A pilot study to analyze descriptive questionnaire responses with natural language processing techniques

Kohji NAKASHIMA
Tokushima University
Suzanne KAMATA
Naruto University of Education

Abstract

The main purpose of the research reported here is to show that the questionnaire responses in text format can be processed and statistically analyzed to some extent by computer programs. We employ unigram, bigram and trigram frequency calculations with the assistance of Python programming in order to analyze the questionnaire responses in text format.

Questionnaire responses chosen from options can be easily statistically analyzed whether they are on a ratio, interval, ordered or nominal scale. In contrast, descriptive questionnaire responses have long been read and interpreted by researchers themselves. They have been analyzed by human labor alone. This could to some extent lead to arbitrary and subjective interpretations and analyses. We demonstrate in this paper that natural language processing techniques in Python could help analyze questionnaire responses in text format more objectively and accurately.

Keywords: natural language processing (NLP), Python, analysis of questionnaire responses in text format, text analysis, reading-while-listening, Listening-only

1. Introduction

It has been widely recognized that a statistical analysis by using a computer can be applied to questionnaire responses in numerical representation (McKinney, 2018). A computer can process numerical data quickly and accurately. On the other hand, descriptive responses in questionnaires have been manually interpreted so far. Until a few decades ago, a computer was not able to handle text data so well, and language

interpretation was exclusively human work. This could inevitably lead to arbitrary and subjectively determined interpretations of questionnaire responses in text format (McLeod, 2018).

A number of language corpus projects in a variety of languages all over the world have been undertaken along with the development of computers and information technology, which has made it easier to collect language data and to analyze and annotate large corpus data (Römer, 2006; Hasko, 2020). Computers have been getting cheaper and cheaper and faster and faster in the last several decades. In academic fields typically allocated a small budget, such as linguistic studies and language education studies (Benneworth & Jongbloed, 2010), computers have been playing an important and requisite role. In addition, since around 2010, artificial intelligence (AI) with machine learning and deep learning has flourished, and natural language processing (NLP) has also become feasible, not only in natural science fields but also in the fields of linguistics and language education.

In this paper, as a pilot study, we introduce programming techniques in Python and its libraries, in order to analyze the questionnaire responses in a more objective way.

2. Method

The questionnaire we designed was in Japanese because the meaning of each question needed to be fully understood by the respondents, Japanese university students. Accordingly, the students' responses were also in Japanese. They were not required to answer the questions only in Japanese, but some Japanese university students don't have enough language skills to express their own feelings or opinions in English (Fukuzawa, 2016). As a result, all the responses turned out to be answered in Japanese, as shown in the Appendix.

In order to analyze Japanese text by a computer program, it is necessary to conduct morphological analysis in preprocessing because Japanese language is an agglutinative language: words are not separated by a space as in English or other European languages. Thanks to several text processing libraries such as *MeCab*, *Janome*, *nagisa* and *spaCy[ja]*, it is possible to conduct morphological analysis in Japanese much easier than before. In this paper, we employ *MeCab* library for this purpose as it can be easily incorporated (or imported) into a Python program. The result of the morphological analysis of one of the questionnaire responses by *MeCab* library is shown below:

回答 1: 聞き取りが合っているかどうかを確認しながらリスニングできるから。
動詞, 自立, *, *, 五段・ラ行, 連用形, 聞き取る, キキトリ, キキトリ

助詞,格助詞,一般,*,*,*,が,ガ,ガ
動詞,自立,*,*,五段・ワ行促音便,連用タ接続,合う,アッ,アッ
助詞,接続助詞,*,*,*,*,て,テ,テ
動詞,非自立,*,*,一段,基本形,いる,イル,イル
助詞,副助詞／並立助詞／終助詞,*,*,*,*,か,カ,カ
副詞,助詞類接続,*,*,*,*,どう,ドウ,ドー
助詞,副助詞／並立助詞／終助詞,*,*,*,*,か,カ,カ
助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
名詞,サ変接続,*,*,*,*,確認,カクニン,カクニン
動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
助詞,接続助詞,*,*,*,*,ながら,ナガラ,ナガラ
名詞,一般,*,*,*,*,リスニング,リスニング,リスニング
動詞,自立,*,*,一段,基本形,できる,デキル,デキル
助詞,接続助詞,*,*,*,*,から,カラ,カラ
記号,句点,*,*,*,*,。,。,。

The result of the analysis output seems to be rather complicated, so we extract only content words, such as nouns, verbs, adjectives and adverbs for simplification. The extracted words are converted into lemmas, i.e., base forms. The first item in parentheses is lemmas of content words, and the second indicates its part-of-speech information.

('聞き取る', '動詞')
('合う', '動詞')
('いる', '動詞')
('どう', '副詞')
('確認', '名詞')
('する', '動詞')
('リスニング', '名詞')
('できる', '動詞')

With these tuple-formed (lemma, pos) associations, we can calculate the frequency of the collocations with the assistance of a computer: frequency of lemmas (unigram frequency), bigram frequency, and trigram frequency. We calculated these frequencies with the following computer programming code written in Python.

tuples of content words

```
pos_select=['名詞','動詞','形容詞','副詞']
wd_pos_tups = [(nd.feature.split(',')[6], nd.feature.split(',')[0])
    for nd in nodes if nd.feature.split(',')[0] in pos_select]
# count the frequency of lemmas
def count_freq(item_lst):
    frq_dic = defaultdict(int)
    for itm in item_lst:
        frq_dic[itm] += 1
    return frq_dic

# show the lemmas in descending order (more than four in frequency)
def show_freq(freq_dic) -> None:
    print("frq: {:5}\t{}".format('word', 'pos'))
    print("--------------------")
    for itm, frq in sorted(freq_dic.items(), key=lambda x:(-x[1],x[0])):
        if frq > 4:
            print("{:3}: {:6}\t{}".format(frq, itm[0], itm[1]))

frq_dic = count_freq(wd_pos_tups)
show_freq(frq_dic)
```

We can get the frequencies of bigrams and trigrams in the following functions respectively:

```
# tuples without pos restrictions
wd_pos_tups = [(nd.feature.split(',')[6], nd.feature.split(',')[0])
                for nd in nodes if nd.feature.split(',')[0] not in ['記号']]

def bigram(nodes):
    bi_nodes = []
    binode_append = bi_nodes.append
    for wp1, wp2 in zip(wd_pos_tups, wd_pos_tups[1:]):
        binode_append((wp1, wp2))
    return bi_nodes

def trigram(nodes):
```

```
tri_nodes = []
trinode_append = tri_nodes.append
for wp1, wp2, wp3 in zip(wd_pos_tups, wd_pos_tups[1:], wd_pos_tups[2:]):
    trinode_append((wp1, wp2, wp3))
return tri_nodes
```

In order to calculate the bigrams and trigrams, *zip* function will be very helpful and convenient to make a program.

For getting the key words or important words in a text, feature extraction and TF-IDF (Term Frequency-Inverse Document Frequency) by vectorizing the frequencies of each word in corpora are widely recognized as useful, but we do not employ these methods in this study, because the corpus size is not balanced or proportionate: the number of students who responded 'listening-only' was easier was only two (one of which did not respond the reason why descriptively), compared to the number of students who responded the other way around, which was 59.

3. Statistical analysis to the questionnaire responses in choice format

We conducted questionnaires to 64 university students after they experienced two modes of listening experiments: Reading-while-listening and Listening-only. (For details of the experiment, see Nakashima et al., 2018) The following questionnaire responses are the result of the question asking which was easier, Listening-while-reading or Listening-only, i.e., listening with a script, or listening without a script (see Figure 1).

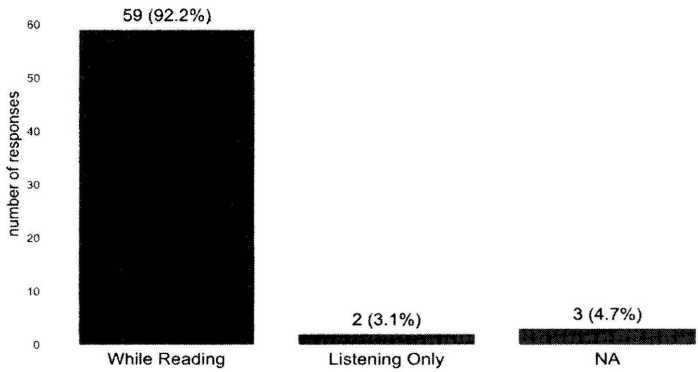

Figure 1. Which was easier, Listening-while-reading or Listening-only?

Figure 1 shows that almost every Japanese university student feels Listening-while-reading is easier than Listening-only. To put it another way, listening to English spoken at a natural speed without looking at a script is much harder for most Japanese university students to understand. Fifty-nine out of sixty-four students responded that Listening-while-reading was easier than Listening-only. Only two students out of 64 responded that they felt Listening-only was easier than Listening-while-reading. Three students did not respond at all, probably because this is the last question of the 27 multiple-choice questions we conducted in the questionnaire and they ran out of time to respond to the question; they did not respond to the last several questions, either. The result is obvious, but just to make sure, we present the contingency table between listening modes and the preference, and the result of Fisher's Exact test, which is a statistical test for the analysis of a 2 x 2 contingency table whose sample sizes are very small, as with this questionnaire.

Table 1. Contingency table between listening modes and their preferences

	easier	others
While Reading	59	5
Listening Only	2	62

# Fisher's Exact Test
p = 1.479703e-27***

The p-value of Fisher's Exact test is significantly low ($p <= 0.001$), which indicates the probability that the categories of the two listening modes are independent of their preferences is significantly low; in other words, the two modes are not equally preferred by the students. The questionnaire results show that the Japanese university students much prefer Listening-while-reading to Listening-only because the former is considerably easier for the Japanese students than the latter.

4. Analysis of the questionnaire responses in text format

Why is Listening-while-reading easier than Listening-only for most students? We asked the reason for each choice in the questionnaire. They responded in Japanese in a descriptive form. In order to calculate the responses by a computer program, we needed to normalize their text responses because some of them included wrong expressions as Japanese and fluctuations of description (allomorphs) such as 'わかる' or '分かる' (to make out), 'おう' or '追う' (to follow), and so forth. We wrote a computer program in Python to

calculate the unigram, bigram and trigram frequencies (see Table 2, 3 and 4) of the base forms of the words in the students' responses. As stated in the previous section, base forms are computed with the help of *MeCab*, which is a Python library for analyzing Japanese text.

Table 2. Word frequency (unigram) of the students' responses (freq >= 5)

frq: word	pos
23: できる	動詞
18: 理解	名詞
16: 分かる	動詞
15: 単語	名詞
15: 聞く	動詞
15: 見る	動詞
14: こと	名詞
13: 読む	動詞
11: いる	動詞
10: する	動詞
9: ある	動詞
8: 文	名詞
8: 文章	名詞
7: 聞き取る	動詞
7: 聞き取れる	動詞
7: 逃す	動詞
6: ため	名詞
6: の	名詞
6: ほう	名詞
6: やすい	形容詞
6: リスニング	名詞
6: 目	名詞
6: 確認	名詞
5: 文字	名詞
5: 英文	名詞
5: 部分	名詞

Table 3. Bigram frequency of the students' responses (freq >= 6)

frq: words

13: できる から
10: た から
10: を 見る
10: 理解 できる
9: が 分かる
9: こと が
8: て いる
7: 分かる ない
7: 聞く 逃す
7: 見る ながら
6: から 聞く
6: が ある
6: ほう が
6: 単語 が
6: 文 を
6: 聞き取れる ない

Table 4. Trigram frequency of the students' responses (freq >= 3)

frq: words

6: 理解 できる から
5: から 聞く 逃す
5: を 見る ながら
4: が ある た
4: こと が できる
4: 聞く 逃す た
3: ある た から
3: から 単語 が
3: から 文 を
3: が 分かる ない
3: が 分かる やすい
3: た から 単語
3: た ほう が

3: だけ だ は
3: て いる か
3: ない から 聞く
3: ない こと が
3: ほう が 分かる
3: やすい た から
3: 文 を 見る
3: 確認 できる から

By using the computer program Python, we were able to compare the frequency of unigrams, bigrams and trigrams with the students' descriptive responses (see Appendix). For example, many students claim they can understand "理解する" or confirm "確認する" the words with the help of reading the script. They also claim that they cannot understand well without the script partly because the natural speed of spoken English is too fast for them to comprehend as shown by the relatively high frequency of "聞く 逃す" in the bigram and the trigram. While the precise reasons for the students' lack of understanding are beyond the scope of this paper, these results suggest areas for further study.

We have shown that questionnaire responses chosen from options can be easily analyzed or categorized if we transform the nominal-scale responses into dummy values represented by 0 and 1. In contrast, descriptive questionnaire responses (text data) have long been read and interpreted by researchers themselves. This could lead to arbitrary and subjective interpretations and analyses.

5. Conclusion

It cannot be denied that computers have been playing an amazing role in analyzing numerical data for decades. Though text analysis by computer is not yet dependable enough, it is now advancing rapidly with the progress of artificial intelligence technology. In order to interpret the descriptive questionnaire responses more accurately and objectively, we need to develop the way to make the most of computers and computer programming techniques. Presenting computer programming codes in a paper will also be helpful to assess the reproducibility of the analyses and interpretations in the study.

References

Benneworth, P., Jongbloed, B.W. (2010). Who matters to universities? A stakeholder perspective on humanities, arts and social sciences valorisation. *Higher Education*, 59, 567–588. https://doi.org/10.1007/s10734-009-9265-2

Bird, S., Klein, E. & Loper, E. (2009). *Natural language processing with Python.* O'Reilly Media, Inc.

Fukuzawa, R. E. (2016). English proficiency and internationalization among Japanese university students. In J. Mock, H. Kawamura & N. Naganuma (Eds.) *The Impact of Internationalization on Japanese Higher Education* (pp. 53-67). Brill.

Hasko, V. (2020) Qualitative corpus analysis. Wiley Online Library. Retrieved from https://doi.org/10.1002/9781405198431.wbeal0974.pub2

Lubanovic, B. (2015). *Introducing Python.* O'Reilly Media, Inc.

Lutz, M. (2014). *Python pocket reference.* O'Reilly Media, Inc.

McKinney, W. (2018). *Python for data analysis, 2nd edition.* O'Reilly Media, Inc.

McLeod, S. (2018). *Questionnaire: Definition, Examples, Design and Types.* Retrieved from https://www.simplypsychology.org/questionnaires.html

Nakashima, K. & Stephens, M. (2017). Do students prefer listening to texts read aloud to them as a group, or to audio-books in solitude online? *Hyperion 63*, 11-25

Nakashima, K., Stephens, M. & Kamata, S. (2018). The interplay of silent reading, reading-while-listening and listening-only. *The Reading Matrix: An International Online Journal, Volume 18, Number 1,* 104-123

Pellegrino F., Coupé, C. & Marsico, E. (2011). A cross-language perspective on speech information rate. *Language 87*, 539-558.

Römer, U. (2006). Where the computer meets language, literature, and pedagogy: Corpus analysis in English studies. In A. Gerbig & A. Müller-Wood (Eds.). 2006. *How globalization affects the teaching of English: Studying culture through texts* (pp. 81-109). Lampeter: E. Mellen Press.

Appendix

Students' questionnaire responses (normalized)

Asterisk in front of the responses indicates the response was by the students who responded Listening-only was easier than Listening-while-reading.

Q27. 英文を見ながらリスニングするのと、見ずにリスニングするのと、どちらが簡単でしたか？

1. 文を見ながらリスニングするほうが簡単だった

2. 文を見ずにリスニングするほうが簡単だった
Q27a. 上の(27)で、そのように答えた理由を説明してください。

(English translation of the questionnaire questions)
Q27. Which was easier, listening while reading or listening only?
1. Listening with a script was easier
2. Listening without a script was easier
Q27a. Please explain why you responded so.

N.B. The following responses are normalized in a way that notational variants are unified (in order to calculate the frequencies of the words and bigrams), and that obvious spelling or grammatical mistakes are corrected. The asterisks in front of the responses indicate that the response corresponds to the second alternative, i.e., "Listening without a script was easier."

Students' responses to Q27a.
- ----
1. 聞き取りが合っているかどうかを確認しながらリスニングできるから。
2. 聞き逃した単語があっても文字を読むことで拾えるから。
3. 文があったほうが分からないときに読むことができる。
4. 自分が聞き取ったことが正確か確認できるから。
5. 何を言っていたのか分からなくても、英文を見ることで単語を拾い理解できるから。
6. 単語のつながりや速さ、難度の高い単語は文章を見ていないと聞き取るのが難しかったから。
*7. 文字を追うのは読むのは長い時間続けられないから。
8. 聞き逃したところも確認することができるから。
9. 音だけでは似たような音の単語の判別ができず、分からないことが多かったから。
10. 単語の発音のつながりが分からないため、英文を見ながらであれば、何の子音が消えているかを理解できるから。
11. 文章を見ながら聞いていたほうが、リラックスして聴けるから。
12. すぐに確認できるから。
13. 言葉の速さと理解のスピードが釣り合っていないので遅れても文章があれば後から理解できるから。
14. NA
15. 文を見ながらだと、聞き逃した部分が分かるため簡単だった。

16. 何を言っているか分かる方が、聞き取りやすい。
17. NA
18. 聞き取れなかった部分も文字を目で追うことで理解できるから。
19. 聞き取りにくい部分でも理解できる。
20. 英語を聞いたとき英語のスペルを思い浮かべてそれを日本語に訳しているイメージなので文を見ながら聞いたほうがスペルを思い浮かべなくてよいぶん楽に訳せる。
21. 聞くだけだと分からない部分もあるから。
22. リスニングだけでは理解が難しかった。
23. リスニングだけではスピードが速すぎて付いていけないことがあった。
24. 視覚的に見ながら音声を聞くことで理解しやすかったから。
25. 単語が聞き取れないことがなくなるから。
26. 文を見ずにリスニングをするのは少し理解が追い付かないから。
27. 聞き逃ししないから。
28. 聞き逃すことがないし聞くより読む方が理解できるから。
29. 内容を聞き漏らしても文章で読み、理解することができたから。
30. 確実に単語を聞き取れるから。
31. 目で追うことができるため。
32. 耳からだけでなく目からも情報を読み取れるから。
33. 意味を考えながら読めるから。
34. 文をたどっていくほうが分かりやすい。
35. まだ未熟なので聞き取れない部分が多かった。
36. NA
37. 聞き取りづらいところも文を読めば理解できたから。
38. 自分の読むスピードと録音のスピードが違うから。
39. 確認できるから。
40. 正しい発音を知らない単語について目で見てスペルから理解できるため。
41. 文字を見ながらのほうが分かりやすかったから。
42. 英文も見ることでどこを読んでいるのかが分かりやすかったから。
43. 文章なら高校までで英文に慣れ親しんでいたが、リスニングになると文脈を理解しながらでないと付いていけない。さらに小説のようなものであると登場人物の名前や登場人物の性格などがとても読みにくかった。
44. 読まれる文章を目視できるから。
45. 単語が分かるから。
46. NA
47. 文章が全部理解できるから。
48. NA

49. 文章の切れ目が分かり、主語を考えながらリスニングできるため。
50. 聞き取りにくい単語があったから。
51. 文字を見る方が音声だけ聞くよりも理解しやすいから。
52. どこ読んでいるか分かるから。
53. 英文が書いてあるから。
*54. NA
55. 聞いて分からない単語でも、目で見て読めば分かる単語が多くあったから。
56. 意味が分からない文が無くなるから。
57. 聞き取れない語があったから。
58. 単語が聞き取れない心配がないから。
59. 目と耳で感じるから。
60. 英語を見ないと理解しづらいから。
61. 聞き取れなかった単語も文を読めば理解できる。
62. 聞き逃したとき、補完できるから。
63. 聞き逃しが少なくなるため。
64. 見ながらのほうが確信をもって確認できる。