テキストマイニングによる英語授業に関する自由記述回答の内容分析

阪　上　辰　也
広島大学外国語教育研究センター

1．はじめに

本稿の目的は，「テキストマイニング」という手法を利用して，大学院生向けに開講された英語授業における受講生の反応がどのようなものであったかを示すことである。なお，テキストマイニングとは，「構造化されていないテキストから目的に応じて情報や知識を掘り出す方法と技術の総称」である（石田・金，2014）。

授業を行っている期間中，あるいは，授業がすべて終了した後にアンケートを実施して学生の授業に対する反応を調査する機会は少なくない。満足度などを問う選択式の設問は数値で程度を表すことから集計が容易であるが，自由記述式の回答は言語データが中心となるために計量的な分析を行うことが難しい。例えば，書かれた内容を解釈し KJ 法を用いて分類することがあるが，分類作業に多くの時間を要するため，結果的に十分な集計・分析を行うことが困難である。授業に対して何を感じ，何に満足したか，困っているのかなど，受講生の詳細な反応を知る上で自由記述の回答を分析することは有用と考えられる。

近年になり，テキストマイニングと呼ばれる手法が広く知られるようになり，さらに，自由記述回答を計量的に分析することを可能にする無償のソフトウェアが利用できるようになったことで，医療・金融等の分野のみならず，人文系の分野においても活用され始めている。そこで本稿では，大学院生向けに実施した英語授業の中で継続的に収集した自由記述回答をテキストマニングの手法を利用して単語の使用傾向などを分析し，学生の授業内外での活動状況や授業内容に対する反応の傾向を明らかにする。

2．対象とした授業の概要

調査対象とした授業の科目名は「Advanced English I」であり，その授業計画は，表 1 の通りである。本授業は，TOEIC テストで測定されるような，成人が国際的で日常的な場面において使用する英語運用能力のうち，リスニングとリーディングに焦点を当て，様々な分野の英語運用の基盤となる能力向上を目指している。受講生は，工学研究科の大学院生を中心とし，授業開始時の TOEIC スコアは 500 点程度である。

基本的な授業展開は，事前に指定された範囲分のオンライン教材と単語学習に取り組み，その復習としての小テストを授業冒頭で実施し，その後，聴解・読解のトレーニングとして，ディクテーションや文法問題などの演習を行うというものである。なお，オンライン教材として指定したのは，TOEIC の問題形式を用いた「ぎゅっと e」と呼ばれる問題集と，広大スタンダード語彙リストを学習するための「オンライン単語学習」の 2 つであり，これらを中心とした自学自習を授業の前後で行うように指示している。

表１　Advanced English Ⅰの授業計画

回	主な活動内容
第 1 回	授業の進め方の説明，実力診断，等
第 2 回	授業の進め方の説明，実力診断，等
第 3 回	TOEIC IP テスト
第 4 回	小テスト（L001–072，G001–042，Ca01-a02），演習
第 5 回	小テスト（L073–144，G043–084，Ca03-a04），演習
第 6 回	小テスト（L145–216，G085–126，Ca05-a06），演習
第 7 回	小テスト（L217–288，G127–168，Ca07-a08），演習
第 8 回	小テスト（L289–360，G169–210，Ca09-a10），演習
第 9 回	中間試験（L001–360，G001–210），演習
第 10 回	小テスト（L361–432，G211–252，Ca11-a12），演習
第 11 回	小テスト（L433–504，G253–294，Ca13-a14），演習
第 12 回	小テスト（L505–576，G295–336，Ca15-a16），演習
第 13 回	小テスト（L577–648，G337–378，Ca17-a18），演習
第 14 回	TOEIC IP テスト（効果測定）
第 15 回	小テスト（L649–720，G379–421，Ca19-a20）
第 16 回	期末試験（L361–720，G211–421）

注：L は「ぎゅっと e」の Listening Section，G は Grammar Section を示す。
C は「広大スタンダード語彙リスト」の範囲を示す（合計 6000 語のうち，2000 語が対象）。

上記の活動に加え，プレゼンテーション技能の改善を目指して，さまざまな分野の専門家や著名人が行うプレゼンテーションを公開している「TED[1)]」（http://ted.com）というウェブサイトから各受講生が興味のあるプレゼンテーションを 1 つ選び，そのプレゼンテーションを再現するという活動を取り入れた。ただし，授業内では学期の半ばと終盤に音声の録音をさせるのみとし，必要なトレーニングは授業外で行うように指示した。また，プレゼンテーションを再現するために必要なディクテーションの実施方法や，プレゼンテーションで使われている単語の学習が行えるウェブサイトを授業内で紹介し，自学自習を促した。

３．自由記述の回答文のデータ分析

テキストマイニングを行うためには，大量のテキストデータが必要となる。今回は，過去 2 期分の授業で得た自由記述の回答データを利用する。次節以降で，その収集方法や処理方法について説明する。

3.1. データの収集

原則として，毎回の授業内で自由記述の回答データを収集した（欠席者には授業時間外に回答を送信するように指示した）。受講生が大学院生であり，コンピュータの利用に支障はないと判断し，データ収集にあたっては，Google Drive のフォーム作成機能を利用して，トレーニングを通じての感想や困難に感じたことなど報告させた。さらに，報告時の注意事項として，「特になし」

という事実上の無回答をしないようにすること，また，トレーニングを行わなかった場合でも，その原因や対策を書くように文言を添えて，多くの回答を得られるようにした（図 1 参照）。

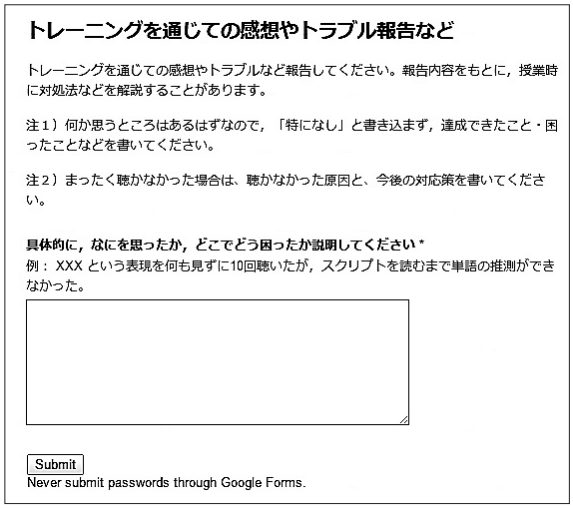

図 1　Google Drive を利用した回答の入力フォーム

3.2. データの加工手順

まず，フォームを経由して受講生によって送られた回答データは Google Drive 上に保存されており，そのデータを Excel 用のファイルとして書き出して保存する（図 2 参照）。直接 CSV 形式での出力も可能であるが，当該の列のみをコピーすることができないため，一時的に Excel 形式で出力した上で，必要な列のみをコピーする手順をとった。

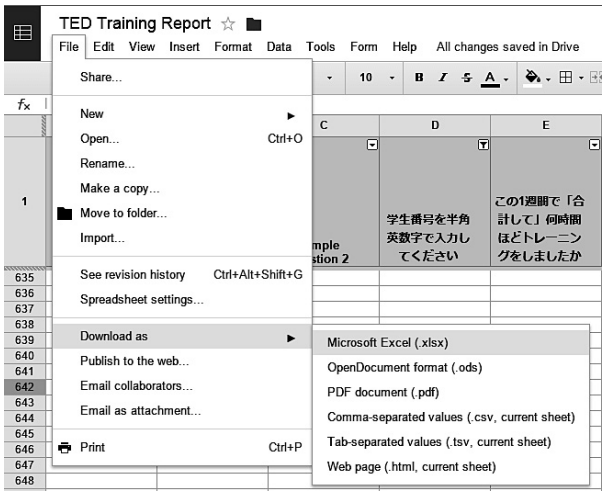

図 2　回答データを Excel 形式のファイルで出力するための画面

次に，ファイル内には，送信された日時や他の質問項目の回答が記録された列も含まれているが，必要な自由記述の回答が記録された列のみを選択して，テキストエディタにコピーする。本研究では，日本語のテキストデータのみを分析対象とするため，留学生が英語のみで記したデータは削除した。

続いて，データの内容を確認しやすくするために，1 行に対し 1 文が並ぶように整形を行った。ここでは，サクラエディタ[2] という無償のテキストエディタの置換機能を利用し，句読点が生じた部分で改行コードを入れる作業を行った（図 3 参照）。具体的には，置換前の欄に「([。．])」，置換後の欄に「$1￥r￥n」という正規表現[3] を入力し，全角文字の句点として「。」か「．」のいずれかが使われていたら，マッチした句点（正規表現の「$1」で表されたもの）に加えて改行コード（正規表現の「￥r￥n」で表されたもの）を追加するという設定を行う。この置換処理により，句点の直後に改行コードが追加され，結果的に 1 行に 1 文が並ぶ形に整形される。

なお，前述の処理をすると，不要な空行が発生するため，それを削除する処理が必要となる（図 4 参照）。具体的には，行頭を示す正規表現の「^」を利用し，行頭に改行コード（￥r￥n）が存在していれば何もないものに置き換える，つまり，置換後の欄を空欄にしておくことで，改行コードを削除する処理になる。

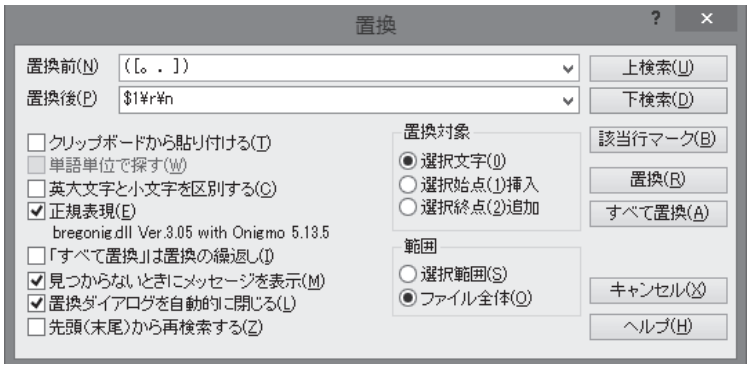

図３　句点を目印に改行コードを挿入するための置換の設定画面

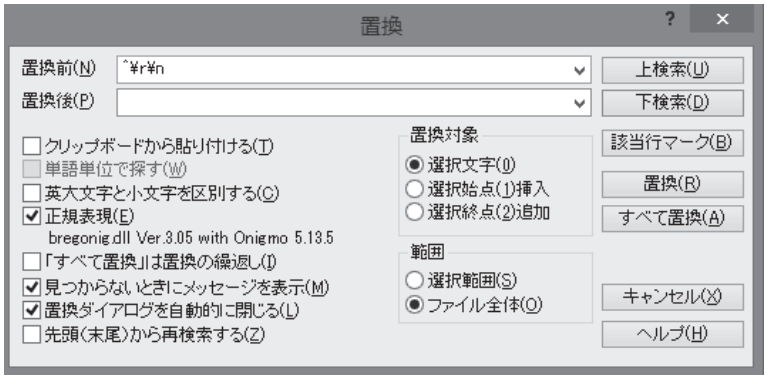

図４　空行を削除するための置換の設定画面

3.3. 分析結果

3.3.1. テキストデータの概要

前節の処理を経て得られたデータファイルの概要は，表 2 の通りである。

表２　回答のデータファイル概要

総語数	15602
異なり語数	1414
文の数	873
延べ回答件数	590

文の数を延べ回答件数で割った値で示される回答 1 件あたり文の数は 1.5 文，また，総語数を文の数で割った値で示される 1 文あたりの単語数は 17.9 語であった。今回のデータでは，5 分程度の時間で自由記述の回答を求めた場合に，2 つ以上の文を回答として書くことが少ないと分かった。

3.3.2. 語彙頻度

データファイルに含まれる単語とその頻度を求めるため，テキストマイニング用のソフトウェアである「KH Coder[4)]」を利用した。頻度順で抽出した際の上位 50 語とその頻度は，図 5 の通りである。

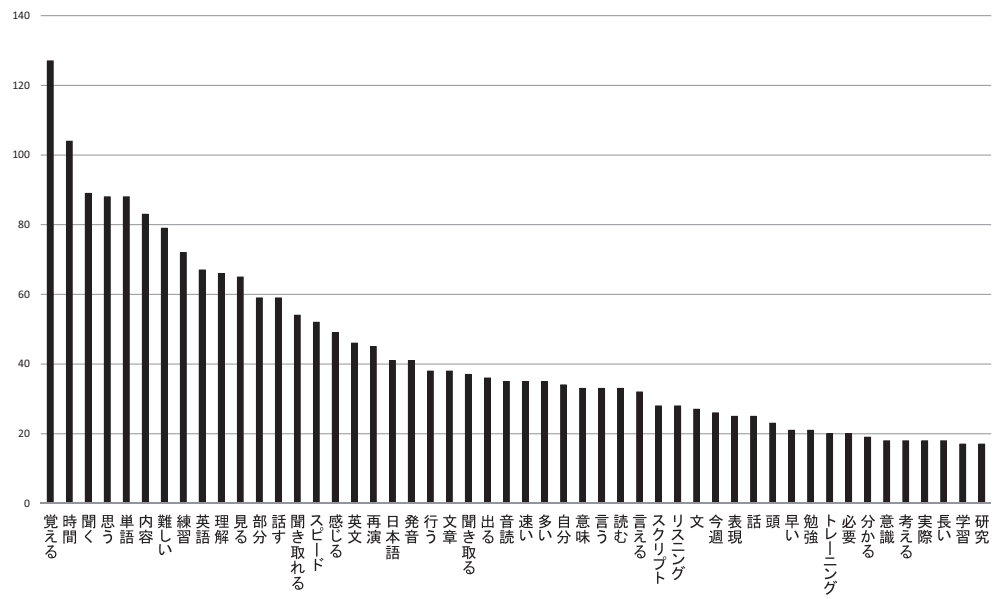

図５　回答データから得られた上位50語の出現頻度

図 5 を見ると，まず，「覚える」という動詞が最頻出の語彙となっていることが分かる。加えて，「聞く」・「思う」・「見る」・「話す」という動詞が上位に現れている。何を覚えようとしていたか，何を話そうとしていたのかなどについては，次節以降で考察する。次に，名詞の「時間」・「単語」・「内容」・「練習」・「理解」が上位に見られるとともに，形容詞の「難しい」も上位にあることが分かる。このように上位にある高頻度の単語同士は回答内で共起している可能性が考えられるが，この点については，次節にてより詳しい分析を行う。

3.3.3. 語彙の共起関係

語彙頻度の観察のみならず，ある語彙がどの語彙と共に用いられていたのかという共起関係を観察することで，どのようなことに対して何を感じていたかを知るための手がかりを得ることができる。まずは，データ全体を対象にして，KH Coder の「共起ネットワーク」を作成する機能を利用し，共起語のつながりを可視化する（図 6 参照）。共起ネットワークとは，ある単語が同じ文中でどの単語と共に使用されているかを図として表現したものである。

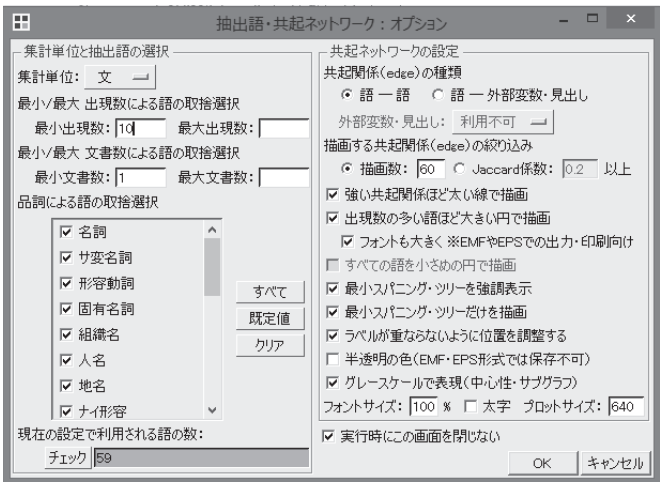

図６　共起ネットワーク作成実行前の設定画面

図 6 に示すような形で，共起ネットワークとして可視化するためのオプションを指定する。注意すべき設定は，集計単位・品詞・描画数である。まず，集計単位を文とした。段落という選択もできるが，共起の範囲が広くなり，共起関係を見つけ出すことが難しくなること，また，今回のデータが回答 1 件あたりにおよそ 1 文という分量であったことから，集計単位を文とした。

次に，品詞については，標準的な設定の場合，否定助動詞（～ない）などの一部の品詞が予め除外されている。データ数の規模が小さいと判断した場合には，より多くの品詞を対象とすることを検討すべきであるが，多くの品詞を対象とすると，より複雑な共起ネットワークが作成されてしまい，特徴的な共起関係を見つけにくくなるおそれがある。今回は，データの規模が大きいものであるとは言いがたいが，より顕著な共起関係を発見できるようにするため，対象とする品詞の範囲を広げず，標準設定のままとした。

最後に，描画数であるが，これは共起関係をいくつ描画するかを設定するものである。品詞の

設定と同様，多くなれば多くなるほど複雑な図となってしまうおそれがあるが，今回のデータは，小規模なものであることから，描画数の値を大きくしても設定した数の描画は行われないと予想されたため，標準設定の 60 のままとした。これらの設定を行って得られた共起ネットワークが図 7 である。基本的に，高頻度で出現した単語は，その円が大きくなる。また，円と円を結ぶ線の太さは共起関係の強さを示している。なお，円の配置はそれぞれの円が重ならぬよう自動で調整されている。

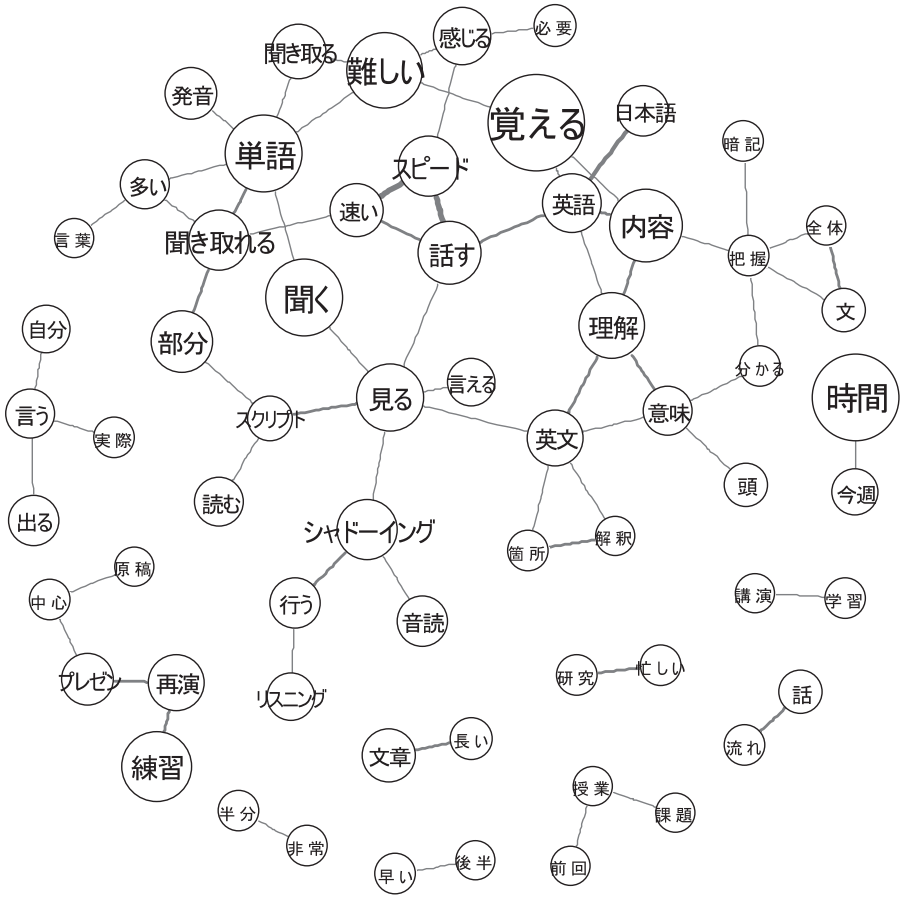

図 7　回答データから得られた共起ネットワーク

図 7 から読み取れることは主に 3 点である。まず，図の上部で最も大きく描かれて位置している「覚える」ことに着目すると，プレゼンテーションの内容や英文そのものを覚えることが困難であったと推察される。実際のところ，内容などを覚えることができたか，できなかったかについては次節にて後述する。

次に，ネットワークの左側中盤にある「聞く」という動詞，また，中心付近に位置している「シャドーイング」・「リスニング」などの名詞が他の単語に比べて大きく描画されていることから，聞く活動を中心としたトレーニングを重視していたことが伺える。

最後に，ネットワークの下部にある「時間」という名詞に共起する単語として，「研究」・「忙しい」という名詞と形容詞が描画されており，トレーニングに取り組めなかった理由が多く書かれたことが分かる。これら 3 点を総合すると，研究活動のためにトレーニングそのものに時間を割くことができず，取り組めたとしても，英文解釈・内容把握・聴く活動に時間を要してしまい，重要な発話のトレーニングに十分に取り組むことができなかったという学生の状況が読み取れる。

3.3.4. 動詞「覚える」の共起表現

「覚える」という動詞が最も高頻度であったことを述べたが，その前後にどのような表現が共起しているのかを観察するため，KH Coder を用いて，KWIC[5] 形式のコンコーダンスラインを出力した（図 8 参照）。なお，動詞の前後の単語でソートした上で出力している。

図8 「覚える」を含むコンコーダンスライン

図 8 から分かるように，「覚える」の右隣に，「（ら）れない」という否定表現が多く出現している。この他にも，「（覚える）のが難しい」・「（覚える）のが大変」・「（覚え）きれず」・「（覚え）にくく」などの否定表現が後続する例が多数を占めていた。このような傾向が見られたことから，受講生の多くは，題材として選んだプレゼンテーションの台詞を覚えることに終始していた可能性が高い。授業内では，台詞を完璧に覚えようとする必要はなく，内容を把握した上で，自らの単語や文法の知識で英文をその場で作り出す感覚を身につけてほしいという旨の助言をしたが，結果的には，選んだプレゼンテーションの語彙・内容・発話速度といった複数の要因が影響し，受講生たちの予想を上回るほど困難なレベルにあったこと，さらに，研究を中心とした生活の中で英語の勉強に十分な時間を割くことができなかったことが重なり，台詞の記憶に終始してしまったのではないかと考えられる。この結果を踏まえ，題材の選び方やトレーニングの方法について，より細かく具体的な指示を与えるべきであったことが反省点として挙げられるだろう。

4．おわりに

本稿では，英語授業における自由記述の回答データに対し，テキストマイニングという手法を用いて受講生の授業内外での活動状況や授業内容に対する反応を調査した。結果として，プレゼンテーションの再現をするという課題に対し，トレーニングのための十分な時間が取れていないこと，また，プレゼンテーションの内容などの難しさが学習を阻害していたことの 2 点が回答データの分析によって浮き彫りとなった。この結果を今後の授業運営の改善に活かすことはもちろん，また，今後も継続して回答データを得ること，さらに，すべての思考を言語化させることは困難であるため，他の質問項目や受講生の属性データと組み合わせることで詳細な分析を行い，どのような状況にある受講生が何に困難さを感じ，何に成功しているかを明らかにすることが今後の課題である。

注

1）TED は，Technology Entertainment Design の略称である。TED は，TED Conference という形で様々な分野・業界の著名人がプレゼンテーションを行う場を設ける組織である。この組織が TED Talks という名でウェブ上にプレゼンテーションを公開するようになり，その内容の専門性やプレゼンテーションとしての質の高さから，語学教材としても注目されている。
2）サクラエディタは，http://sakura-editor.sourceforge.net/ からダウンロード可能である。
3）正規表現とは，特殊文字に意味を持たせ，より少ない文字数で，多くの文字列パタンを表記するための表現方法である。例えば，「覚える」・「覚えよ」・「覚えて」などの動詞の活用形を調べる際に，それぞれの語をそのまま表記せず，「覚え.」のように，ある任意の一文字を意味するピリオドを付けることで，活用形を含んだ表現を一度に検索できるようになる。
4）KH Coder については，http://khc.sourceforge.net/ を参照されたい。
5）KWIC とは，Key Word In Context の頭文字を取ったもので，検索語句を画面中心に並べて共起パタンの発見を支援する表示形式のことである。

参考文献

石田基広・金明哲（2014）.『コーパスとテキストマイニング』共立出版.

ABSTRACT

Analysis of Free Description Data in an English Class Using Text Mining

Tatsuya SAKAUE
Institute for Foreign Language Research and Education
Hiroshima University

The purpose of this paper is to investigate students' reactions to re-presented TED conference presentations. A text mining method was used to analyze the data for this research. Text mining is a way to discover knowledge from text data. It is widely used not only in medicine and finance, but also in the humanities. Free software that allows quantitative analysis of free description responses is available.

In this study, the text data that were collected from the students in an English classroom were analyzed with the "KH Corder" software, which allows processing of the data, calculation of word frequency, extraction of colloquial expressions, and visualization of the data. As a result, the frequencies of particular words such as "time," "remember," "content," and "difficult" were high. This result indicates that the students did not set aside enough time to prepare for the presentation, taking more time for their own research activity than for studying English, and that they had difficulty understanding and keeping up with what the speaker said in a presentation. This means that clear guidance on how to use and learn materials should be provided to students learning languages.