IIAI Open Conference Publication Series
IIAI Letters on Institutional Research
Vol. 001, LIR013
DOI: https://doi.org/10.52731/lir.v001.013

Student's Interests and Career Understanding: A Topic Analysis of First-year Career Courses

Tatsuya Tsumagari [*], Yoko Nakazato [†], Takashi Tsumagari [‡]

Abstract

This study conducted a topic analysis of the free-text reports submitted by the students to examine the outcomes of their first-year career courses. There were two types of reports: 1) asking students why they were interested in a lecture, and 2) how they understood careers to be important as an outcome of the course. In the analysis, both reports were used as the data set, and LDA (Latent Dirichlet Allocation) was used to estimate the topic model, and Gibbs sampling was used to estimate the parameters. Students were classified into five types according to the lectures they were interested in. The analysis confirmed that for the 14 topics extracted, each student type had a unique topic that emerged as the reason for their interest. It was also found that many of the student types tended to have a long-term understanding of career as "their life itself." The result of the analysis also found that there may be a reciprocal phenomenon regarding key topics in the students' interest and career understanding.

Keywords: first year experience, career education, actual state of learning, topic model

1 Introduction

With the universalization of higher education, the importance and significance of first-year experience has been well established among university personnel. One of the aspects of first-year experience is career education. This is one of the most important education programs in today's universities, given that many students enter without a clear idea of their prospects. It is expected to not only make new students aware of their careers, but also to encourage them to actively engage in their studies and student life to achieve their future goals. Almost all universities in Japan offer first-year career courses.

What do university students learn about careers from first-year career courses? In the past, most evaluations of course outcomes have been based on comparisons of pre- and post-surveys. However, this does not necessarily reveal the reality of what students have learned. For this reason, the analysis of students' free descriptions has been attempted [1]. In this study, we focused on students' free writing reports and examined the reality of their learning in first-year career courses through topic analysis.

The reality of learning is not uniform across students. It depends on the type of student taking the course as the learning resources provided are absorbed based on the students' interests. Therefore, understanding the types of students and how they learn can provide useful insights for designing first-year career experience. There are several ways to classify student types. This study classifies them according to their interests which strongly influence their learning. We used the

[*] Seigakuin University, Saitama, Japan
[†] Kagoshima University, Kagoshima, Japan
[‡] Prefectural University of Kumamoto, Kumamoto, Japan

lecture information that the students paid attention to. The first-year career course at a regional university in Japan, which was the sample of the study, consisted of seven lectures. We classified the student types using the information of the lectures the students were particularly interested in.

The students were required to submit two types of reports: one, on the lecture they were interested in and why they were interested in it ("interest report"), and the other, on their understanding of career through the entire lecture ("career understanding report"). This study used both reports lumped together for analysis and estimated a topic model. We discussed the reality of learning for each student type based on the topic distribution.

2 Method

2.1 Sample

The survey was conducted on students who took a first-year career course in 2020 and 2021, which was offered in the first semester as a required course at a public university in Japan. A total of 291 students in 2020 and 292 students in 2021 took the course. The students submitted two types of reports after completing all lectures. One of the reports was an interest report, in which the students chose two lectures that were of interest to them and described the reasons for their choices in any number of words in Japanese. The average number of Japanese characters in the interest report was 152 (standard deviation $\sigma = 111$). The second report was the career understanding report, in which the students were asked to describe their understanding of careers throughout the lecture in approximately 600 Japanese characters. The average number of Japanese characters in this report was 614 ($\sigma=65$). In total, each student submitted reports (two interest reports and one career understanding report) with an average of 918 Japanese characters ($\sigma=233$). For the analysis, we used the reports of 572 students, excluding two students with missing data, out of 574 students who volunteered to participate in the survey. The reports were anonymized prior to analysis.

2.2 Method of Analysis

2.2.1 Classifying students by lectures of interest

The sample first-year career courses for this study consisted of seven lectures. These lectures could be classified into three types based on the contents: A) theory type (four lectures), B) self-understanding type (one lecture), and C) role model type (two lectures).

The theory type is a lecture that explains the meaning of learning at university and the significance of career development for the future, and the self-understanding type is a lecture that encourages self-reflection on their current abilities using the result of the test evaluating generic skills. The last lecture, the role model type, provides stories about senior students' and graduates' experiences who are familiar role models for the first-year students. The students were asked to choose two lectures that they found interesting among the seven lectures. Based on the combination of these choices, the students were classified into the five types shown in Table 1.

Compared to the expectation values of the composition ratios when the students' choice probabilities were same for each of the seven lectures, the composition ratios shown in Table 1 were quite different. The expectation values of the ratios are 28.6% for Type I, 19.0% for Type II, 38.1% for Type III, 9.5% for Type IV, and 4.8% for Type V. In this comparison, the

actual ratio of Type I is very low, while that of Type IV and V is very high. Many of the surveyed students were interested in self-understanding and role model type lectures.

Table 1: The student types (n=572)

Type of Student	Combination of lectures that the students were interested in	Number of students	Ratio (%)
I	A: Theory, A: Theory	12	2.1
II	A: Theory, B: Self-understanding	57	10.0
III	A: Theory, C: Role model	172	30.1
IV	B: Self-understanding, C: Role model	229	40.0
V	C: Role model, C: Role model	102	17.8

2.2.2 Topic model

This study examined the reality of the students' learning by separating the interest reports and career understanding reports into the topics. We used Latent Dirichlet Allocation (LDA) to estimate the topic models and Gibbs sampling to estimate the parameters. We also used the topicmodels package of R [2] as a tool for estimating topic models, and Mecab [3] for morphological analysis of the free descriptions.

(1) Preparation of Data Set

For each of the 572 target students, we had free writing data consisting of two interest reports and one career understanding report. We used the entire free descriptions as a data set for topic model estimation

Morphological analysis of the data for 572 students was conducted, and the data was dissected into 315,290 words consisting of 5,096 different words. Using these results as a reference, we pre-processed the text data for dictionary registration, stop words, and synonyms. First, the extraction of words not registered in the dictionary, such as proper nouns, which are necessary for dictionary registration, was done in the same way as in the pre-processing by Tsumagari et al. [4]. The extracted word list of 292 words was registered in the user dictionary. Subsequently, words containing numbers (8 words) and words such as "do" and "myself" (9 words), which appear frequently in all students' free descriptions and may interfere with the search for topics, were selected as stop words. In addition, a dictionary was created to deal with spelling variations, and a unique word was assigned to each word set.

We extracted nouns (general and proper nouns), verbs (independent), adjectives, adverbs, and unknown words that appeared more than once from the pre-processed data. After this processing, we obtained 1,893 words. For these 1,893 words, we created a dataset in which each row represented each free description and each column represented the number of occurrences of each word in the free description, and estimated a topic model. All analyses were conducted in Japanese, and the final notation was converted to English words.

(2) Estimation of Topic Model

Estimation by Gibbs sampling requires setting two hyperparameters of LDA and the number of topics. The hyperparameters are the hyperparameter α for topic distribution and δ for word distribution.

α is related to the uniformity of the topic distribution; the smaller the value, the more the topic distribution deviates from the uniform distribution and the fewer topics get high scores. α is generally set to 50/k (k is the number of topics) as suggested by Griffiths et al. [5].

However, Jacobi et al. [6] argued that it makes sense to define multiple, clearly distinct topics if the data is assumed to include a variety of events and perspectives, and Jacobi et al. used α = 5/k. The free description, which are the sample in this study, are the data comprising many students' consciousness. We found the significance on defining the multiple clearly distinct topics as Jacobi et al. argued, and used α = 5/k. We used a general value of 0.1 for δ[5].

LDA is required to determine the number of topics, which is generally determined by the perplexity. One of the previous studies argued that the mathematically optimal model is the one with the minimum perplexity; however, it is not necessarily the true model [7]. Jacobi et al. [6] proposed to select the number of topics for which the decrease in perplexity value is gradual, considering "complexity" and "interpretability." This study determined the number of topics the same as Jacobi et al.

In this analysis, the data set was divided into training and validation data in the ratio of 3:7, and perplexity was calculated five times each while changing the value of the number of topics k from k=4 to 70 in 2 increments (Figure 1).The solid line in Figure 1 is the regression curve of the polynomial approximation. As shown in Figure 1, the rate of decrease in the perplexity decreases when the number of topics is between 14 and 16. Since it is desirable to have as few topics as possible for ease of interpretation, the number of topics is set to 14 in this study.

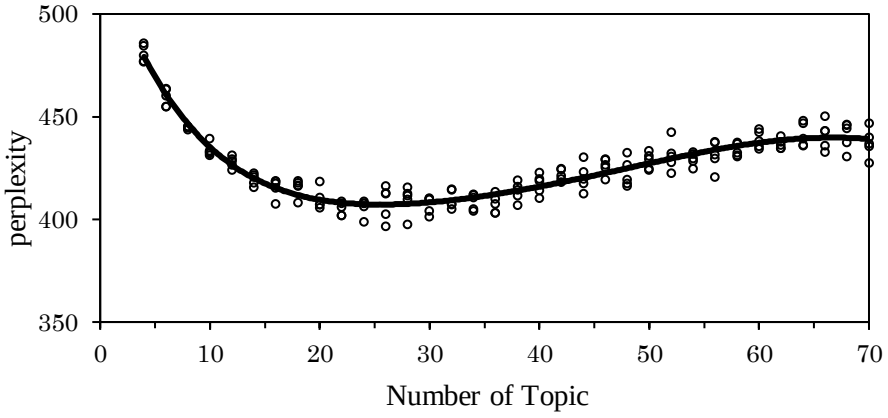

Figure 1: Number of topics and perplexity

(3) Topic Labeling

Topic labeling was based on the top 30 words in the word distribution for each topic and the free descriptions with the highest proportion of applicable topics, with researchers with long career education experience among the authors.

2.2.3 Topic distribution by student type

From the estimated topic model, the topic distribution of the free writing reports is clear. In addition, each free description is associated with information such as student type and type of free writing (interest or career understanding report). Based on this information, we compared the topic distribution of the free descriptions among the student types. The topic ratio for each student type was the simple average of the topic ratio of each student's report, referring to Sun et al [8].

3 Results and Discussion

3.1 The Results of Extracting Topics

Table 2 shows a list of the most frequently occurring words extracted for each topic. These are the topics consisting the entire free writing of the 572 students. Table 2 also shows the labels of the topics as interpreted by the authors. The labels were named after examining the top words and the way the words were used in typical free descriptions.

Table 2: The extracted topics and the frequently occurring words in each topic

Topic	Label	Top frequently occurring words
01	Different careers for different people with different behaviors	people, feel, not, do, high school, are, say, try, around, many, other, talk, little, opportunity, friends
02	Goal setting	goal, do, set, not, find, firm, future, learn, yet, understand, dream, good
03	Accumulating experiences on the student's own	self, accumulate, person, knowledge, human, have, gain, way of life, career, process, polish, work on, get to know
04	Seniors and graduates as role models	senior, listen, upperclassmen, very, actual, spend, future, graduates, good, students
05	Understanding of one's own abilities (strengths and weaknesses)	ability, PROG Test, strength, know, weakness, task, competency, understand, future, take
06	Required abilities in society	ability, oneself, society, acquire, learn, seek, company, student, know, do, knowledge, era, have, talent
07	Being interested in various activities and meeting people	have, people, interest, can, accumulate, confidence, expand, meet, understand, find, vision, field
08	Studying at university with the expectation of entering the workforce	society, learn, get out, life, feel, volunteer, future, not, university student, get, working people
09	Words that left an impression on the student from the senior students	senior, hear, word, remain, feel, impression, very, especially, learn, mind, senior, strong, graduate
10	Interest in Japan and the world	interest, overseas, reason, Japan, world, trigger, know, special lecturer, work, international, very, have, advance
11	Stories from seniors and graduates as an opportunity to think about future dreams and careers	seniors, graduates, listen, know, feel, dream, occupation, get, actual, civil servant, decide, work, aim
12	Learning that a career is "one's life itself"	life, self, learn, own, person, more, concrete, way, not, academic path, role, president, good, spend
13	Understanding of how to achieve personal growth in the future	know, can, future, PROG test, part, good, advantage, opportunity, own, oneself, have, know
14	An opportunity to break free from the sense of stagnation caused by Covid-19	Covid virus, senior, ask, spend, circle activity, virus, university student, situation, future, new type

3.2 Topic Distribution on Interest Report and Career Understanding Report

Figure 2 shows the results of the topic distribution (proportion of each topic) of the Interest Report (Rep. A) and Career Understanding Report (Rep. B) by student type. Table 3 lists the topic numbers with the highest proportions in Figure 2.

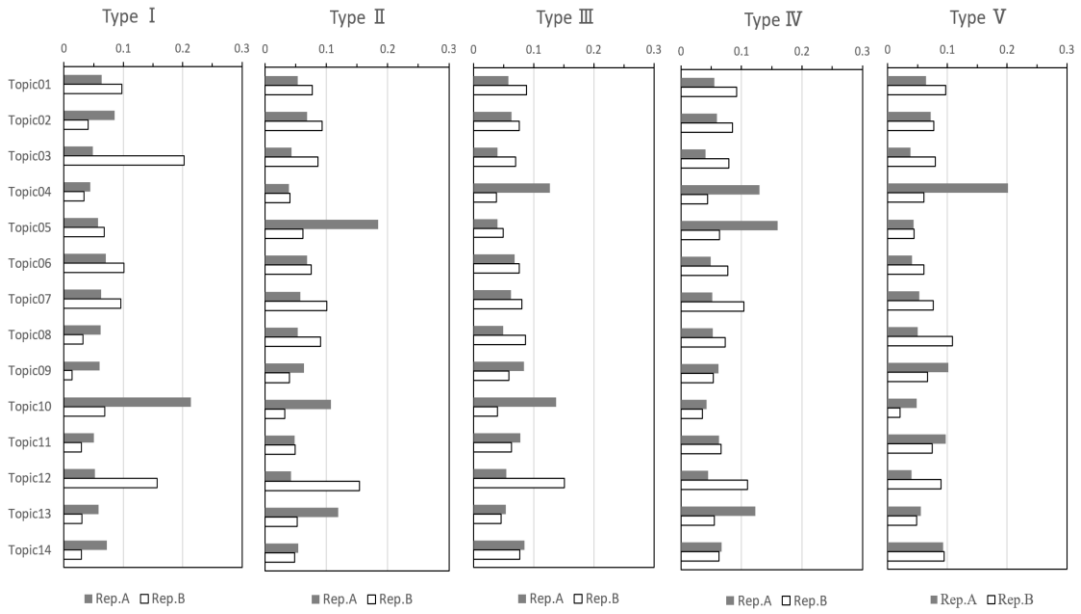

Figure 2: Proportion of each topic in Rep. A and Rep. B by student type

Table 3: List of top proportion topics in Figure 2

	Type I		Type II		Type III		Type IV		Type V	
Rank	Rep. A	Rep. B	A	B	A	B	A	B	A	B
1	Topic 10	Topic 03	05	12	10	12	05	12	04	08
2	Topic 02	Topic 12	13	07	04	01	04	07	09	01
3	Topic 14	Topic 06	10	02	14	08	13	01	11	14

One of the features that can be noted from Figure 2 is that the topics which are specific and depended on the lecture type are prominent in Rep. A. For example, Topic 10 is a prominent topic for students interested in the theory type, while Topic 05 and 13 are for students interested in self-understanding type lectures, and finally, Topic 04 is for students interested in the role model type. Since this is an analysis of reports on the reasons for interest, we can easily expect that specific topics related with the chosen lecture type would be prominent. The result of this analysis adequately expressed this expectation. Figure 2 shows the validity of the analysis results.

The purpose of the first-year career courses is developing the students' career awareness. Figure 2 and Table 3 showed the following about how the students perceived the career. Table 3 showed that Topic 12 has the highest proportion of understanding of career (Rep. B) for all student types. It is especially notable that the proportion is the highest for Type II, III, and IV students who chose two different lecture types. This indicates that most of the students who took this course understood that a career is "one's life itself" and they found that they should deal with it

over a long period of time. However, this tendency was weaker among Type V students than other student types.

Although only about 2% of the total number of students, the distribution of topics for career understanding was particularly characteristic of Type I students. Among Type I students, Topic 03 is prominent. A Type I student's report with higher proportion that included Topic 03 showed that he believed that career refers to the knowledge, experience, and skills that a person gains from various failures, difficulties, and setbacks. It also showed that he believed that participating in an internship or applying for an overseas aid project was not something that someone else did on their own, or something that just happened, but it was actually an opportunity that the person seized. In addition, Type I students were not very interested in self-understanding or role-model lectures, which most students were interested in. This provides us an image of Type I students as proactive individuals who accumulate experiences on their own.

Meanwhile, the students who were interested in role model type lectures tended to have lesser Topic 03, which was different from Type I students. Among the Type V students who were more inclined to role models, Topic 08 ranked first in career understanding, which was different from other student types. A Type V student with the higher proportion that included Topic 08 wrote that he thought that experiences and learning in university were important in the long run in terms of a career; moreover, he believed that this was because university life came with utmost freedom. Although Type V students were aware of the importance of learning at university, they had a lower proportion of Topic 12 than other types, and were relatively less likely to take a long view of their career. It is important for the students' growth to have prospects for their future and understand what they need to do in the present to achieve it. In this sense, the results of this study suggest the need to examine the content of the courses for Type V students. In addition, the ratio of Type V students is larger than the expected value, and the need for consideration of the lecture content is high.

Finally, we discuss the relationship between topics in Rep. A and Rep. B from Figure 2 and Table 3. The prominent topic with high proportion in Rep. A does not rank top in Rep. B among any other student types. In contrast, Topic 12 whose proportion was high in Rep. B was relatively lesser in Rep. A. Each prominent topic in Rep. A and Rep. B seem to have a reciprocal relationship.

It is interesting to note that the topics that students are interested in are not prominent in career understanding, while the topics related to the lecture contents that students are not interested in become prominent in career understanding. We could not confirm why such a phenomenon occurs and whether there is a causal relationship with the entire results of this study. This may be interesting research theme that should be clarified when considering the relationship between students' viewpoints of lectures and the overall course outcomes in the future.

4 Conclusion

By analyzing topics in free-writing reports written by the students who took first-year career courses, we examined the lecture contents and class outcomes that interested each of the five types of students. Fourteen topics were extracted from the free-writing reports. We also examined what the students were learning in the targeted subjects from the topic distribution. The results showed that most of the students understood career to be "their own life itself." The results also showed that the important topics in the lectures that they were interested in became less prominent in their career understanding, and conversely, topics that were important in their career understanding rarely appeared as topics in the lectures that

they were interested in. We are unable to ascertain this causal relationship from the results of this study. Future studies would like to analyze the causal relationship between topics to clarify this reciprocal phenomenon.

Acknowledgement

This work was supported by JSPS KAKENHI Grant No. 21K02634.

References

[1] M. Kikuchi, T. Suda, Y. Tange, and K. Murakami, "Students' learning and Thought-inducing Factors analyzed form their comments on a Career Course," [in Japanese], Jpn. Assoc. for College and University Education, vol. 41, no. 1, 2019, pp.117–156.

[2] B. Grün and K. Hornik, "Topicmodels: An R Package for fitting Topic Models," Journal of Statistical Software, vol. 40, 2011, pp.1–30.

[3] https://taku910.github.io/mecab/

[4] T. Tsumagari, Y. Nakazato, T. Tsumagari, "Analysis of the Actual State of Learning through Career Education as First-Year Experience Using a Topic Model," 2021 10th Int'l Congress on Advanced Applied Informatics(IIAI-AAI), 2020, pp.938–939.

[5] T. L. Griffiths and M. Steyvers, "Finding Scientific Topics," Proc. the National Academy of Sciences of the United States of America, vol. 101(Suppl 1), 2004, 5228–35.

[6] C. Jacobi, W. Atteveldt and K. Welbers, "Quantitative analysis of large amounts of journalistic texts using topic modelling," Digital Journalism, vol. 4, 2015, pp.1–18.

[7] M. Jin, "Basics and Practice of Text Analytics," [in Japanese], Iwanami Shoten, 2021.

[8] L. Sun and Y. Yin, "Discovering themes and trends in transportation research using topic modeling," Transportation Research Part C: Emerging Technologies, vol. 77, 2017, pp.49–66.