現代日本語条件文の条件節の分布にもとづく分類*

瀬戸　義隆

1. はじめに

条件文の機能の特徴的な点は、従属節で条件が提示されるという点にあり、従属節の分類が様々な観点から盛んになされてきた (南 1974、中島 2007、前田 2009 など)。本研究では、Token-based semantic vector space (以下、SVS) を用いて、現代日本語で従属節と主節を接続する条件標識として接続助詞の「バ」を用いる条件文 (以下、バ条件文) の従属節の各用例に生じる語彙の分布情報をもとに、バ条件文の従属節の分類を試みる。第 2 節では、SVS の概要と本研究での分析手法、第 3 節では分析の結果、第 4 節では考察と結論を述べる。

2. 分析手法とデータ

近年では、大規模コーパスの発展によって大量のテキストデータが入手可能であり、それらを用いた言語研究が進んでいる。そのような研究の過程では多くの場合、分析者が設定した特定の意味・機能的な分類基準をもとにして、分析対象の言語表現を含む用例にアノテーションを施すという方法がとられる。そのような分析における問題点のひとつとしてアノテーションの基準が主観的になる可能性が挙げられる。このような問題に対処するためのひとつの方法として SVS がある。以下、2.1.節では SVS の概要を述べる。

2.1. Semantic vector space

SVS の特徴は、特定の言語表現の文脈語との共起関係をもとにして、考察対象となる言語表現間の類似性をもとに、それらの言語表現の距離を捉える点にある。この類似性は、分析対象の言語表現と文脈語の頻度情報をもとにして得られ、分析者による手作業のアノテーションを分析の過程として含まないため、主観的なアノテーションを避けること、および、大量の言語データの分析が可能である。SVS は、特定語彙の性質はコロケーションの検討により明らかになるという Firth (1957: 11) の考えを根底としている。

SVS には、type-based SVS と token-based SVS の 2 種類があり、本研究では後者を採用するが、token-based SVS の作業過程の一部として type-based SVS も含まれる。これは token-based SVS が type-based SVS の短所を補うための手法として開発されたという点に関連している (詳しくは Heylen et al. (2015)を参照)。

SVS は主に、コーパス内の言語表現間の共起表の作成、分析対象となる言語表現の分布の類似度の算出と、その視覚化という段階から構成される。本研究の分析手法は主に、Heylen et al. (2015)と Hilpert and Saaavedra (2017) に基づく。

*本研究は JSPS 科研費 19K13189 の助成を受けたものである。

2.2. 分析: 「犬」「猫」「机」

Token-based SVS の最初の過程では、考察対象となる言語表現をキーワードとして設定し、コーパス内の用例を取得した上で、それらのキーワードの左右に生じる特定範囲内の文脈語 (cw_1) を取得し、共起頻度表を作成する。その後、cw_1 を新たなキーワードとして設定し、コーパス全体から cw_1 の特定範囲内の文脈語 (cw_2) の共起頻度表の作成という過程を踏む。

ここでは、「猫」、「犬」、「机」の 3 つの名詞の意味的類似性の検討を例として分析の過程を示す。ここでは、以下の作例を対象とするが、通常、コーパスの用例を対象とする。それぞれ太字が考察対象のキーワードを表す。

(1) a. 尾の短い**犬**は、撫でられると喜んだ。
b. 尻尾を触られた**猫**は、とても怒った。
c. 部屋の**机**を開いて、ペンを取り出した。

次の行程では各キーワードと、その左右に直接生じる一次共起語 (first-order occurrences 以下、cw_1) の共起表を作成する。通常、機能語などの高頻度語は文脈語から除外し、内容語のみを文脈語とする。上の例では、キーワードから左右 5 語以内の範囲にある名詞・形容詞・動詞を cw_1 とした。表 1 は行に各キーワード、列に cw_1 を含む。

	尾	短い	撫でる	喜ぶ	尻尾	触る	怒る	部屋	開く	ペン	取り出す
(1a) 犬	1	1	1	1	0	0	0	0	0	0	0
(1b) 猫	0	0	0	0	1	1	1	0	0	0	0
(1c) 机	0	0	0	0	0	0	0	1	1	1	1

表 1 キーワード・一次共起語頻度

cw_1 にもとづく分析の問題点は用例中に生起する cw_1 が限定的である点にある。表 1 のように表内のセルは、ほとんどが頻度がゼロで、共通した語彙もない。また、頻度がゼロでなくても、共起頻度は低い。このデータの不足は分析に影響を及ぼす可能性があるため、token-based SVS では cw_1 を新たなキーワードとして、コーパス全体から cw_1 を含む用例から二次共起語 (second-order cooccurrences, 以下、cw_2) を抽出し、キーワードの間接的な共起語として想定するという方法をとる。

	顎髭	散らかる	インク	ちぎれる	頭	脚
尾	0	0	0	0	45	2
短い	1	0	0	0	18	19
撫でる	7	0	0	1	238	3
喜ぶ	0	0	0	0	2	0
尻尾	0	0	0	12	33	1
触る	0	0	0	0	10	3
怒る	0	2	0	0	9	0
部屋	0	32	0	0	7	3
開く	0	1	1	0	26	56
ペン	0	0	15	0	0	0
取り出す	0	0	1	0	5	0

表 2 一次・二次共起語頻度

表 2 は行に cw_1、列に cw_2 の一部を含む。この表は、cw_1 とコーパスに生起する cw_2 の共起頻度を示すため、表 1 の問題である共起頻度の低さは解消される。次の段階として、cw_1 と cw_2 の共起強度を測定するために、Positive Point-wise Mutual Information (PPMI) を用いる。PPMI は PMI という共起強度指標を基盤としており、 cw_1 と cw_2 の PMI は (2) のように求められる。PPMI は、PMI が負の値をとる場合、その値を 0 に変換することで求めることができる。表 3 は、そのようにして得られた PPMI を示す。例えば、表 3 の cw_1 の「撫でる」の行は、「顎髭」、「頭」のような動作の対象となる意味的に結びつきの強い語彙と高い PPMI を示すということを表す。また、「尾」、「尻尾」は、「頭」「脚」が cw_2 の際、それぞれ、近い PPMI の値を示しており、分布の類似性も捉えやすくなる。

(2) $$\mathrm{PMI}\,(cw_1, cw_2) = \log_2 \frac{\text{cw1 と cw2 の共起頻度}}{\text{cw1 と cw2 の共起頻度の期待値}}$$ (cf. Levshina 2015: 327)

以下の表 3 では複数行にわたって、各用例の一次共起語が示されているが、分析の目的は (1a-c) の「犬」「猫」「机」の意味的距離を把握することであるため、次の段階として、各用例を一つの行として表した上で、それぞれの距離を捉える。そのためにまず、キーワードと cw_1 の PPMI を求め、その PPMI を各列に示される cw_1 と cw_2 の PPMI に乗算する。このことで、キーワード、cw_1、cw_2 の結びつきの強さを反映させた値が得られる。そこから、共通した用例に生起する cw_1 と cw_2 に関する PPMI 平均値を、該当する用例でのキーワードの分布を示す表が得られる。例えば (1a) の文脈語の分布を示す値は次のように求められる。

	顎髭	散らかる	インク	ちぎれる	頭	脚
尾	0.00	0.00	0.00	0.00	4.70	3.61
短い	3.65	0.00	0.00	0.00	0.56	4.04
撫でる	8.53	0.00	0.00	5.20	6.35	3.45
喜ぶ	0.00	0.00	0.00	0.00	0.00	0.00
尻尾	0.00	0.00	0.00	9.34	4.06	2.42
触る	0.00	0.00	0.00	0.00	0.84	2.50
怒る	0.00	4.73	0.00	0.00	0.00	0.00
部屋	0.00	6.79	0.00	0.00	0.00	0.00
開く	0.00	1.69	0.62	0.00	0.00	3.86
ペン	0.00	0.00	9.55	0.00	0.00	0.00
取り出す	0.00	0.00	2.78	0.00	0.00	0.00

表 3 二次共起語 PPMI

(3) a. 「犬」(キーワード) と「尻尾」(cw_1) の PPMI を求める (PPMI = 5.43)
b. キーワードごとに表 3 の「尻尾」の行の各列に (3a) の PPMI を乗算
c. 各用例の cw_1 に関して cw_2 の値を平均した表の作成

上の処理の結果、(1a-c) における各キーワードの共起語分布を表 4 のように表した上で、各用例の分布の類似度をコサイン類似度によって求めると、それぞれの位置関係は図 1 のように視覚化される。

	顎髭	散らかる	インク	ちぎれる	頭	脚
(1a) 犬	3.18	0.59	1.25	6.36	6.13	4.74
(1b) 猫	4.22	0.84	0.02	6.99	6.89	4.84
(1c) 机	0.00	1.89	5.26	1.63	0.71	0.42

表 4 各用例の分布値

図 1 では、(1a) と (1b) の「猫」「犬」をキーワードとして含む用例が隣接して位置しており、(1c) の「机」は、その 2 つから離れた位置にある。動物である「猫」と「犬」が近い位置に生じ、物体である「机」が遠い位置にあることは、直観的に理解しやすいと考えられる。このように、token-based SVS を用いて、キーワードとなる表現とその文脈情報にもとづいて、その意味的関係を捉えることが出来る。

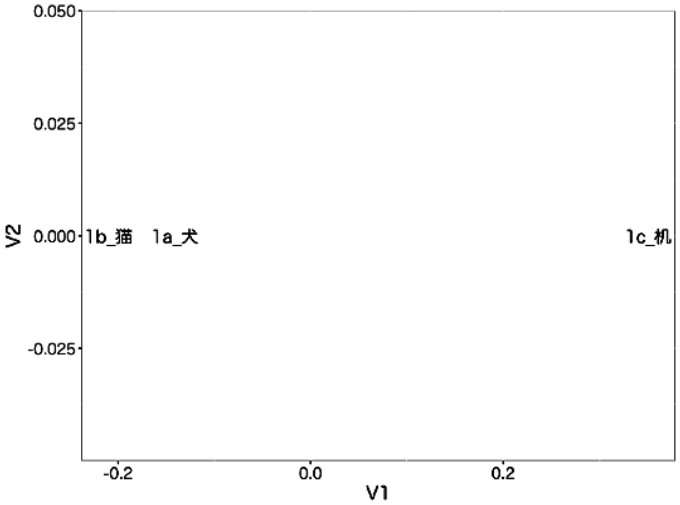

図 1 犬・猫・机の意味関係

3. Semantic vector space を用いたバ条件節分析

前節では SVS が 3 つの名詞に関して文脈語の情報から、それぞれの語彙を分類出来ることを見たが、本節では、従属節と主節が接続助詞の「バ」で接続される条件文 (以下、バ条件文) の従属節を SVS で分類を行う。

3.1. データと分析手法

バ条件文の用例は『日本語書き言葉均衡コーパス有償版』 (国立国語研究所 2012) から接続助詞バの左側に生起する 1 語 (L1) が仮定形の活用形で表される語彙を含むものを抽出した。その中から、代名詞・連体詞・接続詞・感動詞・接頭辞・接尾辞・助動詞・記号、および、従属節の接続助詞バを除いて、従属節内に最低でも 3 語の語彙が含む用例から 273 例を無作為抽出し、SVS の処理を行った。分析にあたっては、処理上の問題により、コーパス内に含まれる上記の品詞以外の上位 2 万語および助詞のみを分析に用いた。その後、前節でとった方法にしたがい、上記のコーパス内に含まれる対象の語彙の共起頻度表と PPMI 表を作成した上で、それぞれの用例に含まれる語彙とその文脈語の情報から、それぞれの用例の語彙分布表を作成した。その分布表から、各用例の類似度をコサイン類似度にもとづいて算出した上で、それぞれの意味的距離を多次元尺度構成法によって視覚化した。プロット上に配置された用例の分類には k 平均法を使用した。分析には、統計ソフトウェア R (R Core Team 2021) を用いた。

3.2. 結果

分析の結果は現代日本語のバ条件文の従属節の用例は、図 2 のように 5 つのクラスターに分類可能であることを示す。図中の右端の凡例は、それぞれ対応するクラスターの番号を示す。プロット上のマーカーの大きさは、各用例における接続助詞「バ」と、その cw_1

の PPMI の最大値を表しており、バ条件節との関連性の高さを反映する。以下、それぞれのクラスターの特徴について述べる。

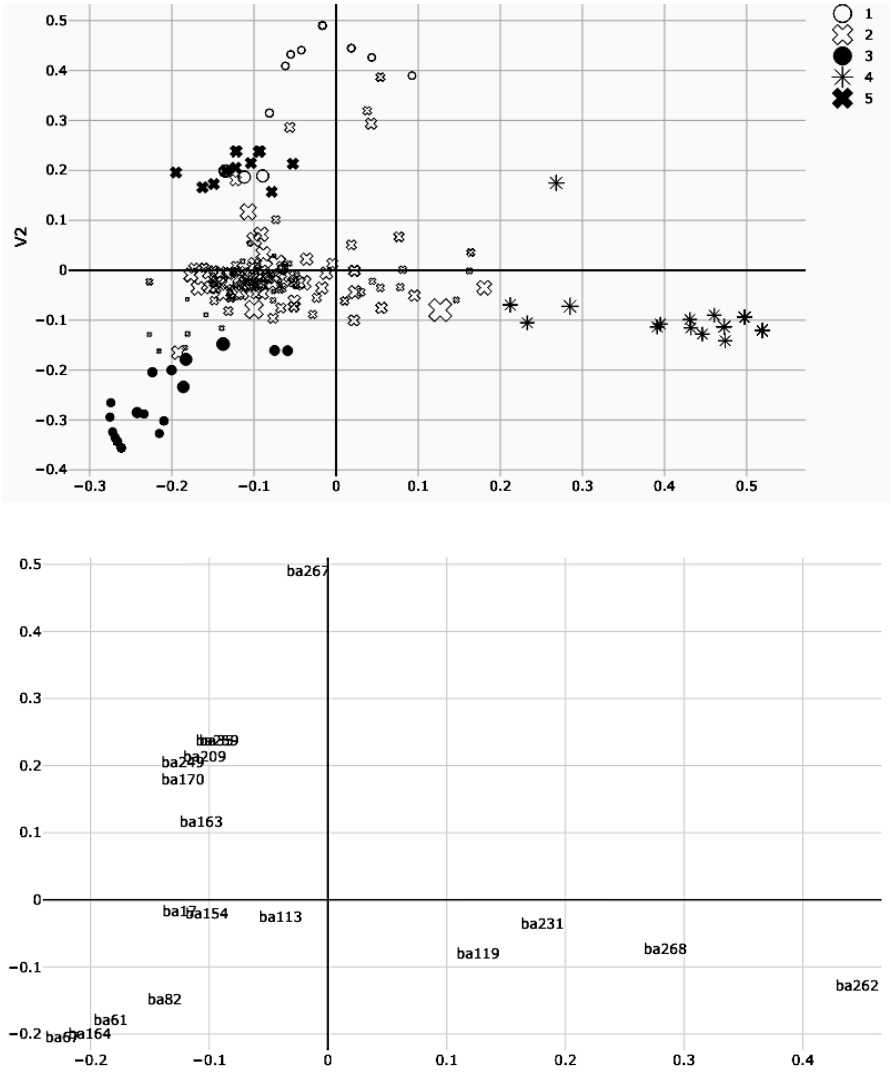

図 2 バ条件文従属節のクラスター分布

3.2.1. クラスター1

クラスター1 は 29 の用例から構成され、図の中央上部に位置する用例は「する」、クラスター5 と隣接した位置の用例は「出来る」を含む。[1] このクラスターに属するすべての用例は、cw_1 として「する」、もしくは「出来る」を含む。そのうち、「出来る」 (PPMI＝1.76) の PPMI は「する」(PPMI=0.68) の場合よりも高い。(4) は cw_1 に「する」のみを含む用例であり、図の最上部に位置づけられる。下線部は用例中の cw_1 の最大値を示す語

[1] 図 2 では一部の用例が表示されていないが、これは、同じ座標上に存在する用例が複数存在することによる。このような状況は、他のクラスターにも見られる。

彙、角括弧は、その PPMI の値を示す。

(4)　　どのようにすれ [0.68] ば… (ba267)[2]

3.2.2. クラスター2

クラスター2 は 149 の用例を含み、5 つのクラスターの中で最も多く用例を含む。クラスター1 よりも、各用例で最も高い PPMI を示す語彙の種類は多く、各用例の PPMI の最大値の幅も大きい。このうち、5 つの用例を除き、全て L1 の位置に生じる動詞が用例中で最大の PPMI を示す。図 3 に示されるように、PPMI が 1.00 以上の値を示す語彙を含む用例は V1 軸の-0.1 付近に集中しており、 (5ab) のような例が挙げられる。

(5) a.　換言すれ [6.06] ば、それは… (ba154)
　　b　　とにかく砂糖をたくさん [0.01] 入れれ [5.09] ば… (ba17)

クラスター2 には、直観的に条件文と関連性の高いと推測される文脈語も見られる。その 1 つとして「もし」を含む (6) が挙げられる。

(6)　　もし [2.11] 心配ならば… (ba113)

(6) は「バ」と高い PPMI を示しており、条件節との関連性が高いことを示す。

3.2.3. クラスター3

クラスター3 は 30 の用例を含み、各用例は「有る」「無い」など、存在に関する語彙を含む。必ずしも、その 2 つの語彙が用例中で最大値の PPMI を示すわけではなく、「暇」、「戦」のような存在対象を示す名詞が、用例中の cw_1 で最も高い PPMI を示す場合もある。

(7) a.　暇 [1.96] があれ [0.80] ば二人は肩を並べて話をした。(ba82)
　　b.　戦 [1.71] があれ [0.80] ば身体を張って村を守ったでしょう… (ba61)

このような例は、バ条件節との関連性が高く、存在を示す表現とも関連性が高い (PPMI(暇, 有る) = 2.24) ことから、このクラスターでの典型例と考えられる。

このクラスターでは、存在対象物が関連した意味を示す用例が存在する場合、それらの用例が近接した位置に配置される。例えば、次のような例が見られる。

[2] 本研究では「ナラバ」の形式を含む用例もバ条件文に含めた。

(8) a. いくらぐらいお金 [1.09] が [0.18] あれ [0.80] ば遊べるのか… (ba164)
b. それどころか機会 [0.99] さえあれ [0.80] ば… (ba67)

「お金」と「機会」は、何らかの好機が存在することを示すという点で意味的に共通性が見られることから、同じクラスターに含まれる用例のうち、近い意味を示す用例は、近い位置に配置されることを、これらの例は示している。

3.2.4. クラスター4

クラスター4 は、47 例から構成され、1 例を除いて「言う」という語彙を含む。そのうち、「言う」が最も高い PPMI を示す (PPMI=1.76) ものが 31 例、「そう」という語彙が「バ」とは最も高い PPMI を示し (PPMI=1.98)、「そう言えば」という形式で示される例が 14 例存在する。「そう」と「言う」の PPMI が 1.98 であることを考えると、「そう」と「バ」、「そう」と「言う」がそれぞれ、高い PPMI を示すことで、それら全ての語彙を含む「そう言えば」という表現が、バ条件文の従属節の中で特徴的な例として存在することがうかがわれる。

今回の分析対象では、特定の人物の発言に関する仮定条件を示す例ではなく、次のように、主節に示される発言内容の前置きとして述べられる用例が多く見られる。

(9) a. あえて [1.99] いえ [1.76] ば仁斎の古義学はそれによって基本的に成立する… (ba225)
b. ひと言 [2.51] でいえ [1.76] ば、時代が変わったからです。 (ba268)
c. はっきり [0.54] 言え [1.76] ば、彼女は皇室に敵意を抱いている。 (ba262)

クラスター4 での各用例の生起位置には、V1 軸の 0.3 付近から 0.4 付近にかけて空白の位置が見られるが、0.3 よりも左側では、(9ab) のように cw_1 が「言う」と「バ」の PPMI の値に勝る語彙を含む用例、0.4 よりも右側では、「言う」が cw_1 の中で最も「バ」との PPMI が高い要素として含む用例が存在するという特徴が見られる。また、クラスター4 に隣接したクラスター2 として分類された用例にも「言う」を含む用例が存在する。

(10) a. しいて [5.02] 言え [1.76] ばズキンズキンって感じでしょうか。(ba119)
b. 簡単に [2.08] 言え [1.76] ば… (ba231)

(10a) は「バ」と高い PPMI を示すとともに、「言う」とも高い PPMI (4.24) を示す。(10b) では、それに比較すると、「言う」との PPMI は低い値 (1.00) を示す。この場合、 (10a) がクラスター2 の中心側に位置するということを考えると、cw_1 と「言う」および「バ」の PPMI との関連性がクラスターの分類に関与していると考えられる。

3.2.5. クラスター5

クラスター5 は 18 の用例を含み、1 例を除いては接続助詞バの L1 の位置に動詞の「成る」が生起し、すべての用例において、この語彙が cw_1 において「バ」との PPMI の最大値を示す (PPMI = 1.21)。他のクラスターと同様、同じような統語環境や意味を示す用例は近接した位置に配置されている。

(11) a. 就職して企業人と [0.19] なれ [1.21] ば… (ba259)
b. “東京ジャーナリズム”が流れを収斂するきっかけ [0.02] となれ [1.21] ば… (ba35)
c. ご覧 [0.46] になれ [1.21] ばわかりますように … (ba249)
d. この魔法書の術を知る [0.22] ようになれ [1.21] ば … (ba209)

(11a-d) のうち、(11ab) は「となれば」という表現が一致していることにより同じ座標位置に生じ、(11cd) では「ご覧」と「知る」という表現に示される動作や状態が引き起こされることを仮定しているという点で類似性が見られることから、近接した位置に生じていると考えられる。

クラスター5 の下部には、クラスター2 が位置しており、その中に「成る」を含む用例が 2 例存在する。

(12) a. 出て いく [1.00] ようになれ [1.21] ば、私たちは… (ba170)
b. 株に興味がある人の助け [2.45] になれ [1.21] ば幸いです。 (ba163)

(12a) では移動動作をとることが従属節で示されており、特定の動作が起こるという変化を表現する点については (11cd) と意味的に関連しており、生起位置も (11cd) と近接している。(12b) は発話者のアドバイスが役立つことが従属節で仮定されており、「になる」という表現を含むため、クラスター5 に分類されても自然であると考えられるが、クラスター2 に分類される背景には、cw_1 の PPMI が関連していると推測される。(11a-d) のクラスター5 に含まれる cw_1 の「成る」以外の PPMI と、クラスター2 の (12ab) のものを比較すると後者の cw_1 がより高い PPMI を示す。このことから、「成る」という表現を L1 に含む用例で、「バ」と高い関連性を示す項目を cw_1 に多く含むと、クラスター2 へと近づくという状況が背景にあると考えられる。

4. 考察と結論

上記の結果は、バ条件文の条件節は生産性の違いを示すクラスターから構成されること、また、それぞれのクラスターが連続的であることを示す。

接続助詞の「バ」と関連性の高い語彙を多く含むクラスターとして、クラスター2 が挙げ

られるが、先に見たように、このクラスターには他のクラスターと共通の語彙を文脈語として含む用例が存在する。例えば、クラスター5 は「成る」、クラスター4 は「言う」という語彙を中心としたクラスターだが、クラスター2 にも「成る」、「言う」を含む用例が存在することが指摘された。そのような場合、クラスター2 は cw_1 として「成る」「言う」よりも「バ」と高い PPMI を示す文脈語を含んでいる。つまり、クラスター2 には、バ条件文の条件節全体の用例を含むこととなる。クラスター2 の各用例中で「バ」との PPMI が高い語彙は L1 に多く、その語彙のタイプ頻度は高いため、意味的な一貫性を規定することは困難である。それに対して、クラスター1, 3, 4, 5 についてはそれぞれ、中心となる L1 の種類は限定的で、「行為・可能」、「存在」、「発言」、「変化」という意味的な一般化が可能である。このように、クラスター2 と、それ以外のクラスターでは、生産性 (Barðdal 2008) の違いがある。

クラスターによってバ条件節と関連性が高い語彙について生産性の違いがあると同時に、上で見たように、2 つのクラスターにおいて、片方のクラスターが、他方のクラスターに特徴的な語彙を含む場合もあることから、明確な境界線を各クラスターの間に引くことは難しく、連続的にクラスターが存在していることを本研究の結果は示している。

SVS を用いた本研究の成果として、バ条件文の条件節が 5 つのクラスターに分類されること、その分類されたクラスターにおける特徴的な語彙には生産性の違いがあること、それぞれのクラスターは連続的に構成されることが明らかになった点が挙げられる。

参考文献・ソフトウェア

Barðdal, Jóhanna (2008) *Productivity: Evidence from case and argument structure in Icelandic.* Amsterdam: John Benjamin Publishing Company.

Firth, John R. (1957) A synopsis of linguistic theory. *Studies in linguistic analysis.* 1-31.

Heylen, Kris, Dirk Speelman, Thomas Wielfaert, Dirk Speelman and Dirk Geeraerts (2015) Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua* 157: 153-172.

Hilpert, Martin and David C. Saaavedra (2017) Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Liguistics and Linguistic Theory.* 1-32.

国立国語研究所 (2012) 『現代日本語書き言葉近郊コーパス有償版』東京: 国立国語研究所.

Levshina, Natalia (2015) *How to do linguistics with R: Data exploration and statistical analysis.* Amsterdam: John Benjamin Publishing Company.

前田直子 (2009) 『日本語の複文−条件文と原因・理由分の記述的研究』東京: くろしお出版.

南不二男 (1974) 『現代日本語の構造』東京: 大修館書店.

中島悦子 (2007) 『条件表現の研究』東京: おうふう.

R Core Team (2021) *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.