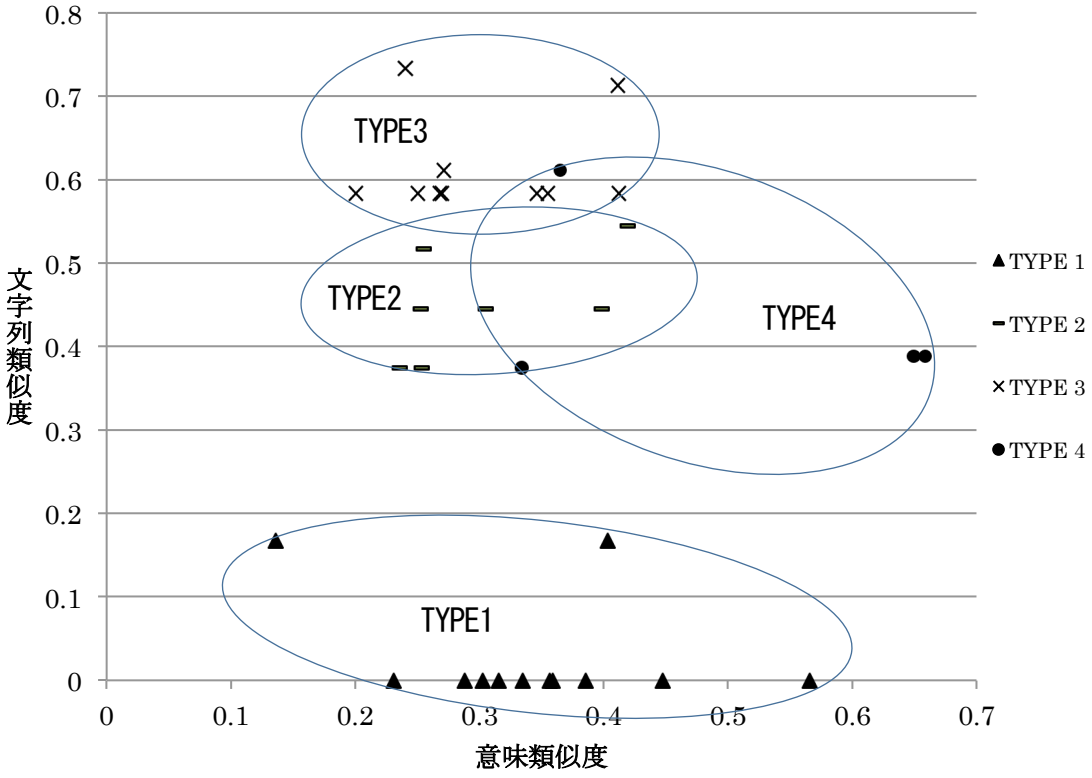
TYPE4 は，意味類似度が高く，文字列類似度は高いものから低いものまである。意味類似度が高いことから，問題文の文脈からの単純な推測だけでは難しく，類義語間の正確な違いの理解が求められるタイプといえる。そして，類義語間の正確な違いが分からなかった場合，文字列類似度が高いものもあるため，その読みを正確に知っているかどうか問われるともいえる。

3.2. 日本語 Wordnet の Wu-Palmer 類似度との関連

類似度を測る手法としては日本語 Wordnet を利用しての研究，調査が自然言語処理の分野を中心にある。ここでは，3.1 で行った『分類語彙表』に基づいた意味類似度を日本語 Wordnet の Wu-Palmer 類似度に変えた場合に散布図上でどのような分布となるか確認した。

本稿の基礎データ内の選択肢にある 4 語すべてが日本語 Wordnet に登録されているものは 80 問中 33 問であった。図 3 はこの 33 問の散布図である。データ数は少ないが，全体的に意味類似度のばらつきが大きくなっていることから，『分類語彙表』と日本語 Wordnet の相関は低いことが推測される。

図 3.　日本語 Wordnet の Wu-Palmer 類似度と文字列類似度の散布図（33 問）



4.　まとめ

漢字の読み問題（4 選択肢）の誤答選択肢を自動的に生成するシステム作成のために，第一段階としてどのような選択肢パターンがあるかを意味類似度と文字列類似度の視点から見てきたが，今回利用した 80 問からは，4 つのタイプに分けられるであろうことが確認された。

また，『分類語彙表』に基づいた Wu-Palmer アルゴリズムの計算結果と日本語 Wordnet での Wu-Palmer 類似度の相関はない，または低いであろうことが推測された。実際，今回の調査の前に，800 単語（319,600 対）で両者の類似度で相関係数を見たところ，相関がない，またはかなり低い相関という結果であった。しかしながら，選択肢の自動生成に日本語 Wordnet を利用できる可能性については検討すべきであるし，日本語 Wordnet と『分類語彙表』の連携についても今後も調査，分析を行っていきたい。

【模擬試験問題集】

国際交流基金（2012）『日本語能力試験 公式問題集 N1』凡人社

国際交流基金（2012）『日本語能力試験 公式問題集 N2』凡人社

千駄ヶ谷日本語教育研究所（2013）『日本語能力試験N1 模擬テスト〈3〉』スリーエーネットワーク

千駄ヶ谷日本語教育研究所（2011）『日本語能力試験N1 模擬テスト〈2〉』スリーe-ネットワーク

新JLPT研究会（2010）『日本語能力試験 模試と対策 N1』アスク出版

新JLPT研究会（2010）『日本語能力試験 模試と対策 N2』アスク出版

ユーキャン日本語能力試験研究会（2010）『U-CANの日本語能力試験N1予想問題集』U-CAN

ユーキャン日本語能力試験研究会（2010）『U-CANの日本語能力試験N予想問題集』U-CAN

参考文献

1）伊藤博美・佐藤 洋之・倉元直樹（2003）「日本語基礎能力テストの特性（1）国語教育から見た語彙理解力測定項目の内容評価」『教育情報学研究』（1）, pp.15-23

2）白砂大・松香敏彦・本田秀仁・植田一博 （2017）「なじみ深さのマッチング：認知プロセスと生態学的合理性の実験的検討」『認知科学』 24（3）, pp.328-343

3）陳 豊・亀山渉（2014）「日本語 WordNet を利用した単語間の類似度計算による画像検索システムに関する研究」,『研究報告オーディオビジュアル複合情報処理（AVM）』,2014（7）, pp.1-5

4）濱田美和（2018）「漢字教材開発のための基礎資料：中・上級日本語学習者の漢字読みテストにおける誤答」『富山大学国際機構紀』1, pp.19-33

（琉球大学 国際教育センター）