第１章　類似文書抽出による自由記述分析と授業改善について

松　田　　健[1]

概　要

サイバー大学（以下，本学という）ではFD活動の一環として学生による授業評価アンケートを活用とした授業改善の取り組みを行っている。自由記述の分析では大量のデータを扱う必要があるため，このような分析を支援するシステムがあると便利である。自由記述の分析にはテキストマイニングを用いた様々な手法が考案されており，自由記述の実際のアンケートデータの分析に利用されている。本小文では，集計した自由記述データに対して類似性の高い文書を抽出するアルゴリズムを提案・適用して分析を行った上で，これらの情報を用いてどのような授業改善が可能であるかということについて考察を行う。なお本研究では，提案アルゴリズムを用いて本学で実施している自由記述形式アンケートの分類・整理した上で授業改善の支援を行うことを目的としているため，他の既存のアルゴリズムとの有効性に関する比較評価ついてはこの小文では述べないこととする。

1. はじめに

「学生による授業評価」というキーワードをネットで検索すると，各大学がどのようなアンケートを行っているかという情報や，学生アンケートの内容を教員はどのように捉えるべきかという情報を得ることができる[1, 2]。本学においても教員各自の授業改善を目的として学生による授業評価アンケートを実施しており，本学のFD委員会で授業評価アンケートの結果を収集・集計し大学ホームページ上で公開している[３]。また，授業評価アンケートの自由記述に書き込まれている内容から本学での授業運営において役立たせたいものを集めて作成したtips集「学生の声を活かしたサイバー大学ティーチングティップス集」を作成し，教員に配布するなどの取り組みも行っている。本学はすべての授業をオンラインで実施しているオンライン大学である。以下，オンライン大学において教員はどのように授業を行い，どのようにして授業改善に取り組んでいるかということについて述べるために簡単に本学の授業運営システムについて紹介を行う。

本学のようなオンライン大学では，学生はインターネットを利用して授業を受講している。学生の受講に関する情報はLMS（Learning Management System）で一元管理されており，教員はLMSにアクセスすることで学生の学習進捗状況をリアルタイムに確認す

1　IT総合学部講師

ることができる。また，学生は LMS に設置されている掲示板を用いて教員に質問（Q & A 掲示板）を行ったり，他の学生と議論を行ったりコミュニケーション（ディベートルーム）をとったりすることができる。さらに，大学独自の SNS（Social Networking Service），e-mail, Skype, 手紙などによって学生・教員間でコミュニケーションをとっている。本題の授業評価アンケートについても本学の LMS 上で実施されている。授業評価アンケートは各授業の最終回の受講の前に回答し，その後授業視聴・期末試験受験が可能となる。授業評価アンケートの内容は 2 章で紹介するが，5 段階評価で回答する項目が 17 個，自由記述で回答する項目が 4 個の全 21 問で構成されている。5 段階評価の項目についてはアンケートデータの集計を行うことは容易であり，データをグラフなどで表して可視化できるといったメリットがある。しかしながら，回答の選択肢が「良くない・分からない・難しい」といった単純な感想であることが多いため，具体的な改善策を立てにくいという側面をもっている。これに対して，学生の生の声である自由記述からは例えば以下のような改善を行うための手掛かりとなる具体的な情報を得ることができる。

・授業のどの部分に満足し，どの部分に不満をもっているか
・授業のどの部分が分かりやすく，どの部分が分かりにくかったか

自由記述の分析を行うには大量の回答データを取り扱わなければならない上に，自由記述の回答は 5 段階評価のように数値データとして得られるわけではないため，分析・集計を行う際にはどうしても人の手で行うことが必要になる。このような分類を行う必要がある自由記述データが大量にある場合は，テキストマイニングの手法を取り入れることがしばしば行われている。自由記述データの分析をする場合，それらのデータの特徴を表すいくつかの項目を定義し，その項目ごとにデータを分類していくという手法がよく用いられる［4, 5, 6］。また，テキストマイニングの方法としてベクトル空間モデルを用いて文書の類似度を測る手法も提案されている［7］。テキストマイニングを行うためのツールについては有料のものとして SPSS, 無料のものとして Chasen や MeCab, RMeCab などが有名である。しかし有料のものは商用のために開発されていることが多いため，教育に関するデータの分析ツールとして相応しくないという指摘もある。また，教育現場での使用を想定しているツールとしては Just Systems 社の MiningAssistant/R2 という有料のものも開発されている。

本小文では自由記述文の分析を支援するために，統計的手法を用いて異なる 2 つの文の類似度を測るアルゴリズムを提案し，アルゴリズムのアンケートの自由記述分析への適用方法とそれを用いた授業改善に関する考察を行う。なお，自由記述の分析には構文解析を行うための無料のツールである RMeCab を用いた。以下，第 2 章で授業評価アンケートの実施方法・内容・回収結果等について紹介し，第 3 章で自由記述の分析を支援するアルゴリズムについて概説し，第 4 章で自由記述に提案アルゴリズムを適用した結果について考察を行う。その際に LMS から得られる情報について紹介を行い，授業改善を行うため

の方法についても考察を行う。最後に第 5 章でまとめと今後の課題について述べる。

2. 学生による授業評価アンケート

本学では全ての開講科目について授業評価アンケートを実施しており，学生はこのアンケートに回答することで期末試験の受験ができるシステムとなっている。アンケートの内容については科目区分（講義・演習・卒業研究）によって若干異なる部分があるがほとんどの項目は同じ内容のため以下のようなアンケートを実施している。

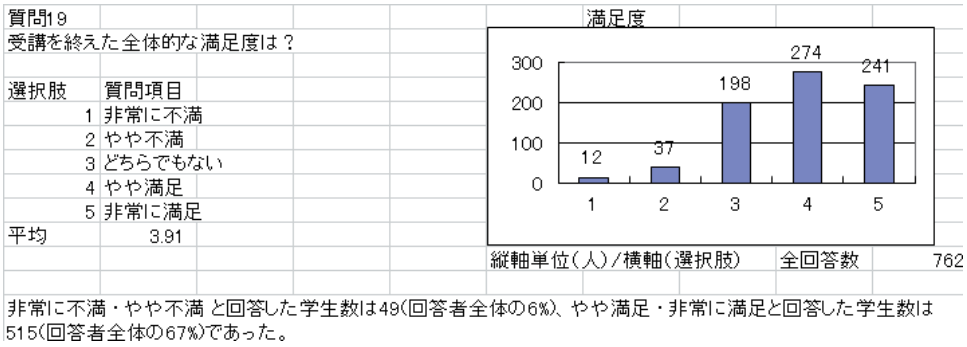

図 1　2010 年度春学期基礎講義　質問 19 回答集計結果

表 1　アンケート項目（5 段階評価）

番号	質　　問	講義	演習	卒業研究
1	授業の学習目標が理解できた	○	○	○
2	授業の内容が理解できた	○	○	○
3	授業を通して新たに専門的知識を習得できた	○	○	○
4	スライドの文字は見やすかった	○	○	
5	スライドの画像（図・写真）は見やすかった	○	○	
6	スライドの内容は授業の理解に役立った	○	○	
7	音声の大きさは適切だった	○	○	
8	教員の話し方の速さやテンポは適切だった	○	○	
9	学習の理解に役立つ学習資料の提供があった	○	○	○
10	教員の教育に対する熱意が感じられた	○	○	○
11	教員は学生の質問に適切に答えていた	○	○	○
12	メンターは学生の質問に適切に対応していた	○	○	○
13	メンターの励ましは学習の継続に役立った	○	○	○
14	ディベートルームは発言しやすい雰囲気だった	○	○	○
15	授業を受けて，より関心が高まった	○	○	○
16	授業内容について，自分自身でもさらに深く調べた	○	○	○
17	小テストの難易度は？	○		
18	演習課題の難易度は？		○	○
19	授業コンテンツ（ビデオ・スライド）の学習内容量は？	○		
20	演習課題の量は？		○	○
21	受講を終えた全体的な満足度は？	○	○	○

表 2　アンケート項目（自由記述）

番号	質　　問
22	よかったと思う点や来学期も続けてほしいと思う点があれば，ご記入ください。
23	来学期は改善してほしいと思う点があれば，ご記入ください。
24	その他ご意見があればご自由にご記入ください。

学生は，講義科目については 1〜17, 19, 21〜24 の 22 項目，演習科目・卒業研究科目については 1〜16, 18, 20〜24 の 22 項目の回答を行う。質問 1〜21 は 5 段階評価で回答する方式であり，項目 22〜24 は自由記述で回答する方式となっている。アンケートの内容は教養科目・外国語科目・専門科目で共通のものとしている。なお，講義科目については項目 1〜17 が質問 1〜17, 項目 19 が質問 18, 項目 21〜24 が質問 19-22 に対応し，演習科目・卒業研究科目については項目 1〜16 が質問 1〜16, 項目 18 が質問 17, 項目 20〜24 が質問 18〜22 に対応する。図 1 は FD 委員会による報告書[3]から授業の満足度に関する回答の集計結果を抜粋したものである。

学生の授業評価アンケートの結果は，各科目毎に集計したものについては担当教員に個別に配布し，科目区分毎に集計したものについては大学ホームページに公開されている。本学の教員はこれらの情報をもとにその学期の授業運営について振り返りを行い，次学期の目標をたて大学事務に報告を行っている。

3. 類似文書抽出アルゴリズム

この章では，学生の授業評価アンケートのうち自由記述の分析支援を行うための手順について概説する。なお，付録 A で本章で利用する統計量 χ^2 値の数学的な性質に関するまとめを行う。

近年，このような自由記述の分析手段としてテキストマイニングを用いた方法が広く利用されている。テキストマイニングを行うためのツールやその使い方について様々な手法が考案されている[5]。本小文で紹介する方法もそのようなテキストマイニングの一手法である。まず計算に使用する記号や関数の定義を行う。

アンケートデータを d, 分析を行うアンケートデータを集めてできる集合を D とする。また，アンケートデータの集合 D に含まれる単語を w とし，各単語の出現頻度を $|w|$, アンケートデータ d に含まれる単語数を $|D|$, アンケートデータの集合 D に含まれる単語の総数を $|D|$ で表す。さらに集合 D から重要語 x を抽出し，重要語からなる集合 X と次の 2 つの関数を定義する。

$$f(w) = \sum_{x \in X} \frac{(x(w) - |w|_D |x|_D)^2}{|w|_D |x|_D}$$

$$F(d) = \sum_{w \in d} f(w)$$

ただし，

$$|w|_D = \sum_{w \in D} |d|$$

$$|w|_D = \sum_{w \in d} \frac{\Sigma_{x \in d}|d|}{|D|}$$

で，$|w|_D$ は d を含む回答データの単語数の合計，$|x|_D$ はアンケートデータ全体の単語数に対する x を含む回答データの単語数の合計の割合，$x(w)$ は単語 w と重要語 x の共起頻度を表す。

以下，分析を行う手順について述べる。

表 3　類似文書抽出アルゴリズム

Step	作　業　内　容
【Step 1】	分析するデータを集めて構文解析し，各単語の出現頻度を計算する
【Step 2】	データに含まれる単語を構文解析ソフトの辞書にある単語に置換する
【Step 3】	データの回答数，各回答に含まれる単語数とその合計をそれぞれ数える
【Step 4】	出現頻度の高い単語から重要語を抽出する
【Step 5】	重要語とデータに含まれる単語の関連度を測る指標 $f(w)$ を計算する
【Step 6】	すべての $d \in D$ に対し，$F(d)$ の値を計算して降順に並べ替える

【Step 1】はテキストマイニングで必ず行われるもので，アンケートデータの集合 D に対して形態素解析を行い，D に含まれる単語を品詞別に分類してそれぞれの単語が D の中に何回出現したかということを数える。

【Step 2】では，日本語のもつ曖昧さによって同じ意味の文章でも微妙な表現の違いがうまれてしまうため，それを少なくするために行う作業である。

【Step 3】は【Step 5】以降で扱う統計データを扱うための準備である。

【Step 4】ではアンケートの質問内容に応じて重要であると考えられる単語の選定を行う。しかしながら，出現頻度が多くても質問項目の特徴となる重要語であると考えにくいものも多数含まれる場合もあるため重要語の選択には注意が必要である。

【Step 5】では χ^2 値と呼ばれる統計量の計算を行う。

式中の $|w|_D|x|_D$ は，単語 w が出現したときに単語 x が w とペアになって出現する頻度の期待値に値し，2 つの単語の関連性を理論値として表現したものである。この理論値は分析を行うアンケートデータ全体から得られるものであるため，分析を行うデータに依存して変化する数値データである。$x(w)$ は 2 つの単語 w と x の実際の共起頻度であり，実際の頻度と頻度の理論値の差異を χ^2 値を用いて測る。

以下，本研究で χ^2 値を用いる理由について説明する。2 つの異なる単語 A, B がある文章に含まれているとする。もし単語 A と B が何らかの意味的な関連性をもたなければ，その文章における単語 A と B の出現頻度はその文の中でほとんど同じであると考えられる。一方，単語 A または B のどちらかがその文章の特徴を表すキーワードであり，かつ単語 A と B が何らかの意味的な関連性をもっているとすると重要なキーワードは頻繁に

出現するという意味で 2 つの単語 A と B の出現頻度には偏りがあると考えられる。そのような偏りを見つけるために本研究では χ^2 値を用いる手法を採用した。

【Step 6】では【Step 5】の計算結果を用いて各アンケートデータを数値データとして扱い，降順にソートしてアンケートデータの分類を行う。

4. アンケートデータへの提案アルゴリズムの応用

本章では，前章で提案したアルゴリズムを実際の授業評価アンケートに適用した実験結果についてまとめる。実験条件は次の通りである。

［実験条件］
使用したデータ：2011 年春学期専門基礎科目質問 21, 22
履修登録者数：564 人
対象科目数：24 科目
質問 21 の回答数：184
質問 22 の回答数：144
構文解析ソフト：MeCab, RMeCab

自由記述のデータを RMeCab で構文解析したあと，重要語の抽出を次のように行った。

・出現頻度の高い単語から選択
・形容詞，名詞，動詞，感嘆詞以外の品詞をほとんど除外

4. 1. 実験結果

4. 1. 1. 質問 21 への適用結果

質問 21 は，授業で良かった点やこれからも続けて欲しい内容について記述する項目である。質問 21 における各単語の出現頻度は以下の通りであった。

これらの単語に対して「理解，講義，できる，良い，説明，聞き取る，解説，ありがとう」などの単語を含む 25 個の単語を重要語として抽出し，前章で紹介したアルゴリズム

表 4　質問 21 における単語の出現頻度の分布

単　　語	出 現 頻 度
す　る	107
思　う	72
講　義	43
やすい	42
良　い	39
とても	37
授　業	36
できる	32

を適用した。質問 21 は, 学生が授業の良かった点について記述を行うためアンケートデータには基本的に良いことが書かれているはずである。実験結果は【Step 6】で行う計算 $F(d)$ の値が大きければ良いコメントが集められていたが，$F(d)$ の値が小さいものに対しては良いコメントではあるが授業に対する要望が含まれている回答が多く含まれるといった結果を得ることができた。$F(d)$ の値が大きなものの中には例えば次のようなコメントが多く含まれていた。

・講義のテンポが良かった
・わかりやすい構成だった
・レポートに対して素早い評価をいただきありがとうございました
・例題が良い

一方，$F(d)$ の値が小さいものの中には

・最新の情報を取り入れて欲しい
・資料が複雑なものを見やすくして欲しい
・アプリケーションについてもっと中身が知りたかった

など改善を要望するコメントが含まれていた。$F(d)$ の値が近いコメントでは，完全ではないものの同じような意味をもつ傾向があることも確認することができた。

4. 1. 2. 質問 22 への適用結果

質問 22 は，授業で改善をして欲しい内容について記述する項目である。質問 22 における各単語の出現頻度は以下の通りであった。

表 5　質問 22 における単語の出現頻度の分布

単　　語	出　現　頻　度
す　る	180
思　う	89
あ　る	84
授　業	47
理　解	33
な　い	29
できる	28

これらの単語に対して「理解，授業，できる，ない，説明，問題，多い，わかる，にくい」などの単語を含む 20 個の単語を重要語として抽出し，前章で紹介したアルゴリズムを適用した。質問 22 の実験結果では，$F(d)$ の値を降順に並べたときに顕著な特徴があるとは言えない結果になった。しかしながら，$F(d)$ の値が小さいデータでは科目特有の内容の書き込みが多いという傾向をみることができた。

4. 2. 授業改善への活用

4. 1. 1 と 4. 1. 2 の結果から 3 章で紹介したアルゴリズムにより $F(d)$ の値を計算した結

果を降順に並べることで，ばらばらに並んでいる自由記述回答データをある程度意味的な繋がりをもったデータの塊に整理することができることがわかった。実際に並べ替えを行ったデータを tips 集の作成を担当したスタッフなどに確認してもらったところ，tips 集のような資料を作成する上で役に立つのではないかという感想を得ることができた。このようにデータを整理してみたところ，授業内容に関わる要望や授業コンテンツに関する意見など，授業改善に利用できそうな学生の意見も多数みられた。これらの情報の中には日ごろの授業運営の中，例えば LMS での学習資料の配布やディベートでの書き込みなどの対応などで解決可能な問題は多く見受けられる。

この小文では，授業改善に役立つ情報を学生による授業評価アンケートの自由記述回答を分析から得ることを目的としているが，LMS に記録されている他のデータも授業改善のための参考情報となるものが多く存在している。例えばディベートルームでは，学生が他の学生にとっても有益となる情報を書き込んだり，こういった資料が欲しいという具体的な書き込みがされることもある。ディベートルームの書き込みと自由記述の回答データを比較してみると，推測の範囲であるが「ディベートルームにも自由記述にも回答している学生」,「ディベートルームには書き込みをしないが自由記述には回答する学生」が存在するのではないかと考えることができる。これらの情報の中にも授業改善に利用可能な情報が埋もれている可能性が高いためディベートルームの書き込みについて分析を行う必要性は高いと言える。

また，2011 年の春学期から IT 総合学部では新しいカリキュラムがスタートしたことにより，解説を聞きながら学習する VOD（Video on Demand）型の授業コンテンツだけでなく学習者自身が教材を読みながら解説の動画や音声の視聴を行う WBT（Web Based Training）型の授業コンテンツを利用した科目も開講されている。WBT 型の授業コンテンツは数学系の科目で利用されており，数学系の科目ではこの形式の授業コンテンツで学習したいと答える学生が多数いる一方で，どのように学習を進めれば良いかわからないという意見の学生もおり，今後はすべての学生が利用しやすいコンテンツ制作の方法について検討していきたい。

5. ま と め

本小文では，学生による授業評価アンケートデータ分析を支援するための類似文書の抽出を行うアルゴリズムを提案し，実際のアンケートデータへの適用を行った。アルゴリズムの適用結果から十分な効果が得られているといいにくい面も多いため，アルゴリズムの改良していくことが今後の課題である。またアンケートの自由記述の回答率を見ると，まだ分析を行うために十分と言えるほどのデータが集まっていないとも考えられる。しかしながら，LMS 上にあるデータの分析から何らかの有益な情報を得られる可能性もあるため，そのようなデータの解析を行うことも今後の課題である。

参考文献

［1］ J・ローマン　阿部美哉監訳　塩崎千枝子他訳：大学のティーチング，玉川大学出版部，(1987).
［2］ 田岡智志　渡辺敏正：*Web* システムによる学生授業評価アンケートの実施方法とその検証，電子情報通信学会技術研究報告．LOIS, ライフインテリジェンスとオフィス情報システム：IEICE technical report 110（450），pp. 185-190（2011）.
［3］ サイバー大学 FD　委員会平成 21 年度　授業評価アンケート調査結果，http://www.cyber-u.ac.jp/outline/self-check.html.
［4］ 飯塚雄一　ケイン・エレナ　小玉容子　松本亥智江：テキストマイニングによる短期海外研修の自由記述の分析，『総合政策論叢』第 17 号　島根県立大学　総合政策学会，(2009).
［5］ 石田基広：*R* によるテキストマイニング入門，森北出版，(2008).
［6］ 渡辺智幸：自由記述文書データからの知識発見手法に関する研究，http://www.it.mgmt.waseda.ac.jp/mi-tech/activity/student1/0681038.pdf.
［7］ Schutze, H., Hull, D. A., and Pedersen, J. O.: *A comparison of classifiers and document representations for the routing problem*, in Proc. ACM SIGIR '95（1995）.
［8］ Ferguson T. S.: *A Course in Large Sample Theory*, Chapman & Hall（1995）.

A　付録：χ^2 値について

この小文では単語 A, B が何らかの関連性をもつかどうかを χ^2 値を用いて評価を行った。このような手法は文献[7]でも利用されいるが，ここではこの小文においてどのような考え方に基づいて χ^2 値を利用したかということについて簡単にまとめておく。また，参考のために確率変数 χ^2 値の和が χ^2 分布に従うことの導出過程の紹介も行う。r 個の重要語 $x_1, x_2, \cdots, x_r$ と，n 個の任意の単語 $y_1, y_2, \cdots, y_n$ を考える。単語 y_j が重要語 x_i と共起する確率を $p_i (i = 1, 2, \cdots, r-1)$ と定義する。なお，p_i は 3 章では

$$p_i = |x|_D$$

と定義しており，$p_r = 1 - \sum_{i-1}^{r-1} p_i$ と定義すれば

$$\sum_{i=1}^{r} p_i = 1$$

が成り立つことに注意する。いま $i = 1, 2, \cdots, r$ に対して，ある文章における単語 y の重要語 x_i に対する共起頻度を q_i とおくと，確率変数 $(q_1, q_2, \cdots q_r)$ はパラメータ $(p_1, p_2, \cdots p_r)$ をもつ多項分布に従う。したがって，単語 y と重要語 x_i の共起頻度 q_i の期待値 $E(q_i)$ と分散 $V(q_i)$ は

$$E(q_i) = np_i$$
$$V(q_i) = np_i(1-p_i)$$

となる。ここで，q_i の期待値 $E(q_i)$ は単語 y と重要語 x_i の共起頻度 q_i の理論値であると考えると，実際の実測値である共起頻度 q_i と理論値 $E(q_i)$ がどの程度ずれているかということが問題となる。そのような実測値と理論値のズレを測るための指標としてピアソンは次の式で表される χ^2 値を定義した。

$$\sum_{i=1}^{r} \frac{(q_i - E(q_i))^2}{E(q_i)} \qquad (1)$$

χ^2 値は，実測値である q_i の値が理論値である q_i の期待値 $E(q_i)$ の値と近い場合は小さくなり，q_i の値が $E(q_i)$ の値と大きく異なる場合は大きくなるという性質をもつことが定義の式からわかる。ピアソンは χ^2 値が次のような性質をもつという事を示している。

Theorem 1.　上述の定義のもとで

$$\bar{q} = \frac{\Sigma_{i=1}^{r} q_i}{n}$$

は自由度 $(r-1)$ の χ^2 分布に従う確率変数である。

この定理の証明は例えば[8]などにある。ここではこの定理の証明の方針を述べることにする。その際に，次に示す中心極限定理が重要な役割を果たす。

Theorem 2.　確率変数 $q_1, q_2, \cdots q_r$ が独立に同じ分布に従い，$i = 1, 2, \cdots, r$ に対して

$$E(q_i) < \infty$$
$$V(q_i) < \infty$$

であるとする。このとき，

$$\bar{q} = \frac{\Sigma_{i=1}^{r} q_i}{n}$$

とおくと

$$\lim_{n \to \infty} P\left[\frac{\sqrt{n}(\bar{q} - E(q_i))}{\sqrt{V(q_i)}} \leq t\right] = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

が成り立つ。つまり，確率変数

$$\frac{\sqrt{n}(\bar{q} - E(q_i))}{\sqrt{V(q_i)}}$$

は平均 0, 分散 1 である標準正規分布 $N(0, 1)$ に弱収束する。

多項分布に従う確率変数$q_1, q_2, \cdots q_r$ の期待値と分散はそれぞれ $E(q_i) = np_i$, $V(q_i) = np_i(1-p_i)$ であるから q_i を標準化した確率変数は

$$\frac{q_i - np_i}{\sqrt{np_i(1-p_i)}} = \frac{q_i - np_i}{\sqrt{np_i}} \frac{1}{\sqrt{1-p_i}}$$

であり中心極限定理から，この確率変数は平均 0, 分散 1 の標準正規分布 $N(0, 1)$ に従う。したがって，確率変数

$$\frac{q_i - np_i}{\sqrt{np_i}} \qquad (2)$$

は平均 0, 分散 $1 - p_i$ の正規分布 $N(0, 1 - p_i)$ に従う。この式 (2) は実測値と理論値のずれを測るための指標である式 (1) と同じである。これを平均 0, 分散 $1 - p_i$ の正規分布 $N(0, 1 - p_i)$ に従う確率変数 z_i を用いて

$$\frac{q_i - np_i}{\sqrt{np_i}} \to z_i$$

と表すことにする。ここで，ある単語の共起頻度を求めるときに 2 つの異なる単語 A, B のみを考慮し，他の単語 C との共起頻度は考えない，つまり 3 つ以上の単語 A, B, C の共起頻度については考えないものとする。このとき，共分散

$$E\left(\frac{q_i-np_i}{\sqrt{np_i}}\cdot\frac{q_j-np_j}{\sqrt{np_j}}\right)=\frac{1}{n\sqrt{p_ip_j}}\left(E(q_iq_j)-n^2p_ip_j\right)$$
$$=\frac{1}{n\sqrt{p_ip_j}}\left(n(n-1)p_ip_j-n^2p_ip_j\right)$$
$$=-\sqrt{p_ip_j}$$

であることがわかる。以下，確率変数

$$\sum_{i=1}^{r}\frac{(q_i-np_i)^2}{np_i}\to z_i^2$$

について考える。標準正規分布に従う r 個の確率変数 $u_1, u_2, \cdots, u_r$ からなるベクトル $u=(u_1, u_2, \cdots, u_r)$ とベクトル $p=(\sqrt{p_1}, \sqrt{p_2}, \cdots, \sqrt{p_n})$ を定義し，これらのベクトルに対して

$$v=u-(u\cdot p)p$$

を考える。ここで，$u\cdot p$ はベクトルの内積を表すものとする。ベクトル v の i 番目と j 番目の要素はそれぞれ，

$$v_i=u_i-\sum_{k=1}^{r}u_k\sqrt{p_k}\sqrt{p_i}$$
$$v_j=u_j-\sum_{k=1}^{r}u_k\sqrt{p_k}\sqrt{p_j}$$

であり，共分散を計算すると

$$E(v_iu_j)=-\sqrt{p_i}\sqrt{p_j}$$

となる。また，

$$E(v_i^2)=1-p_i$$

も成り立つことから，

$$\sum_{i=1}^{r}\frac{(q_i-np_i)^2}{np_i}\to|v|^2$$

ベクトル v の直交変換により，$|v|^2=w_1^2+w_2^2+\cdots+w_{r-1}^2$ となる標準正規分布に独立に従う確率変数 $w_1, w_2, \cdots, w_{r-1}, w_r=0$ をとることができる。最後の式は，定義より自由度 $r-1$ の χ^2 分布に従う。