『総合政策論叢』第17号（2009年３月）
島根県立大学 総合政策学会

〔研究ノート〕

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

飯塚　雄一
ケイン・エレナ
小玉　容子
松本　亥智江

１．目的

近年、学生の短期海外研修への応募が減少傾向である。高い研修費用をどうするかという検討も重要であるが、どうすれば学生の短期海外研修へ興味や関心を高めてやるのか方策を検討する必要がある。1996年から出雲キャンパスで始まった語学・看護学海外研修及び浜田キャンパス、松江キャンパスでの研修に対する学生の感想文（自由記述）を資料として検討する。

本研究では、短期海外研修の自由記述の報告書を、最近の新しい手法であるテキストマイニングによって検討した。複数年度にわたる大量の自由記述感想文の中から有用な情報を選別し取り出すために、テキストマイニングによって大量の発言内容をコンピュータにより処理すれば、代表的な感想や傾向を把握することができる。また処理された言語は数量化されているので様々な分析にかけることができる。これを活用することで、学生たちの希望や不満などを把握することもできる。また学生たちがどのような収穫を得たかを探

ることもできる。

そこでこれまでに浜田、松江、出雲の各キャンパスで海外研修へ参加した学生たちがどのような感想を述べているかを調べる。つまり出雲キャンパス、松江キャンパス、浜田キャンパスの既存の海外研修報告書の自由記述感想を、使われている単語という側面からイメージを分析する。更に年度やキャンパスによる違いがあるかなども調べる。海外情報がマスコミなどであふれている現状で、実際に海外へ行くことから得られるものが何かを探り、これを学生への動機づけの一助として活用できないか検討する。

2．研究方法

⑴対象

島根県立大学、短大部出雲キャンパス、松江キャンパスで実施された短期海外研修の報告書を対象とする。松江キャンパス1991年度（ワシントン州エレンズバーグとシアトル）、島根県立大学は2007年度（カリフォルニア州モントレー）、出雲キャンパスは1996、1998、2000、2003、2006、2007の各年度（ワシントン州ワナチとシアトル）の報告書をとりあげた。なお、松江キャンパスの記録は、テキストマイニングの分析対象年度1991年に加え、1992、1997、1998、2002、2007、2008の各年度（ワシントン州エレンズバーグとシアトル他）の報告書も参考資料とした。但し、2007年度と2008年度は、学年暦の変更により研修期間が25日間から19日間となり、報告書（日誌）枚数もそれに従い減っている。３キャンパス間の報告書の数（語数）には入手可能性の違いにより大きなばらつきができたり、参加者は女性が多数である、などのため、本報告ではキャンパス、性による違いよりも、主に、全体的な傾向を把握することにした。

⑵方法

質的データ解析はテキストマイニングのソフトウエアであるTMStudio 3.0（数理システム）を用い分析した。テキストマイニングでは、記述された文章を分解し、分解した記述語ひとつひとつを変数とみなし、数量データと同じように扱っている。すなわち、自由記述文から得られたテキスト型データをまず分かち書きし、単語（構成要素）に分ける（形態素解析）。例えば、「私はアメリカに行きました」という自然言語文に形態素解析を実行した場合、「私（名詞）、は（助詞）、アメリカ（名詞）、に（助詞）、行き（動詞）、まし（助動詞）、た（助動詞）」のようになる。さらに互いに依存関係（係り受け関係）にある文節の組を作り（構文解析）、特徴表現分析や文章分類、評判分析などを行った。

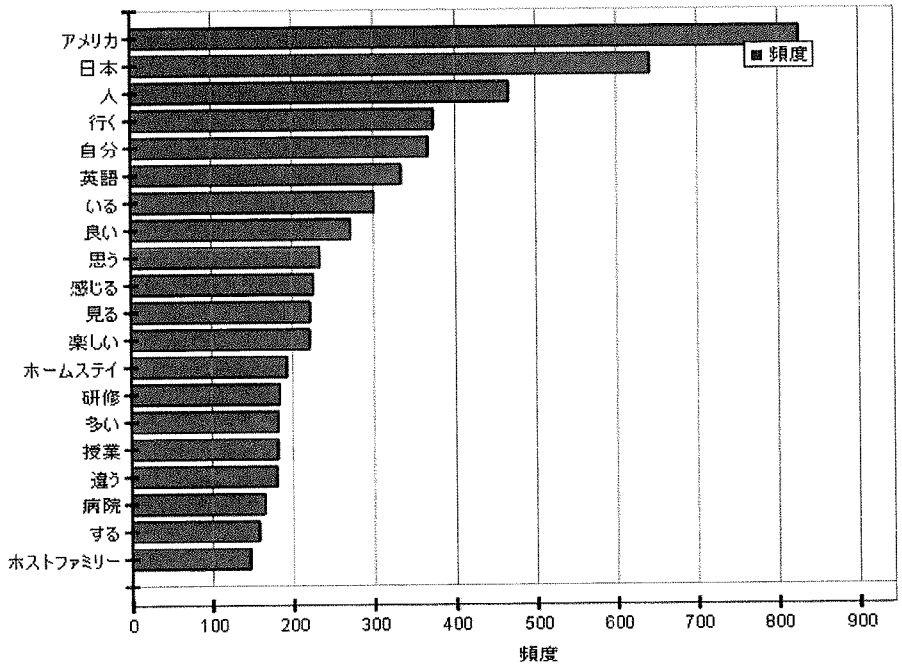
3．結果―頻度分析―

⑴単語頻度解析

テキスト情報の品詞別出現回数では、頻出順に、名詞、動詞、副詞、形容詞、などで感動詞も多かった。

まず、報告書にどういう単語が多く出現しているかをみるために、単語頻度解析を行った（**図１**）。すべての報告書を合わせて最も頻度の多い単語は、「アメリカ」、「日本」、「人」、「行く」、「自分」、「英語」などの単語が多く出ていることがわかる。

図1　単語出現頻度



この頻度解析をもとに、松江キャンパスの記録で使用されている使用頻度の高い語を使用回数という面からさらに比較検討した。**表1**はワード（Word）の検索機能を用い、使用回数を年度別にまとめたものである。使用回数が2桁になっているものを取り上げたが、比較検討のため、同様の語で使用頻度の低いものも表に入れている。

17年間にわたっているが、時の経過による使用語の特徴的変化は見られず、多少の偏りは参加者の属性によると考えられる。感情を表す語としては「楽しい・嬉しい」が、感想を表す語としては「良かった」が高い頻度で用いられている。さらに、研修中の活動としては、「授業」の他に、「話す」「買う」「食べる」ことへの関心の高さがみられる。また、「体験」と結びつくと考えられる「貴重な」は、まず「初めて」体験し、「貴重な」という価値を計る言葉の使用回数は低くなっている。「すごく」という副詞の使用回数の高さも、実体験が心、気持ちに与える刺激の高さを示していると考えられる。

表1　単語出現頻度（松江キャンパス）

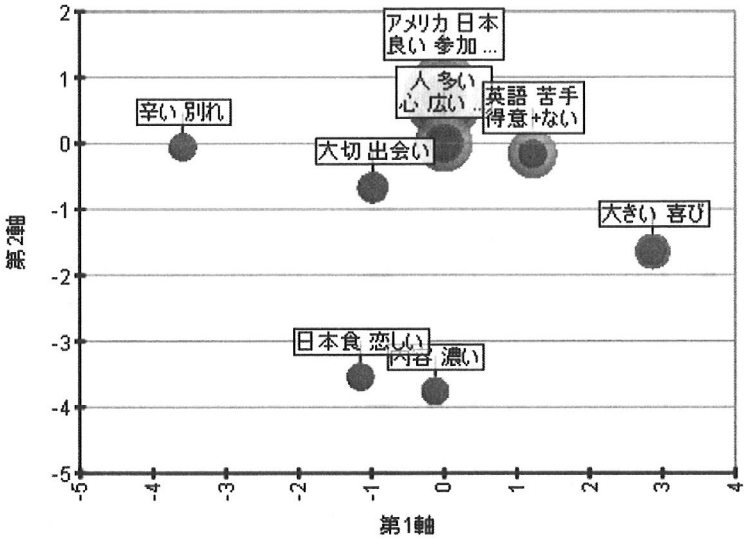
	1991	1992	1997	1998	2002	2007	2008
英語　English	4	6	4	8	15	10	6
アメリカ（人）	16	36	6	11	30	7	14
日本（語）	10	30	13	22	27	15	19
授業	20	35	4	16	27	16	13
楽しい	15	13	13	13	31	24	22
うれし（嬉し）かった	11	7	2	6	11	3	11

良（よ）かった、良（よ）い	18	16	11	11	19	20	20
おもしろかった＋興味深かった＋感動	2+3+2	5+1+1	0+0+0	3+0+0	1+4+7	0+1+2	0+0+2
すごく	17	17	5	14	33	19	9
大きい（な）（でかい,big）	6	6	2	3	15	5	6
小さい（な）	4	0	0	2	3	2	1
いろんな＋いっぱい	4+4	3+4	5+4	0+7	8+5	1+2	6+5
貴重な（大切な）	0	1	1	1	2	1	1
経験（体験）	2	6	4	3	5	5	5
初めて	5	10	9	5	9	6	1
話す（した）	22	15	17	20	39	10	25
買う（買い物、ショッピング）	23	26	19	26	36	8	16
食べ（食事など）	38	24	20	30	40	13	21
店	8	13	9	7	11	5	4
歩く	6	18	1	11	12	0	2
疲れ（て、る）	2	5	11	8	7	2	3
寝て	7	7	12	11	3	6	9
ありがとう（感謝）	1	2	7	0	8	5	6
友達	9	4	3	11	6	3	2
もっと	2	2	0	0	9	6	6
泣く（涙）＋悲しい	1+0	7+4	4+3	0+0	1+0	1+0	3+4

(2)係り受け頻度解析

単語頻度解析では、大体のイメージをつかむために１語に絞ってみた。さらに詳細に内容をみるため互いに依存関係（係り受け関係）にある文節の組を作る。感想文の中に表れている係り受け表現について、係り元単語と係り先単語の頻度を求め、対応分析によりデータを２次元上に配置してみた結果である（**図２**）。対応分析では関連のあるものは、近い点に配置される。「参加は良いもので、喜びも大きく、出会いが大切である」、そして、「別れは辛い」、「英語が苦手」というメッセージが読み取れる。

図２　係り―受けの関係



⑶文章分類

次に、感想文の中のすべての文を、言葉の使われ方が似ているもの同士の５つに分類してみた。文章分類とは、使われている単語に応じて、テキストの行もしくは文章をクラスターに分類し、クラスターごとの特徴単語を表示することである。分類１は「自分が行って見るのは良い」というクラスター（**図３**）、分類２は、「アメリカへ自分が行く」というクラスター（**図４**）、分類３は、「英語の授業やホームステイは楽しい」というクラスター（**図５**）、分類４は、「人」などの単語を多く含む文章の集まり（**図６**）、分類５は「日本とアメリカの病院は違う」というクラスター（**図７**）があった。

図３　文章分類１―自分

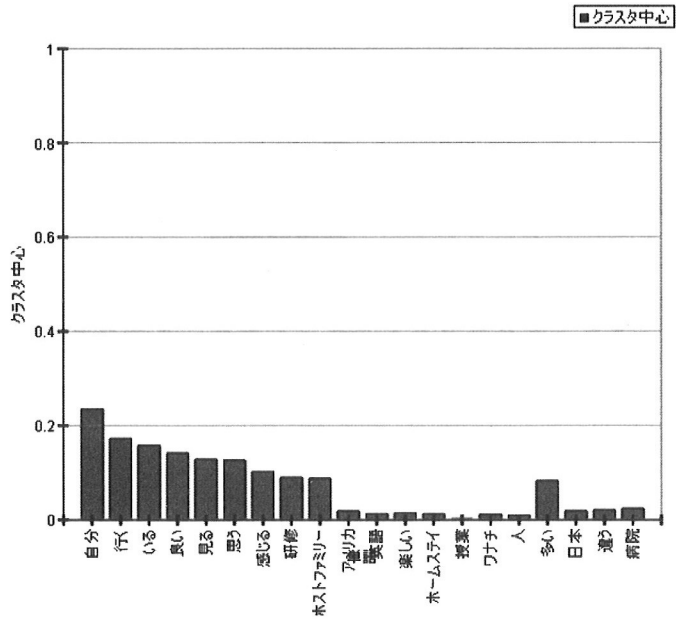


図４　文章分類２―アメリカ

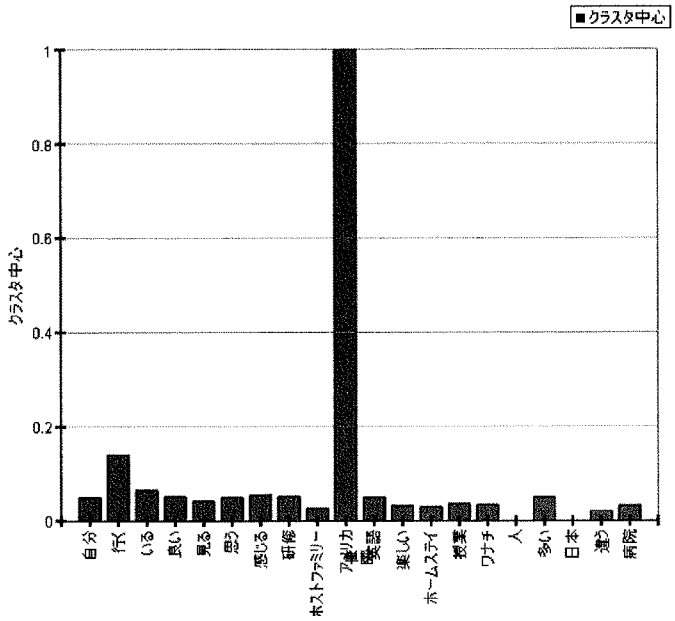


図５　文章分類３―英語

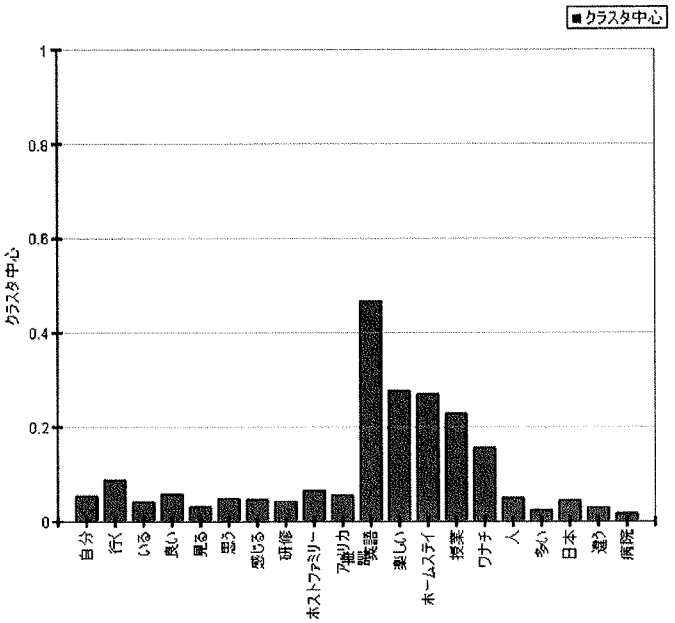


図6　文章分類4―人

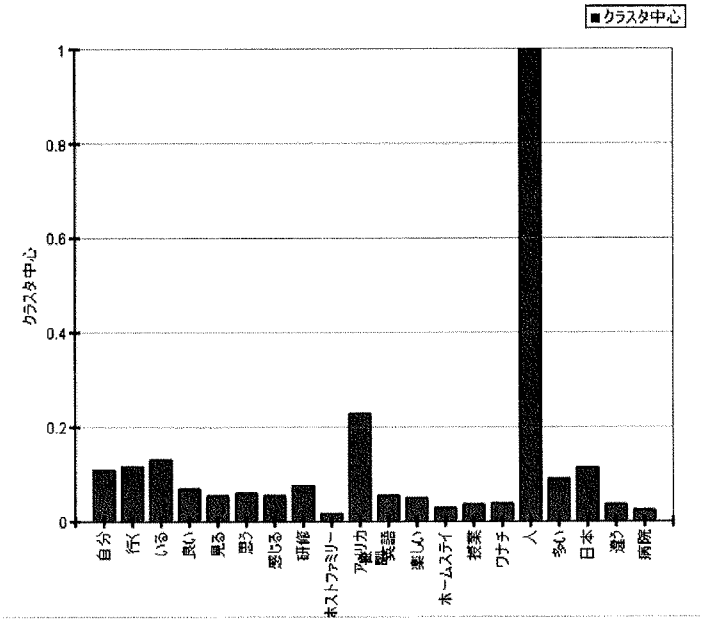
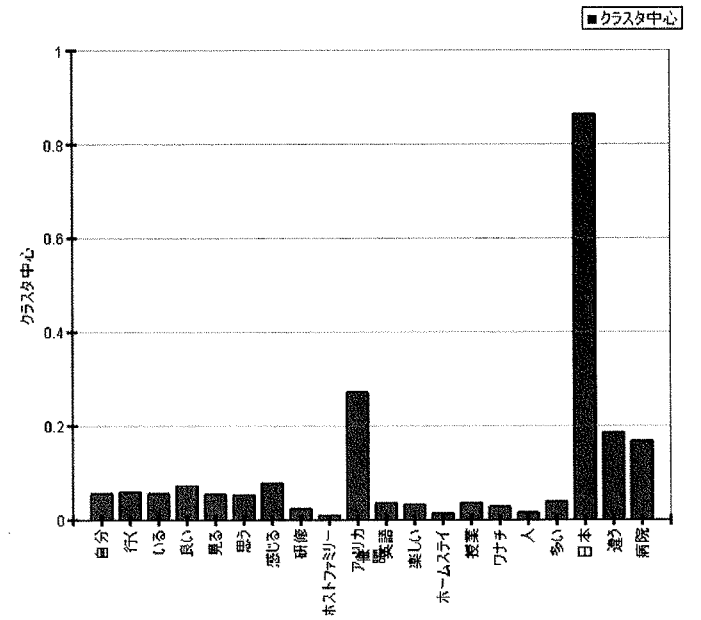


図7　文章分類5―日本



(4)評判分析

次に、良いイメージで語られている単語、悪いイメージで語られている単語を調べてみた（評判抽出)。これは、単語に対して好意的な表現や非好意的な表現それぞれ語られた回数をカウントし、それをもとに、好評語・不評語のランキングを作成する。つまり単語