2007－NL－178 (3)
2007／3／28

コロケーションに着目した日本語テクストのメッセージ分析

國府久嗣 園田勝英
北海道大学大学院国際広報メディア研究科
E-mail: ccoe@mac.com, sonoda@ilcs.hokudai.ac.jp

あらまし 日本語テクストに含まれる語彙項目間のコロケーションに着目し、その状況を視覚化することでメッセージ分析を行なう方法について考察した。このとき統計手法としては主に多次元尺度構成法を用いている。本発表ではコロケーション定義のうち重要な部位をなすspanについて、値や判定法を変化させた際の分析結果との相関について検討した。これによって語彙項目以外をspanに含まない方式には、分析結果がspanの値によって過敏には左右されない特徴があることを明らかにしている。対象テクストが恒常的に有していると考えられるメッセージを抽出し分析するという観点からはこの性質はのぞましい点にも言及した。

Collocation and Message interpretation of a Japanese Text

Hisatsugu Kokubu and Katsuhide Sonoda
Graduate School of International Media and Communication, Hokkaido University
E-mail: ccoe@mac.com, sonoda@ilcs.hokudai.ac.jp

Abstract In this paper we will suggest that it will be useful for interpreting the message(s) of a Japanese text to visualize its frequencies of lexical collocations. The visualization is based on MDS. We explore into the effects of various settings of span. Span is currently considered to be the central parameter of the notion "collocation" in that two elements are said to be in collocation when they cooccur in a certain specified span. It is shown that various settings of the span length do not significantly affect the final configurations obtained through visualization, when span is defined with non-lexical, i.e. functinal, elements excluded. The result supports our initial suggestion because the message of a text we are trying to capture is one of its constant properties.

1 はじめに

本研究の目的は日本語テクストの内容を語彙項目（lexical item）の共起関係に着目し、それを多次元尺度構成法（multidimensional scaling）などの統計手法を用いて視覚化することで把握する方法について考察するところにある。本稿ではこのうち特にコロケーション（collocation）について着目し、日本語テクストを扱う際にその定義と結果がどのような相関にあるのかを検討した。具体的にはコロケーション定義の構成要素の一つであるspanに焦点をあて、spanの判定やその値設定が分析結果に与える影響について述べている。

以下、spanに関する具体的考察の前に本研究で用いる分析手法と概念規定、定義などについて簡単に言及し、その後spanの値を変えながら多次元尺度構成法によって視覚化した語彙項目群布置図表を基に相関の度合いや傾向を分析する。

2 分析手法と定義

ある程度一貫した内容を持つと予測されるテクストからその内容（メッセージ）を数理言語学的な裏付けに基づいて何らかの数値化を行なう事で抽出し、それについて分析しようと目論んだ場合、もっとも有効な視点となるのはコロケーションであると考えられる。ただし、ここで述べるコロケーションは通常よりも広義のものである。英語圏で

の先行研究としてはSinclair（1966）[5]を含む一連のもの[6] [7]をあげることができる。

2.1 コロケーション

コロケーション定義については諸説あるが、それは次の2点に集約される。一つは「何がコロケーションを形成しているか」というものであり、次に「どこまでがコロケーションか」というものである。

前者について別の言い方をすれば、それは「何が語彙項目であるか」ということになる。後者は「どのようにspanを規定するか」に換言される。この二つは別個に独立したものではなく、双方は不可分な関連にある。

コロケーションに関する問題そのものは非常に広範なもの[2] [3]であるため一度にすべてを扱うことは難しい。ここではそのうちspan定義に焦点をあてて、メッセージ分析への貢献という点からのみ考察していく。以下span以外の項目について説明の必要上触れざるを得ないものについて、本研究で用いた手法や各種定義について簡単に述べる。

2.2 視覚化手法

本研究で用いている視覚化手法は、共起頻度の高い語彙項目同士をその度合いによって「より近い関係」と看做して数値化する考えに基づいている。数値化は以下に示す「非類似度（dissimilarity）」を算出することで行なっている。局所最適解や退化した解に陥る抵抗力が大きい手法[8]という利点から多次元尺度構成法のうち特に計量MDS（古典的MDS）を用いており、その都合で非類似度を距離的性質を持つ値となるよう前処理している。

ここでいう非類似度 $dissim(t,t')$ [1]は次のように表せる[2]。

$$dissim(t,t') = \left|\log \frac{freq(t \cap t') + a}{freq(t')}\right|$$

$freq(t)$ は語彙項目 t の出現数を表す。また $freq(t \cap t')$ はspan内で t と別の語彙項目 t' が共起した数である。分母には出現頻度の少ない方を使う。a は頻繁に発生が予想できる「共起無し」の場合でも値にバラつきを出す為に設けた定数である[3]。

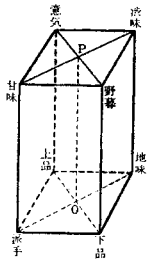

Fig. 1: 趣味構造直六面体

2.3 対象テクストと語彙項目

本稿では分析対象テクストとして青空文庫[4]で公開されている『「いき」の構造』（九鬼周造）を用いた。spanを変更した際に生じる変化を、視覚化された結果から追いやすい特徴的な語彙項目の使用がみられることからこの作品を利用した。特徴的な語彙項目としては、まず第一に標題ともなっている「いき」があげられる。通常の日本語テクストにおいてこの語が出現頻度で最上位となることは考え難い。また分析に用いた形態素解析器[5]の採用する文法基準に起因して「する」が最頻出語彙項目となっているが、ここで「いき」はそれに次ぐ頻出語彙項目である。

また内容や文体面での特徴に関連するが、このテクストの第三章では主に「いき」という概念に関連する趣味項目を8つあげ、それらの関係をFig.1に引用した図[9]に結実するよう解説している。ここで用いられた語彙項目[6]群は他の部分での語彙項目の使用とやや差があるため、この状況が分析結果の図表にどう反映されているかで相関評価に使用できると考えた。

[1] 値は $0 \leq dissim(t,t') < \infty$ の範囲をとる。
[2] $(freq(t) \geq freq(t'), freq(t \cap t') \neq freq(t'))$
[3] 完全に共起した場合にはこの値は加算していない。a 値として適切なのは $0 < a < 1$ の範囲。
[4] http://www.aozora.gr.jp/
[5] 日本語形態素解析システム『茶筌』version-2.3.3 (http://chasen.aist-nara.ac.jp/hiki/ChaSen/)
[6] ただし「派手」の対極にある「地味」は語彙項目に漏れたため出現しない。

なお語彙項目については「出現頻度が大きい」「(所謂) 内容語」とここでは定義した。自然言語においては極少数の word type が word token の大部分を占めるということがよく知られている [1]。これに該当する word type でなおかつ機能語ではないものであれば、対象テクストが持つ特定の意味への寄与度合いが高いと考えて基準を設けた。本研究での基準で選定したところここでは以下の 78 項目が該当した。

「する」「いき」「意味」「ある」「いう」「もつ」「場合」「表現」「存在」「媚態」「なる」「関係」「味」「客観」「芸術」「ない」「形式」「二元」「模様」「渋」「民族」「意気」「できる」「見る」「我々」「対立」「異性」「特殊」「他」「価値」「表わす」「色」「現象」「意識」「形」「可能」「趣味」「考える」「上品」「野暮」「下品」「体験」「語」「示す」「色彩」「自然」「甘味」「横縞」「女」「様態」「縞」「派手」「文化」「因」「建築」「平行」「対」「粋」「諦める」「構造」「自己」「内容」「江戸」「理想」「茶」「一般」「理解」「現実」「縦縞」「成立」「規定」「具体」「取る」「有する」「意気地」「形相」「言」「把握」。

3 span

span の条件変更はまず、非語彙項目をカウントする場合において 1 から 3 までの範囲で行なった。$span = 1$ というのは前後に隣接する関係である。間に一つも他の語を挟んでいない状態を指す。2 で隣接又は一つ挟んだ状態までをカウントする。前記の語彙項目に該当してもしなくても区別せず数える。

次に語彙項目以外をカウントしない方法で、やはり同様に 1 から 3 まで値を変化させた。$span = 1$ は前後に隣接する関係である点は先の場合と同じだが、語彙項目同士の隣接は間に非語彙項目がいくつ入っていても無視して成立する。実質的な span は非語彙項目を数える場合よりも広くなることはあっても狭くなることはない。

そして最後に非語彙項目をどうするか、という問題が発生しない方式として sentence を span とした場合で分析を行なった。これらの結果を以下に示し、結果について検討と考察を行なう。

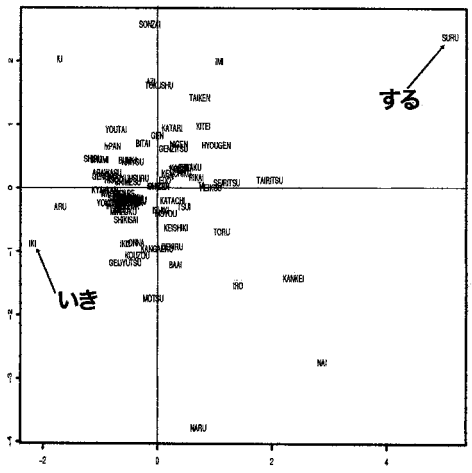

Fig. 2: 非語彙項目を含む $span = 1$ の場合

3.1 非語彙項目を含む場合

非語彙項目を span 測定に含む場合、値を変化させるとどのような結果になるのか実験した。結果は Fig.2 から Fig.4 で示した通りである。顕著な傾向としては、span 値が大きくなるにつれて頻出語彙項目「いき」と「する」が原点近くに接近して配置されるようになっていく点があげられる。また Fig.2 では重なり合いが酷く図示できなかった趣味直六面体語彙が Fig3 以降では判別可能となっている。いずれにせよ、span の値を変えることで算出される座標が大きく変化していることが伺い知れる。

Fig.2 において語彙項目が判別不可能なほど密集している原因は、この条件ではほとんど共起が発生しないせいだと考えられる。また最頻出語彙項目であり他の語彙項目と多く共起してコロケーションを形成すると予測される「いき」と「する」が原点付近ではなく周縁に配される結果となっているのも、共起そのものが発生し難いことを示している。

Fig3 ではそれがやや改善され、Fig4 ではもっと改善されたと一見みなせるようだが、それは妥当とはいえない。この三種類の図表における語彙項目間の布置座標から、それが導きだされて来た元の対象がもつはずの、なんらかの一貫した構造を読み取ることができない（か非常に難しい）ため

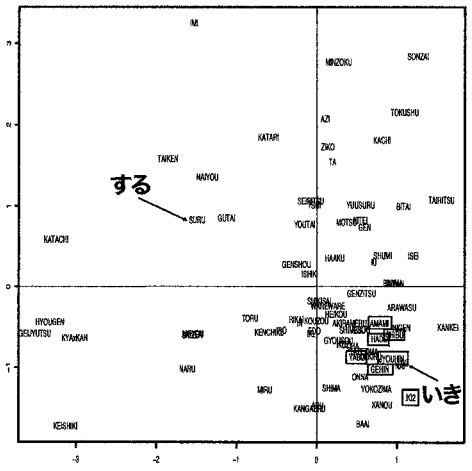

Fig. 3: 非語彙項目を含む $span = 2$ の場合

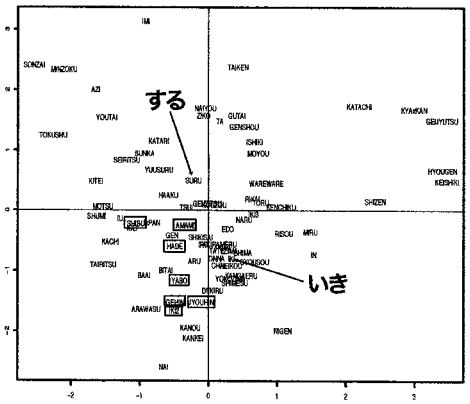

Fig. 4: 非語彙項目を含む $span = 3$ の場合

である。

3.2 語彙項目のみ

一方、語彙項目同士の間に非語彙項目がいくつ挟まっていても span の値としてはカウントしない方式で同様の実験を行なったところ Fig.5 から Fig.7 で作図した通りの結果になった。

Fig.2 から Fig.4 までと比べて顕著なのは、「いき」「する」の位置や趣味直六面体語彙項目の配置に関して span 値の違いによる影響がさほど劇的ではない点である。いずれの場合でも最頻出語彙項目である「いき」と「する」は原点近くに置かれ、趣味直六面体語彙項目も周縁部分に他の語彙項目群とはやや離れた位置に固まって布置される傾向が一貫している。

3.3 sentence

非語彙項目を無視すべきかどうかという問題を回避する尺度として sentence[7] を採択したして同様の実験を行なった。結果は Fig.8 に見られる通りである。語彙項目は Fig.7 を 180 度回転させたものと良く似た布置をしている。

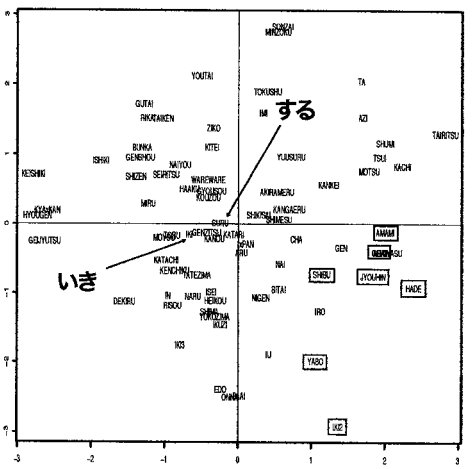

Fig. 5: 語彙項目間 $span = 1$ の場合

[7] ここでは簡単のため句点と句点で挟まれた文字列と定義する。

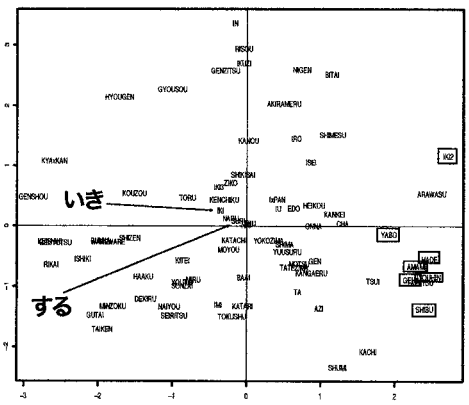

Fig. 6: 語彙項目間 $span = 2$ の場合

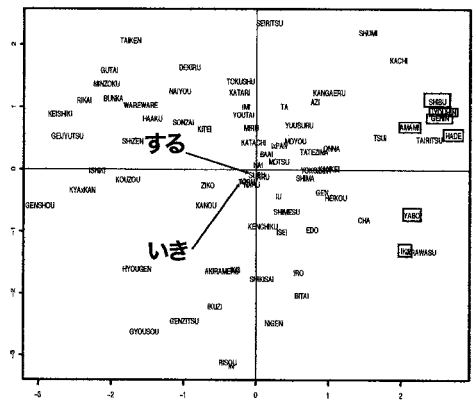

Fig. 7: 語彙項目間 $span = 3$ の場合

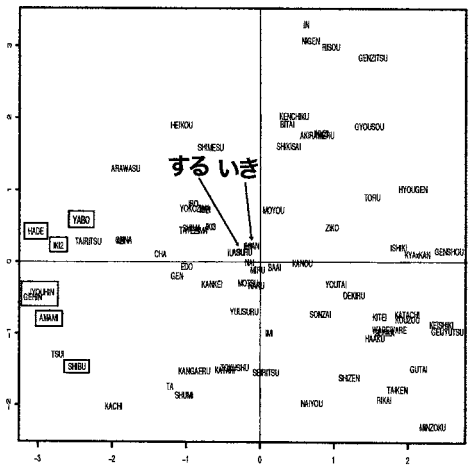

Fig. 8: *sentence* の場合

sentence を span とする定義では前の二つのケースとは異なり、span 値は一定していない。文の長さは一定ではなく、そこに含まれる語彙項目数も一定しないからである。にもかかわらず布置にある種の一貫した構造が見て取れるのは、語彙項目間でのみ span をカウントした際に値を変えてもある程度一貫した構造 Fig.5 から Fig.7 まで共通してみられたことと同様の原因が考えられる。

4 まとめ

コロケーション認定に用いる共起 span について、これまで見て来たような非語彙項目を含むか否かと言う両方式による結果の違いを、「メッセージ分析」に用いるという観点から見た場合、より優れているのは後者の方式だといえる。

メッセージ、あるいはその具体的な構成要素である（ある対象がもつ特定の意味を形成する）コロケーションは、ある程度一貫した形態を持たねばならない。こうした性質を持つものを抽出する手法にも同様の性質が必要となる。

本稿では紙面の関係もあり記述を省略したが、分析の際には多次元尺度法で作成した布置座標から階層クラスター分析（hierarchical cluster analysis）[8]を用いて語彙項目同士のグループ分けを行う作業

[8] Ward 法

過程がある。ここでもそれを行ない、算出した複雑な樹状図（dendrogram）から５～６個程度のクラスターを切り出して [4] 相互に含まれる語彙項目の検討をおこなった。

この分析結果では語彙項目「いき」と「する」は非語彙項目を含むケース以外のすべてのspanで同じクラスター（中央にあるクラスター）に分類されている。また趣味直六面体語彙項目の方はすべてのspanで概ね同一クラスターに分類されているが、こちらの場合重要なのは同一クラスターかどうかではなく（非語彙項目を含む場合とspan=1以外では）趣味直六面体語彙項目が、それのみを主体とするクラスターに分類されたという点である。これはうまく分析対象テクストが持つ語彙構造を反映した結果と考えられる。

統語論的側面に着目してspanを考える場合、非語彙項目を無視してカウントしたりsentenceをspanとすることはやや思慮にかける（理論的に正当と言えない、妥当性を欠く）行為とみえるかもしれない。しかしながら意味の側面について言えば、この方向、つまりspanを比較的広く、緩やかに定義することでメッセージ分析の安定を得ることが可能になると考えて良いのではないかと思える。

参考文献

[1] Baayen, R. H.: 2001, *Word Frequency Distributions*, Kluwer Academic Publishers, Dordrecht.

[2] Siepmann, D.: 2005, Collocation, colligation and encoding dictionaries. Part 1: Lexicogical aspects, *International Journal of Lexicography*, vol.18 No.4, 409-443.

[3] Siepmann, D.: 2006, Collocation, colligation and encoding dictionaries. Part 2: Lexicogical aspects, *International Journal of Lexicography*, Vol. 19 No. 1, 1-39.

[4] Lebart, L., Salem, A. and Berry, L.: 1998, *Exploring Textual Data*, Kluwer Academic Publishers, Dordrecht.

[5] Sinclair, J. M.: 1966, Beginning the Study of Lexis, *In Memory of J. R. Firth*, 410-430, Longmans, London and Beccles.

[6] Sinclair, J. M and Daley, R.: 1970, *English Collocation Studies: The OSTI Report*, Continuum, London.

[7] Sinclair, J. M.: 1974, English Lexical Collocations: A study in computational linguistics, *Cahiers de Lexicologie*, 15-61 .

[8] Kruskal, J. B. and Wish, M.: 1978, *Multidimentional Scaling: Quantitative Applications in the Social Sciences*, Sage Publications, Beverly Hills.

[9] 九鬼周造（1979）『「いき」の構造』岩波文庫.