P-2

日本語並列名詞句における語順決定要因の検討

吉田眞三（dod17930@kwansei.ac.jp），中野陽子(y-k.nakano@kwansei.ac.jp)
関西学院大学　言語コミュニケーション文化研究科

要旨

日本語コーパス jpTenTen11 から並列名詞句「名詞 A と名詞 B」を抽出し、名詞 A（以降 A）と名詞 B（以降 B）の 2 つの名詞の語順と名詞句「A と B」の頻度の関係および A と B の語順を決定する要因について検討した。コーパス上の調整頻度から相対頻度を算出し、調整頻度と相対頻度の関係を検討し、また web 上で調整頻度の異なる並列名詞句について、A と B の語順について容認度調査を行った。いずれの検討においても調整頻度が高い表現では語順は固定傾向となり、調整頻度が低い表現では語順の可変性が高いという結果が得られた。さらにコーパスから抽出した並列名詞句について、意味的制約、モーラ数のパターン、単語頻度のパターン、調整頻度を予測変数とし、各表現の相対頻度を応答変数として重回帰分析を行ったところ、意味的制約が語順の形成に大きな影響を与えていることが示された。頻度の低い並列名詞句「A と B」と高頻度の並列名詞句では、処理過程が異なる可能性が示唆された。

1．緒言

単語認知における頻度効果についてはよく知られており、出現頻度（使用頻度）の高い単語ほど認知されやすいとされている。近年では単語のみならず多単語表現(multi-word expression)においても、高頻度の表現については、学習・記憶されそのまま再利用されることにより処理速度の向上につながっているのではないかとする議論がなされている(Morgan & Levy 2016)。 Arnon and Snider (2010)は、”Don’t have to worry.”などの非常に高頻度で使用される表現と、表現を構成する単語の頻度は一致しているが、フレーズとしてより低頻度の”Don’t have to wait.”といった表現、及びフィラー表現を用いて、語彙決定試験と同様な方法で、それが英語としてフレーズをなしているかどうかを判定させるフレーズ決定試験(phrase-decision test)を行っている。高頻度表現で処理速度が早いことを示しており、表現全体として記憶・再利用されるのではないかと報告している。英語の過去形産出における議論においても、使用頻度の非常に高い規則動詞では、過去形も心内辞書に記憶されている可能性を示す研究がある（Taft, 1979)。”A and B”という英語二項表現において Morgan and Levy (2016)は、高頻度表現ではその語順の選好に直接の言語経験がより大きな影響を与えることを実験結果として示している。さらに英語二項表現については、その語順の決定に関わる各種制約について多くの研究の蓄積があり、Cooper and Ross (1975)において集約的な検討がなされ、「ここに今あるもの、成人、男性に関係するもの、積極性、生あるもの、動作主、友好的、愛国的」といった事象に関連する単語が、その対極にあるものの前に来るという”Me first"原則が提唱されている。

一方、英語二項表現に対応すると考えられる日本語並列名詞句において、その語順と表現頻度の関係

について検討した報告は、文献を渉猟した限りはない。また語順に関わる要因について体系的に検討した報告も極めて乏しい。このような観点から、今回我々は、コーパスから異なる頻度の表現を抽出し、出現頻度（使用頻度）が「A」と「B」の語順にどのような影響を与えるのかということについて検討を行った。さらに「A と B」の語順形成に影響を与える要因について検討をおこなったので報告する。

２．方法及び結果

2.1 コーパスからの並列名詞句の抽出

2.1.1 調整頻度上位群の抽出

日本語コーパス jpTenTen11(sample version)を用い、検索ソフト Sketch Engine[1]のコンコーダンス機能を用いて CQR([tag="N.c.g"][word="と"][tag="N.c.g"])として並列名詞句を抽出した。329,642 表現(893.67/million)がヒットし、これらを頻度でソートした後、この上位 1000 表現をダウンロードした。解析に不適当と考えられる表現を削除した後[2]、残った 440 表現についてその上位 50 表現を高頻度表現、下位 50 表現を中頻度表現とした。これら 440 表現の調整頻度(100 万 token あたりの表現の度数)の分布を図 1 に示す。図 1 のグラフは 440 表現を順に並べ、縦軸にその調整頻度を示している。

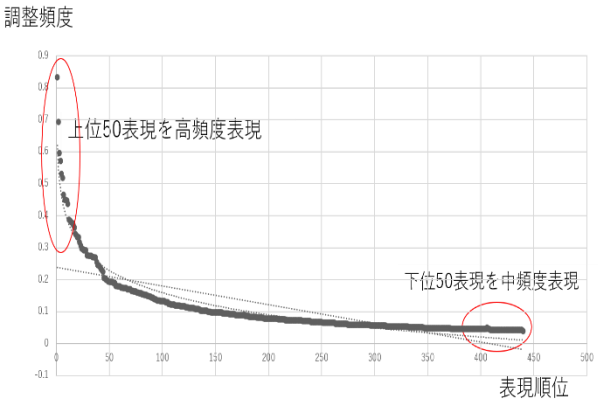

図１　2.1.1でコーパスより抽出した440表現

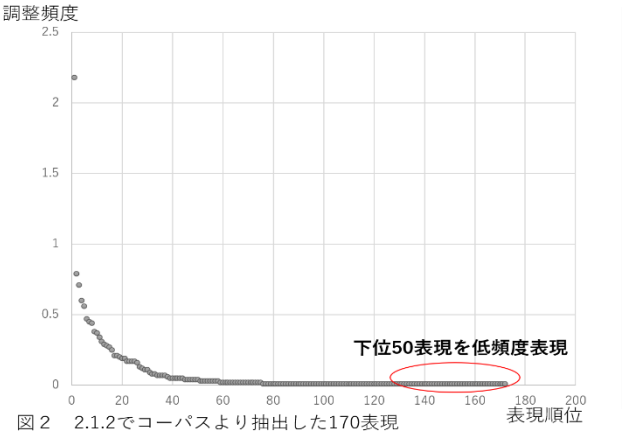

図２　2.1.2でコーパスより抽出した170表現

2.1.2 調整頻度下位群の抽出

2.1.1 で抽出した表現よりもさらに下位の表現を抽出すべく、同様に検索しヒットした表現 329,642 表現について 1000 表現をランダムサンプリングし、これをダウンロードした。①と同様の方針で解析に不適当な表現を削除した後、残った 170 表現について、その下位 50 表現を低頻度表現とした（図２）。

---

[1] Sketch Engine (https://www.sketchengine.eu/)

[2] ①リストに「A と B」「B と A」の両者がある場合は頻度の低い方を削除　②「A と A」は削除　③「罪と罰」「千と千尋」などの映画，本のタイトルなどは削除，④「メリットとデメリット」などの仮名のみの表現は削除　⑤「青と赤」などの色の組み合わせは個人的嗜好が多いと判断し削除，⑥「簗」など常用漢字に含まれない漢字を含む表現も削除した。③～⑥は後に読み課題を行う上で問題となると判断した。

ちなみにこれら下位 50 表現のコーパス上の出現度数はすべて 1 であった。

2.2.　調整頻度と相対頻度の関連についての検討

2.1.1 で得た調整頻度上位 440 表現について、相対頻度（「A と B」の調整頻度/（「A と B」の調整頻度＋「B と A」の調整頻度））を算出した。440 表現について、相対頻度と調整頻度の関連を散布図に示したのが図 3 である。図 3 では表現「A と B」を五十音語順で示している。例えば「光と影」という表現を例にとると、その相対頻度は 0.98 であるが、五十音順では「影と光」となりその相対頻度は 0.02 となり、図 3 の赤丸の位置となる。横軸は表現「A と B」の相対頻度を示し、0 または 1 に近い程、その語順は固定傾向となり、0.5 付近では語順は可変で一定の傾向を示さないことを表す。縦軸は表現「A と B」の調整頻度を示す。図からは表現の調整頻度が上昇するにつれて相対頻度は 0 または 1 に近づく傾向が窺え、語順が固定する傾向にあることが示唆される。

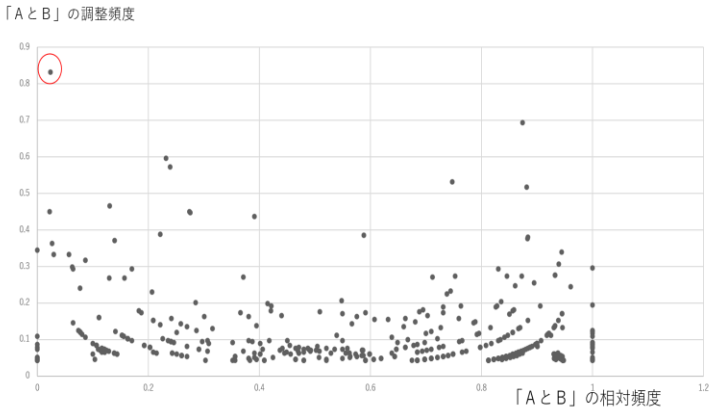

図 3 2.1.1で抽出した440表現について、相対頻度と調整頻度を散布図で示す

この結果を踏まえて、さらに詳細に検討すべく、2.1.1 で得た高頻度表現 50，中頻度表現 50 について、その相対頻度について検討したものが図 4 、図 5 である。横軸は相対頻度を表し、縦軸にそれぞれの相対頻度の区間毎の度数を示している。図 4 の高頻度表現では相対頻度が両端の 0 または 1 に近い表現が多い傾向が認められるのに対し、図 5 の中頻度表現では相対頻度が 0.5 の中央付近の表現が多い傾向が認められた。

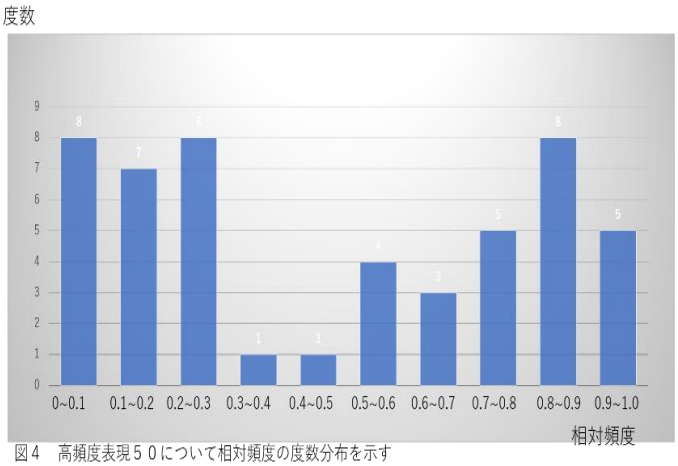

図 4　高頻度表現５０について相対頻度の度数分布を示す

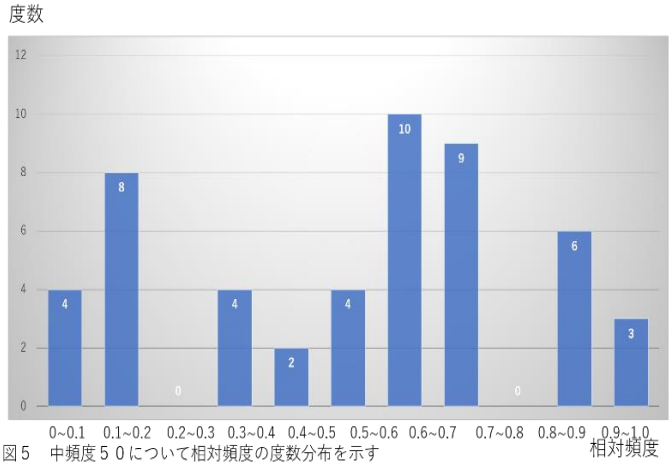

図 5　中頻度５０について相対頻度の度数分布を示す

2.3 容認性判断課題による検討

2.2 で認められたコーパスから抽出した表現についての調整頻度と語順の固定性の関係について実際に人における選好を検討するためにクラウドソーシングを利用して web 上で容認性判断課題を行った。2.1 で得た高頻度表現 50、中頻度表現 50、低頻度表現 50 について、日本語母語話者を対象とし図 6 に示す方法で正語順と逆語順 でそれぞれ表現を提示し、選好について回答を得た。コーパスから抽出した表現の語順を正語順とし、これと反対の語順を逆語順とした。被験者はランサーズ（株）を通じて募集した 39 人であった（男性 28 人、女性 11 人、平均年齢 41 歳)。正語順を先（選択肢 A として）に提示するか、後（選択肢 B として）に提示するかはランダム化し、参加者一人当たり 150 表現について回答してもらった。結果は R(ver 4.0.1)を用いて一般化線形混合効果回帰分析モデルにより解析した。解析には lme4 パッケージ（Bates et al. 2015)を使用した。表現の頻度 f1（高頻度/中頻度/低頻度)、表現の提示の仕方 f2（正語順を先/逆語順を先）を固定要因、被験者 subject と項目 item をランダム要因とし、語順の選好 pre_order（コーパスと同語順を選好/コーパスと逆語順を選好）を従属変数とした。減数法(Baayan et al. 2008)を用いて最適モデルを得た（式 1)。

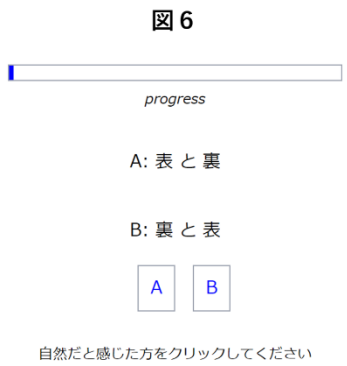

図 6　容認性判断課題

式（ 1 ） glmer(pre_order~f1*f2+(1+f1*f2|subject)+(1+f1|item),data=y,family=“binomial”)

全体分析では，表現頻度において主効果が見られた（β =2.4380, SE=0.2857，z =8.5331，p<.001)。表現の提示の仕方においても主効果が見られた（β =0.24340, SE=0.06823, z=3.567, p<0.001)。下位分析では正語順を先に提示した条件で高頻度表現・中頻度表現間（β =1.9753 ，SE=0.3421, z=5.774, p<.001)、高頻度表現・低頻度表現間（β =2.5790 , SE=0.3548, z =7.269, p<.001)、中頻度表現･低頻度表現間(β = 0.6280, SE=0.2966，z =2.118, p<.05)に語順の選好において、統計的に有意な差を認めた。また逆語順を先に提示した条件でも同様に高頻度表現・中頻度表現間（β =1.5764, SE=0.2464, z=6.399, p<.001)、高頻度（β =2.2211, SE=0.2847, z=7.303, p<.001)、中頻度・低頻度表現間（β =0.6529, SE=0.2654, z=2.460, p<.05)で語順の選好において有意な差を認めた（図 7)。高頻度表現では正語順の選択は

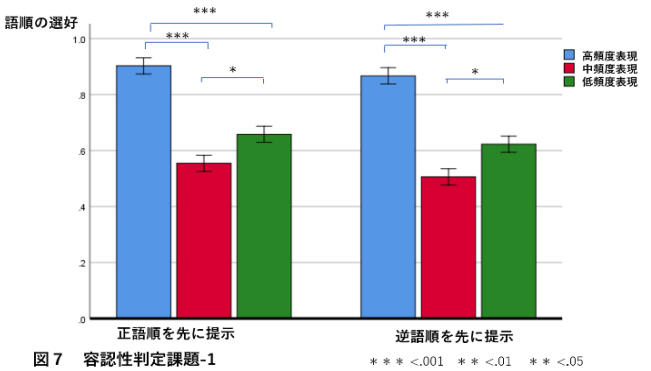

図 7　容認性判定課題-1
＊＊＊ <.001　＊＊ <.01　＊＊ <.05

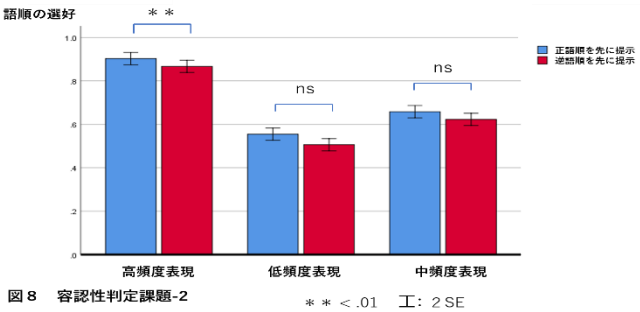

図 8　容認性判定課題-2
＊＊ < .01　工: 2 SE

88.5±1.0%(mean±se)であったのに対し、低頻度表現では語順の選好はチャンスレベルに近かった(53.0±1.0%)。中頻度表現は両者の中間(64.1±1.0%)に位置した。また高頻度表現では正語順、逆語順の提示順序により語順の選好に統計的に有意差が見られたのに対し、中頻度表現・低頻度表現では提示順序による差は認められなかった（図８)。図７，８の縦軸は語順の選好を示す。1 に近づく程、語順が固定傾向となり、0.5 に近づく程語順が可変となることを示す。

2.4 語順に影響を与える要因についての検討

これまでの検討で表現の経験頻度と語順の固定に関連があることが示されたが、語順に影響を与えると想定されると思われる要因についてさらに検討した。2.1.1 で得た 440 表現について以下の要因と語順の関連について検討した。(ⅰ)AとBのモーラ数の関係　(ⅱ)A と B の頻度の関係　(ⅲ)意味的制約（Morgan & Levy(2016) における”power”,” perceptual markedness”, “iconic sequencing”,”formal markedness”の意味的制約を採用）　(ⅳ)コーパス上の調整頻度　について検討した。(ⅰ) のモーラ数については、A<B の場合を+1 ,A=B の場合を 0, A>B を-1 としてそれぞれの表現についてコーディングした。(ⅱ)の単語の頻度については、(A の頻度)＞（B の頻度）を+1, (A の頻度)<(B の頻度)を-1 とした。(ⅲ）の意味的制約については上述した 4 つの要素について、表現「A と B」の語順に対して、その制約が有効であり語順がそれに従っている場合を+1, 制約が有効でない場合を 0, 制約は有効であるが語順がそれに反している場合を-1 として 4 つの要素についてそれぞれコーディングを行った。例えば、「光と影」という表現では、モーラ数の関係は A>B で-1, 単語頻度の関係は A＞B で+1 と設定, 意味的要因については、perceptual markedness が+1, その他は 0 と設定、調整頻度は 0.83、相対頻度は 0.98 であった。表現を五十音順表示とし、「光と影」は「影と光」としてコーディングを行い、モーラ数の関係、単語の頻度の関係、意味的制約は逆転したものを設定、また調整頻度は新たにコーパスより 0.02 と算出、また相対頻度は 1-0.98=0.02 と算出した。このような操作を 440 表現について行い、これら 7 つの要因を予測変数、表現「A と B」の相対頻度を応答変数として重回帰分析を行った。分析にはＳＰＳＳversion26 を用いた。変数の強制投入により有意なモデルが得られた (F (7,431) ＝125.230、p＜.000, R^2=.670)。検討したすべての予測変数について有意な効果(p<.01)が認められたが、その影響については意味的制約 power がもっとも大きく (β =.469)、ついで perceptual markedness(β =.300)であった (表１)。応答変数であるところの相対頻度は下記の式 (2) で近似された。

式 (2) 相対頻度(estimated)=0.441+0.249×power+0.203×Percep.markedness+0.294×Iconicity+0.186×Formal markedness+0.874×調整頻度+0.092×モーラ変数+0.026×単語頻度変数

図 9 に 440 の表現それぞれについて、コーパスから算出した相対頻度と重回帰分析から得られた近似式から算出した相対頻度をグラフに示す。|r|=0.8237, R^2=0.6704 となり、当てはまりが良いモデルが得られた。

表1
語順の形成に関係すると思われる要因についての重回帰分析結果

	係数				
	B値	SE	β 値	t値	有意確率
Power	0.441	0.016	0.469	15.942	0.000
Percep.markedness	0.203	0.019	0.300	10.631	0.000
Iconicity	0.294	0.033	0.256	8.994	0.000
Foramal markedness	0.186	0.063	0.083	2.956	0.003
調整頻度	0.874	0.125	0.218	7.011	0.000
モーラ数	0.092	0.012	0.230	7.822	0.000
単語の頻度	0.026	0.009	0.084	2.880	0.004

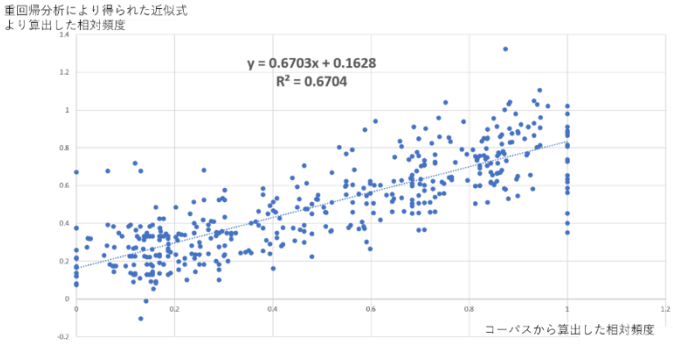

図 9　コーパスから算出した相対頻度と重回帰分析より得られた近似式から算出した相対頻度の相関

3. 考察

本研究ではコーパスから抽出した「A と B」という並列名詞句について、その調整頻度と相対頻度の関係について検討し、さらに人でその語順についての容認性判断課題を行ったが、いずれの検討においても、高頻度表現の語順の選好は固定傾向となるのが観察され、これに対して低頻度表現になるほど、その語順は可変性が高いという結果が得られた。単語レベルのみならず多単語表現においても、その頻度が認知処理に関係するとする指摘は Bybee (2006), Arnon and Snider (2010), Morgan and Levy (2016)などの報告に見られる。今回の検討結果からは、並列名詞句「A と B」において、頻度の低いものと高いものでは処理過程が異なるのではないかということが示唆された。すなわち前者ではそれを構成する「A」「B」二つの語彙とこれを「と」でつなぐという規則に則り、表現が機会毎に生成されるという要素が強くなるのに対し、後者では「A」と「B」の結合が強く、「A と B」という形で心内辞書において一つの語彙として記憶・貯蔵されている、あるいは「A」の提示により「B」が活性化されやすくなっている可能性が考えられた。高頻度表現では中頻度表現・低頻度表現とは異なり、正語順・逆語順の提示順序により語順の選好に有意差が見られたこともこれを示唆しているのではないかと考えられた。

語順に影響を与える要因についての検討では、「A」と「B」の結合については意味的制約が大きな影響を及ぼすことが示された。日本語並列名詞句においてその語順の問題を体系的に検討した研究はほとんどないが、Lowman and Takada (2014)は、日本語話し言葉コーパス(Corpus of Spontaneous Japanese、CSJ)[3] から 861 表現の並列名詞句を収集し、本研究と同様に意味的要因にモーラ数、単語の頻度を加えてロジスティック回帰モデルにより検討している。その語順に与える影響については我々の検討と同様に意味的要因の影響が大きいとしており、またモーラ数の問題についても short before long のパターンが統計的に有意に多いことを示している。また単語の頻度の影響については我々の検討と同様にその影響は小さいことを報告している。著者らのロジスティック回帰モデルによる検討では予測の正解率が 66%と低かったことを報告しているが、本研究においてはコーパスから算出した相対頻度と重回帰モデ

[3] 日本語話し言葉コーパスの詳細については以下 URL を参照
https://ccd.ninjal.ac.jp/csj/misc/preliminary/4.html

ルにおける近似式から算出した相対頻度が、良い相関を示した(|r|=0.8237)のとは異なる。我々の検討においては各表現の相対頻度を応答変数としたことが良好な相関が得られたひとつの要因と考える。また本研究においては調整頻度を予測変数に加えていること、用いたコーパスの違いが関係している可能性も考えられた。

本研究では、名詞並列表現「A と B」を構成するそれぞれの単語の頻度が語順に与える影響は小さかった。また Lowman and Takada (2014)の研究では統計的に有意な影響は示されていない。これは英語の二項表現における検討で一般に“more frequent before less frequent”として知られている頻度の高い語が頻度の低い語より前に来るという制約に反する結果である。特に Fenk-Oczlo(1989)の報告では、検討した表現の 84%でこの原則が見られ、検討した制約の中で最も拘束力の高いものであったとしている。英語における制約の研究はその多くが、慣用句として語順がほぼ固定したいわゆる”frozen”と呼ばれる表現を検討対象にしているものが多く、Fenk-Oczlon の報告も”frozen”の並列表現 400 を対象としたものである。本研究、Lohmann らの研究はいずれも語順が可変な表現を多数含むものであり、対象を語順が固定したものに限ればまた結果は異なる可能性がある。

参考文献

1) Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*, 67-82.

2) Baayan, R.H., & Davidson, D.J, & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412

3) Bates, D., & Maechler, M., & Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*, 1-48.

4) Bybee, J. (2006). From Usage to Grammar: The Mind’s Response to Repetition. *Language. 82*, 711-733.

5) Cooper, W.E. & Ross,J.R. (1975). World order. In R.E.Grossman, L.J.San, & T.J.Vance(Eds.), Papers from the parasession on functionalism(pp. 63-111), Chicago Linguistics Society

6) Fenk-Oczlon, G. (1989). Word frequency and word order in freezes. *Linguistics, 27 ,* 517-556.

7) Lohmann, A., & Takada, T. (2014). Order in NP conjuncts in spoken English and Japanese. *Lingua 152,* 48-64.

8) Morgan, E., & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition157* , 384-402.

9) Taft, M.（1979）. Recognition of affixed words and the word frequency effect. *Memory & Cognition 7* , 263·272.