Visualizing Categorical Data: An Introduction to Correspondence Analysis for Technical Communication Researchers

Chris Lam
University of North Texas
Christopher.lam@unt.edu

Abstract – Technical communication researchers often collect categorical, or non-numerical, data. However, when analyzing this type of data, researchers are typically limited to reporting descriptive statistics or simple contingency tables. One advanced statistical technique that allows researchers to more powerfully explore complex categorical data sets is correspondence analysis. This article will first define correspondence analysis and then walk through a tutorial for conducting correspondence analysis in XLSTAT.

Index Terms – Correspondence analysis, data visualization, quantitative research, technical communication research

INTRODUCTION

Exploring and visualizing data has become an important component in technical communication research and practice. In fact, major journals and societies in the field have dedicated special journal issues and conferences to the topic of data visualization. However, even with all of this attention, additional scholarship is needed to provide researchers with the tools and knowledge to realize the full potential of their data sets. To address this need, I will present an introduction to correspondence analysis—a technique to explore and visualize complex categorical data sets.

Categorical data is commonly used in technical communication research. For instance, any study using a content analysis methodology will likely involve categorical data. The same is likely true for studies involving linguistic variables. Categorical data is data that can be coded into distinct categories and is non-numeric, which contrasts ordinal-level or interval-level data. Typically, researchers use contingency tables to examine relationships between categorical variables. For instance, a technical communication researcher might be interested in the relationship between a student's discipline (STEM, humanities, or social science) and technical writing experience (beginner, intermediate, or expert). Examining these two variables results in a 3x3 contingency table. See Table 1 for this contingency table populated with sample data.

TABLE 1. 3X3 CONTINGENCY TABLE FOR DISCIPLINE AND TECHNICAL WRITING EXPERTISE.

Discipline	**Level of Technical Writing Expertise**		
	Beginner	**Intermediate**	**Expert**
STEM	10	13	35
Soc.Sci.	47	12	9
HUM	11	17	25

As can be seen in Table 1, there seems to be some patterns arising. For instance, *STEM* and *humanities* (HUM) majors tend to have more experience in technical writing, while *social science* majors tend to have less experience. While a simple contingency table may be adequate for describing this trend, trends would be significantly more difficult to explore if we examined variables that had more than three categories. It is not uncommon for technical communication researchers to study variables that have significantly more than three categories. For instance, a recent study examined categorical variables—one with 13 categories and another with 5 categories [1]. Analyzing these two variables results in a 65-cell table—a table that would be virtually impossible to spot trends with using simple observation. Correspondence analysis (CA), therefore, allows researchers to explore and spot trends in data where simple observation of frequencies is not feasible.

The remainder of this paper is divided into two key parts. First, I will define CA and describe some basic concepts related to CA, including assumptions of CA, the exploratory nature of CA, row and column profiles, and CA visualization. Next, I'll take the reader through a tutorial for conducting and interpreting a CA in XLSTAT.

978-1-4799-3749-3/14/$31.00

Defining Correspondence Analysis

CA is a geometric technique used to analyze two-way and multi-way tables containing some measure of correspondence between the rows and columns [2]. The mathematical foundations for CA cannot be fully discussed in this article, but additional resources are available [3] – [5]. CA is widely used in corpus linguistics, marketing, and ecological research, but it has only recently been introduced to the field of technical communication [1]. The approach reveals patterns in complex data sets and provides output that can help researchers interpret these patterns. The most powerful tool in CA is its ability to visualize row points and column points onto a multi-dimensional graphical map called a biplot. Rows with comparable patterns, also known as profiles, will be placed in close proximity on the biplot. Similarly, columns with comparable profiles will be placed in close proximity on the biplot. When plotted together, the visualization allows a researcher to examine associations among row and column points.

Four Assumptions of Correspondence Analysis

There are four major assumptions when using CA [6]. First, CA should be used only when all of the variables of interest (both independent and dependent) are categorical. Whenever you have ordinal or interval-level variables, CA will not be the appropriate statistical technique. Instead, techniques like multiple regression analysis, principal components analysis, or factor analysis will be more appropriate depending on your research question(s) and data set. The second assumption is that all variables must have at least three categories, as any less would not require CA for interpretation. A third assumption is homogeneity of variance among row and column variables. This is most often violated if a variable has a value of zero for all of its entries. Finally, none of the data analyzed in a CA can be negative.

As a side note, this paper focuses on simple CA, which examines two categorical variables with any number of categories. Multiple CA is a separate technique that examines three or more categorical variables. The interpretation for a multiple CA differs from simple CA. For additional reading on multiple CA, see [7].

Exploratory Nature of Correspondence Analysis

One essential concept of CA is that it is an exploratory method, and, therefore, is not used for hypothesis testing. Instead, researchers use this method to reveal patterns that would not otherwise be readily apparent through simple observation. Results from CA can possibly lead to additional inquiry that can be examined using an approach that allows for hypothesis testing [8].

Table 2. Row Profile for Discipline and Technical Writing experience

	Level of Technical Writing Expertise			
	Beginner	**Intermediate**	**Expert**	**Sum**
STEM	0.123	0.193	0.684	1
Soc.Sci.	0.651	0.206	0.143	1
HUM	0.200	0.455	0.345	1

Table 3. column Profile for Discipline and Technical Writing experience

Discipline	**Level of Technical Writing Expertise**		
	Beginner	**Intermediate**	**Expert**
STEM	0.119	0.224	0.582
Soc.Sci.	0.695	0.265	0.134
HUM	0.186	0.510	0.284
SUM	1	1	1

Even though CA is not used for hypothesis testing, it does use the chi-square statistic, or a weighted Euclidian distance, to measure the distance between points [6]. It should be noted that CA is a non-parametric statistic and has no theoretical distribution to compare observed distances [6]. Therefore, findings must be reported as exploratory because statistical significance is not used to describe associations in CA.

The concept of row and column profiles

Another concept essential to understanding CA is row and column profiles. In the example presented previously in Table 1, I was able to identify trends based simply on observed frequencies in each cell. CA examines frequencies relative to row and column totals, defined as row and column profiles. That is, CA examines ratios to determine more specifically the impact of a particular cell. It creates a ratio separately for rows and columns, as can be seen in Tables 2 and 3. For instance, the *STEM/Expert* cell in Table 2 is 0.684, but it is 0.582 in Table 3. This is because row totals and column totals differ. Interpreting the cell in the row profile would read, "68.4% of *STEM* majors were also *experts*". For the column profile, it would read, "58.2% of *expert*s were also *STEM* majors".

It is these row and column profiles that are eventually mapped onto a two-dimensional or three-dimensional plot. Each row point, which has a unique profile, is mapped, and other row points with similar profiles will be mapped in close proximity. The same is true for each column point. I will discuss the visualization process in the next section of this paper. Compare Table 1 to Tables 2 and 3 in order to familiarize yourself with the concept of row and column profiles.

THE CONCEPT OF VISUALIZING ROWS AND COLUMNS

One final concept relevant to CA is the visualization of the row and column profiles. As stated previously, the raw frequencies aren't plotted, but instead it is the distances between row and column profiles that are plotted. For any given CA, the perfect visualization is one that accounts for 100% of the inertia, or variance in the model. Mathematically, this is only possible by visualizing n-1 dimensions for the variable with the most categories. For instance, if a CA examines a variable with 15 categories, a 14-dimensional plot is the only way to visualize 100% of the inertia. However, a 14-dimensional visualization would be virtually impossible to interpret. Therefore, it is the purpose of CA to reduce the number of dimensions analyzed by only visualizing the smallest number of dimensions while retaining as much inertia as possible.

The best CA's are visualized on a two-dimensional biplot. On this biplot, individual points from the row and column profiles are plotted together. Typically, but not always, when row points and column points are in close proximity to one another in the visualization, it indicates that the two points correspond. Figure 1 provides an example of a CA biplot from sample data presented in Tables 2 and 3. Notice three clusters with 1) *STEM* and *expert*, 2) *social science* and *beginner*, and 3) *humanities* and *intermediate*. This suggests that these variables are associated, but this is not always true. Often times, because of inadequate sample size or other statistical reasons like outliers, closely clustered variables can be meaningless. Therefore, additional statistical output must be consulted to properly interpret a CA. The process of interpretation will be explicitly described in the final section of this paper when I present the CA tutorial.

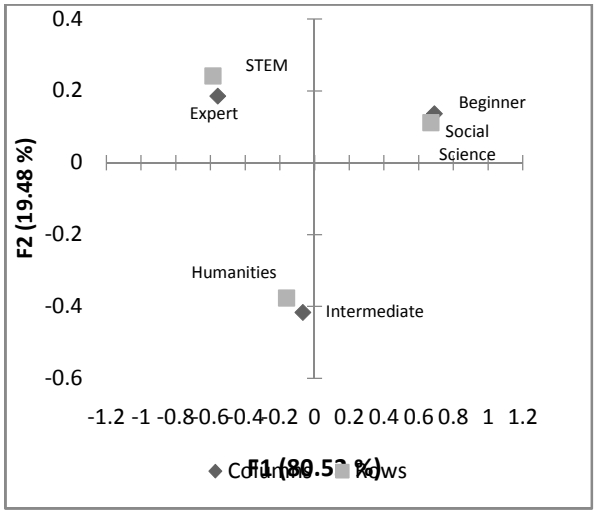

FIGURE 1. CA BIPLOT BASED ON TABLES 2 AND 3

CORRESPONDENCE ANALYSIS TUTORIAL

The remainder of the paper will describe a tutorial for conducting a CA. For the sake of continuity, I'll continue to use the example presented in the first section of this paper. However, I've changed some of the data and added a category to each variable resulting in a 4x4 table. This will allow for a more complex (though not overly complex) illustration of CA. Table 4 shows the contingency table that I'll use for the CA. Suppose we want to explore relationships between a student's discipline and technical writing experience in order to define prerequisite coursework for a technical communication certificate program.

TABLE 4. SAMPLE CONTINGENCY TABLE FOR CA

Major	**Level of Technical Writing Expertise**			
	None	**Beginner**	**Intermedi-ate**	**Expert**
STEM	8	7	11	39
Soc.Sci.	11	41	13	9
HUM	9	11	25	19
Interdis-ciplinary	15	9	21	14

CHOOSING SOFTWARE FOR CA

Since CA is an advanced statistical technique, it must be conducted in a statistical software program—for example SPSS, STATA, or R. This paper, however, will walk through a CA using XLSTAT, a cost-effective add-on for Microsoft Excel. I selected XLSTAT primarily because of the mainstream accessibility of Microsoft Excel.

PREPARING THE DATA

When preparing data in Excel, there is no need to recode the categorical variables into numerical codes, as you would have to do in a software program like SPSS. However, to ensure that your data can be adequately analyzed, it's essential that you check your data for spelling and case errors; all categories must be coded identically.

There are two ways to prepare data for the CA. First, you can keep data in an observations or variable table. In this preparation, each column in the table represents a separate variable. The second way to prepare the table is by creating a contingency table, also known as a pivot table in Microsoft Excel.

CONDUCTING THE CA IN XLSTAT

I will now walk through the actual steps to conduct a CA in XLSTAT. In the section that follows, I will walk through the detailed output and interpretation of results.

Opening the CA Package

1) Open the categorical data set in Microsoft Excel.
2) Find "**Analyzing Data"** from the **XLSTAT** menu bar.
3) Select "Correspondence Analysis."

Selecting the Output Options

1) Select "**data format**" or "**observations/variable table**" depending on the format of your data.
2) Highlight the data from your Excel spreadsheet that you want to analyze.
3) Check "**Variable Labels**" if your data is labeled. If not, leave this box unchecked.
4) Click the "**Outputs**" tab. Check the boxes as presented in Figure 2.
5) Click the "**Charts**" tab. Check the boxes as presented in Figure 3.
6) Click "**OK**" to run the correspondence analysis.

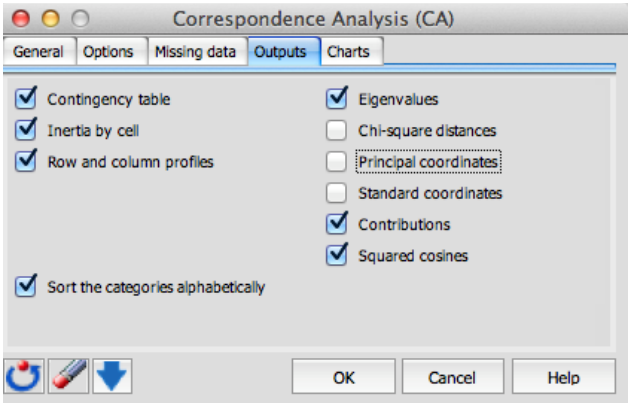

FIGURE 2. OUTPUTS OPTIONS IN XLSTAT

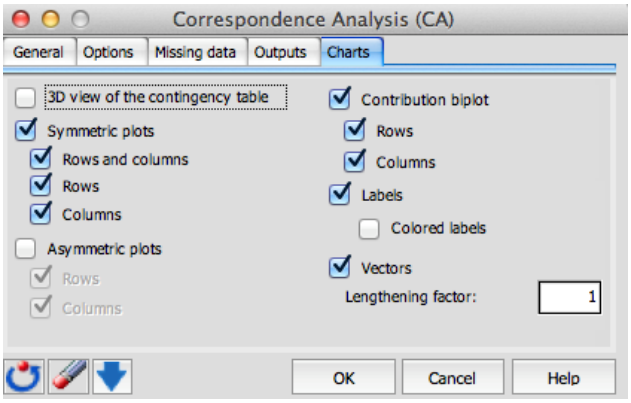

FIGURE 3. CHARTS OPTIONS IN XLSTAT

INTERPRETING CA OUTPUT

I will describe four major steps to interpreting a CA. Generally speaking, interpretation involves gathering and analyzing all of the available data to properly interpret the visualization. Therefore, while the steps are presented linearly in this paper, the process may not always follow the steps in the order presented.

I. Determining Significance of Dependencies

The first step in interpreting a CA is to examine whether the variables in question are related. If they aren't related, there is no reason to further interpret the CA. To determine whether or not the two variables are related, and to what extent they are related, we examine two major statistics. First, we examine the chi-square statistic and corresponding signifcance level to determine if the rows and columns are independent. If there is a signficant chi-square value, the rows and columns are not independent of each other.

Using the sample data, Table 5 displays the chi-square output presented in XLSTAT. As you can see, there is a chi-square value of 76.175 and a signifcance value of $p < .001$. This indicates that the two variables are not independent.

We then examine the total inertia value. Inertia in CA is the eqivalent to variance used in other statistical techniques and is reported a 0 to 1 scale. Obviously, a higher inertia value indicates that the variables explain more variance in the overall model. There is no documented reporting threshold for an acceptable total inertia value. Instead, the results from CA's with relatively low inertia should be hedged, or at the very least contextualized.

In the sample CA, the inertia level is 0.291. This indicates that the variables in question explain 29.1% of the total inertia.

TABLE 5. XLSTAT CHI-SQUARE TABLE

Chi-square (Observed value)	76.175
Chi-square (Critical value)	16.919
DF	9
p-value	< 0.0001

II. Selecting Dimensions

Because CA visualizes data in a multi-dimensional space, the next step is determining the smallest number of dimensions that adequately explains the greatest intertia in the anlaysis. As a reminder, for any CA, 100% of the intertia is explained by n-1 categories of your variable with the most categories.

In the sample CA, one of the outputs will be the percent of inertia table shown in Table 6. As can be seen, a cumulative of 96.714% of the inertia is explained by the first two dimensions. Therefore, for this CA, a two-dimensional solution is appropriate. Glynn suggests that additional dimensions should be added if the total inertia for the first two dimensions is less than 75% [8]. Table 6 also shows how much inertia each individual dimension adds to the overall model. In this example, dimension 1

(labeled F1) adds 71.672% of unique inertia to the overall model, while dimension 2 adds 25.042% of unique inertia. In contrast, dimension 3 only adds 3.286% of unique inertia, and therefore is excluded from further interpretation.

TABLE 6. XLSTAT PERCENT OF INERTIA TABLE

	F1	**F2**	**F3**
Eigenvalue	0.208	0.073	0.010
Inertia (%)	71.672	25.042	3.286
Cumulative %	71.672	96.714	100.000

III. Examining Row and Column Contributions

Once the number of dimensions is selected, contributions of row and column points should be examined. To accomplish this, the contributions table (Table 7) provided in the XLSTAT output can be analyzed. Before examining the table, we determine a contribution threshold by dividing 100 by the total number of rows (100/4= 25). Any row that contributes more than this number contributes more to the dimension than expected.

As seen in Table 7, *STEM* (35%) and *social science* (61.6%) both contribute more than expected to the first dimension. For the second dimension, *STEM* (38.8%) and *interdisciplinary* (35.1%) both contribute more than expected. These row points are highlighted in Table 7. It will be important to keep these row points in mind when it comes time to examine and interpret the CA visualization.

In addition to the contribution scores, another statistic that adds context to the interpretation is the quality score. The quality score indidcates how accurately the actual point is visualized or plotted on the biplot. XLSTAT provides a squared cosines table as seen in Table 8. The quality score is derived from the sum of squared cosines for the number of dimensions analyzed in the CA (in our example, the first two dimensions). For instance, the quality score for *STEM* would be 99.9% (72 + 27.9). We can conclude, then, that *STEM* is almost perfectly displayed on the biplot seen in Figure 4. Any quality score below 50% indicates that the given row or column point may not be accurately displayed on the biplot [8]. As can be seen in Table 8, the quality score well exceeds 50% for all four row points. Therefore, we conclude that the four row points are accurately plotted on the biplot.

The same process of examining contribution scores and then quality scores is applied to column points. As highlighted in Table 9, *beginner* (59.4%) and *expert* (39.6%) contribute most to dimension 1. For dimension 2, *intermediate* (46.6%) and *expert* (29.3%) both contribute more than expected. Again, it is important to keep these column points in mind when it comes time to interpret the visualization.

TABLE 7. ROW CONTRIBUTIONS

	Weight (relative)	**F1**	**F2**	**F3**
STEM	0.248	0.350	0.388	0.014
Soc.Sci.	0.282	0.616	0.101	0.000
HUM	0.244	0.024	0.160	0.572
Interdisc-iplinary	0.225	0.009	0.351	0.414

TABLE 8. SQUARED COSINES FOR ROWS

	F1	**F2**	**Quality Score**
STEM	0.720	0.279	99.9
Soc.Sci.	0.946	0.054	100
HUM	0.226	0.527	75.3
Interdisciplina-ry	0.062	0.812	87.4

TABLE 9. COLUMN CONTRIBUTIONS

	Weight	**F1**	**F2**	**F3**
None	0.164	0.000	0.109	0.727
Beginner	0.260	0.594	0.132	0.014
Intermed-iate	0.267	0.009	0.466	0.258
Expert	0.309	0.396	0.293	0.002

To examine quality scores for column points, Table 10 shows that all column points except for *none* are well above the threshold of 50%. The quality of *none*, however, is only 53.4%. Therefore, we must be very cautious about the accuracy of *none* on the biplot and consider that there is a chance that it is not as accurately plotted the way all of the other row and column points are. I'll provide more detail about this column point in the following section

TABLE 10. SQUARED COSINES FOR COLUMNS

	F1	**F2**	**Quality Score**
None	0.004	0.530	53.4
Beginner	0.927	0.072	99.9
Intermediate	0.051	0.885	93.6
Expert	0.795	0.205	100

Step 4: Interpreting the Visualization

To interpret the visualization, we begin with dimension 1, visualized along the x-axis. As noted, *STEM* and *social science* were rows that contributed most to the dimension. Notice in Figure 4 that *STEM* and *social science* are plotted on the opposite sides of the x-axis, which strongly suggests that the two disciplines have very different profiles. Similarly, *beginner* and *expert* were columns that

contributed most to the first dimension and are also plotted on opposite sides of the x-axis. *STEM* and *expert* are clustered on the left side of the x-axis, while *social science* and *beginner* are clustered on the right side. Since all four points are accurately displayed based on quality scores, we conclude that these points correspond. One final step to confirm this finding is to examine the row and column profiles. Table 11 shows that 60% of *STEM* students were *experts*. Similarly, Table 12 shows 48.1% of *experts* were *STEM* majors. Also, 55.4% of *social science* majors were *beginners*, and 60.3% of *beginners* were *social science* majors. The profiles, therefore, confirm the interpretation.

For the second dimension visualized along the y-axis, *STEM* and *interdisciplinary* were rows that contributed most. Notice how they are plotted on opposite sides of the y-axis. For columns, *intermediate* and *expert* contributed most and are also plotted on opposite sides of the y-axis. Since both *intermediate* and *interdisciplinary* both contributed heavily to the second dimension, are well represented based on their quality scores, and are close in proximity on the biplot, we conclude that the two are associated. The row and column profiles also support this interpretation as 35.6% of *interdisciplinary* students had *intermediate* experience.

Humanities is a special case to interpret as it didn't strongly contribute to either dimension. In these cases, we reexamine the squared cosines table (Table 8) for *humanities*. The squared cosines were 22.6% and 52.7% for the first and second dimensions respectively. This indicates that *humanities* is more closely associated with the second dimension. Since *intermediate* contributed significantly to the second dimension, it's reasonable to conclude that *humanities* and *intermediate* are also associated—though not as strongly as the previous associations. This is confirmed after examining the row profile for *humanities*, as 39.1% of *humanities* students had *intermediate* experience.

Finally, it's important to determine if *none* is a point worth interpreting in the CA. As stated, *none* barely meets the threshold for quality score, which indicates that it may not be accurately represented in the visualization. Since *humanities* is close to *none*, examining the row profile for *humanities* is logical. Only 14.1% of *humanities* students had no experience, and only 20.9% of students with no experience were *humanities* majors. Therefore, we conclude that *none* does not strongly associate with *humanities* and is likely plotted inaccurately in our two-dimensional visualization. A final clue that *none* is not interpretable is its high contribution to the third dimension (72.7%), as seen in Table 9. Since that dimension only contributes 3.286% of the total inertia, it was not examined.

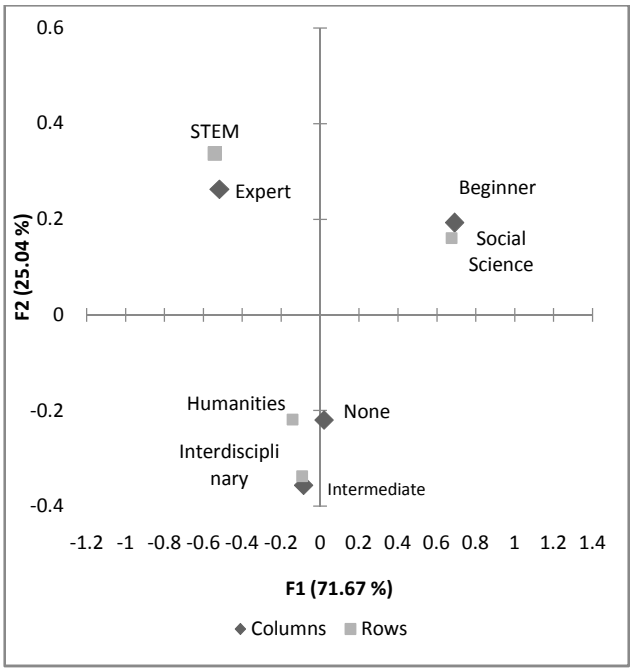

FIGURE 4. CA BIPLOT FOR DISCIPLINE AND EXPERIENCE

TABLE 11. ROW PROFILES FOR DISCIPLINE AND EXPERIENCE

	None	**Begin-ner**	**Inter-media-te**	**Expe-rt**	**Sum**
STEM	0.123	0.108	0.169	0.600	1
Soc.Sci.	0.149	0.554	0.176	0.122	1
HUM	0.141	0.172	0.391	0.297	1
Interdisc-iplinary	0.254	0.153	0.356	0.237	1
Mean	0.167	0.247	0.273	0.314	1

TABLE 12. COLUMN PROFILES FOR DISCIPLINE AND EXPERIENCE

	None	**Begin-ner**	**Inter-media-te**	**Expe-rt**	**Mean**
STEM	0.186	0.103	0.157	0.481	0.232
Soc.Sci.	0.256	0.603	0.186	0.111	0.289
HUM	0.209	0.162	0.357	0.235	0.241
Interdisc-iplinary	0.349	0.132	0.300	0.173	0.239
Sum	1	1	1	1	1

CONCLUSION

As shown in this article, CA is a technique that can provide immense value to technical and professional communication researchers who analyze categorical

variables. It allows researchers to powerfully explore relationships between categorical variables. While this article is simply an introduction to the method, it provides enough guidance for technical communication researchers to properly use and interpret the output provided in the technique.

REFERENCES

[1] R. Boettger and C. Lam, "An overview of experimental and quasi-experimental research in technical communication journals (1992-2011)," *IEEE Trans. Professional Communication*, vol. 56, no. 4, pp. 272–293, 2013.

[2] M. Greenacre, *Theory and Application of Correspondence Analysis*. London: Academic Press, 1984.

[3] M. Greenacre, *Correspondence Analysis in Practice*. Boca Raton, FL: Chapman & Hall, 2007.

[4] M. Greenacre and O. Nenadic. (2012). *Simple, multiple and joint correspondence analysis (R package)*. [Online] Available: http://ftp.ussg.iu.edu/CRAN/

[5] D.L. Hoffman and G.R. Franke, "Correspondence analysis: Graphical representation of categorical data in marketing research," *J. Marketing Research,* vol. 23, no. 3, pp. 213–227, 1986.

[6] L.Doey and J. Kurta, "Correspondence analysis applied to psychological research," *Tutorials in Quantitative Methods for Psychology*, vol. 7, no. 1, pp. 5–14, 2011.

[7] H. Abdi and D. Valentin, "Multiple correspondence analysis" in *Encyclopedia of Measurement and Statistics*, N. Salkind, Ed. Thousand Oaks, CA: Sage, 2007, pp. 1-13.

[8] D. Glynn, "Correspondence analysis: exploring data and identifying patterns," in *Polysemy and Synonymy: Corpus Methods and Applications in Cognitive Linguistics*, D. Glynn and J. Robinson, Eds. Amsterdam & Philadelphia: John Benjamins, to appear.

ABOUT THE AUTHOR

Chris Lam is an Assistant Professor of Technical Communication at the University of North Texas where he studies team communication in technical communication projects. He also examines research in professional and technical communication using quantitative methods.