AUTOMATIC CONTENT ANALYSIS AND BINARY CHARACTERISTICS DETECTION OF DREAMS

by

Xiaofang Zheng

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Psychological Sciences

West Lafayette, Indiana

May 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Richard Schweickert, Chair
Department of Psychological Sciences

Dr. Anne B. Sereno
Department of Psychological Sciences

Dr. Robert W. Proctor
Department of Psychological Sciences

Dr. Sebastien Hélie
Department of Psychological Sciences

Approved by:
Dr. Kimberly P. Kinzig

ACKNOWLEDGMENTS

First, I would like to express my gratitude to my supervisor, Dr. Richard Schweickert, for his invaluable advice and help. I really appreciate the opportunity he gave me to start the journey of Ph.D. life and his continuous support for me along the journey. Also, I want to thank my parents, Jun Zheng and Lirong Wang, for giving me life and love. I wouldn't be able to make it without their tremendous encouragement and supports. Additionally, my appreciation goes to my beloved husband, Wenbin Zhu, for his accompany, understanding, and love. Last but not least, I want to say thank you to my other family members, friends, research mates, and committee faculties for all of their kind help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

For dream content analysis, automatic quantitative analysis techniques not only can be faster than traditional hand-coding, but also be lower in coding errors and bias caused by humans. Linguistic Inquiry and Word Count (LIWC, Pennebaker, Boyd, Jordan, & Blackburn, 2015) is an automatic technique possibly useful for dream research. We tested the suitability of LIWC for dream content analysis by comparing results by LIWC and Hall Van de Castle coding system (HVdC, Hall & Van de Castle, 1966) using canonical correlation analysis. Moreover, we analyzed the consistencies and inconsistencies between dreaming and waking by comparing the word frequencies in reports. Last, we introduced machine learning techniques to dream research and built support vector machines to achieve the binary characteristics detection of dreamers (e.g., female or male, blind or sighted, waking activity or dream) based on dream content. Our theoretical and methodological contributions to dream research would not only deepen people's understanding about dreams but also introduce new methods for scientific research on dreams.

OVERALL INTRODUCTION

A famous legend about ancient China says that Duke Zhou, the founder of the Zhou Dynasty (1600-1046 BC), compiled a dream interpretation dictionary named "Dreaming of Duke Zhou" (*Zhougong Jie Meng*). It has been over three thousand years from then until now, and people have been increasingly curious about the interpretation of dreams. Dreams are believed to have adaptive functions for human beings (e.g., Blagrove, 1996; Ellman & Weinstein, 1991) and seem to be able to express uncensored thoughts, emotions, and fears (e.g., Bulkeley & Graves, 2018). Scientific studies also showed that dream content is meaningfully and predictably connected to the major concerns of the individual's waking life (e.g., Domhoff & Schneider, 2008a, 2008b).

Despite people's continuous interest and curiosity, dreams are far from well understood because of their uncertainty, instability, and indecisiveness. Instead of pure qualitative analysis, more and more researchers tend to quantitatively examine the relationship between dream contents and factors such as gender, age, mental health, and personality. Linguistic Inquiry and Word Count (LIWC, Pennebaker, Boyd, Jordan, & Blackburn, 2015) is an automatic technique possibly useful for dream research. To evaluate the suitability of information produced by LIWC for dream studies, Chapter 1 compared the outputs by the Hall Van de Castle (HVdC) coding system (Hall & Van de Castle, 1966), the traditional and most often used hand-coding system in dream research, and LIWC, by means of canonical correlation analysis (CCA) using a classic set of dream reports collected by Hall and Van de Castle (1966).

The continuity hypothesis is "There is considerable congruence between what a person dreams about at night and what he does or thinks about when he is awake" (Hall & Nordby, 1972, p. 125.). Despite recent controversies (e.g., Domhoff, 2017; Erdelyi, 2017), the hypothesis has been largely supported by dream research (e.g., Domhoff, 1996; Schredl & Hofmann, 2003; Strauch & Meier, 1996). For example, dream contents were found to be related to people's personality traits (e.g., Bernstein & Roberts, 1995; Hall, 1953; Hartmann, Elkin, & Garg, 1991; Hawkins & Boyd, 2017; Schredl, Schäfer, Hofmann, & Jacob, 1999), attitudes and beliefs (e.g., Bulkeley, 2009), life events (e.g., Proksch & Schredl, 1999), and emotions (e.g., Breger, Hunter, & Lane, 1971). However, there still remains a lot to be discovered about the correspondence between dream contents and waking activities. In Chapter 2, we tested the consistencies between waking life and dreams in terms of gender, age, and blind vs. sighted differences by word frequency analysis using LIWC. Besides, dreams and waking life are not always correspondent. Only a few studies have ever tried to explore the differences between waking life texts and dream reports by quantitative analysis (e.g., Bulkeley & Graves, 2018). Hence, in Chapter 3, the discontinuities between dreams and waking life were analyzed from the aspects of social contents and cognitive functions by doubly repeated mixed effect models.

Further, machine learning techniques have occasionally been utilized to automatically score dream content (e.g., Fogli, Maria Aiello, Quercia, 2020; Wong, Amini, & De Koninck, 2016). Such methods can undoubtedly promote the compatibility and speed of coding dream reports, which deserve further application in dream research. In Chapters 2 and 3, we gave an introduction of the support vector machine (SVM), a machine learning technique, to dream research and built SVM models to precisely detect binary characteristics of dreamers and dreams based on the word frequencies of LIWC categories. The three chapters are written separately, and they can be read independently. A manuscript based on Chapter 1 has been submitted for publication. Our studies introduce methods for automatically analyzing dream contents to quantitatively measure important aspects of dreams and also enrich knowledge about the associations, consistencies and inconsistencies, between waking life and dreaming.

CHAPTER 1: COMPARING HALL VAN DE CASTLE CODING AND LINGUISTIC INQUIRY AND WORD COUNT USING CANONICAL CORRELATION ANALYSIS

Introduction

Increasing amounts of dream materials, e.g., surveys, experiments, and reports, are available for dream researchers. Such materials contain useful information, but it is not easy to extract it with traditional hand-coding. Technology and artificial intelligence help, with automatic quantitative analysis of dream contents. Automatic techniques are obviously faster than hand-coding and have lower biases caused by human error. But they produce different information. Before we lean on new tools, we should carefully compare their results with results of traditional ones. The automatic technique most often used so far for dream analysis is Linguistic Inquiry and Word Count (LIWC, Pennebaker, Boyd, Jordan, & Blackburn, 2015), and the hand-coding technique most often used is the Hall Van de Castle system (HVdC, Hall & Van de Castle, 1966). Here we compare the two techniques with a multivariate statistical analysis, canonical correlation.

Canonical correlation analysis, developed by Hotelling (1936), is a statistical method for identifying and measuring associations between two sets of variables (Knapp, 1978). Canonical correlation analysis is potentially greatly useful in dream studies. As a multivariate method, it has several advantages over univariate analysis in some contexts (e.g., Sherry & Henson, 2005). For one thing, many dream research topics are about associations between multivariate quantities. An example is the association between personality and dream content, each of which has multiple aspects. Separately testing each personality variable with each dream variable ignores the complexity of each set of variables, and may only give partial answers to researchers' most natural questions. Instead of analyzing variables in pairs, multiple regression could be used to, say, predict each dream content variable from all the personality variables. But Type I errors possibly arise from each multiple regression conducted. Canonical correlation analysis can curtail these, because it allows for multiple simultaneous comparisons among variables. Canonical correlation generalizes commonly used techniques such as Analysis of Variance and multiple regression and can sometimes be advantageously used in place of them (e.g., Henson, 2000; Knapp, 1978; Sherry & Henson, 2005; Thompson, 1991). A drawback to canonical correlation analysis is that sometimes the variables it constructs are not easy to interpret. Another is that often inputting a large number of variables yields little of significance to report. Of the variables constructed from a canonical correlation analysis, “the first one or two pairs are often significant and the remaining ones are not” (Tabachnick & Fidell, 2013, p. 573). Despite drawbacks, canonical correlation analysis has the potential to be well suited to the complexity of dream content. Hence, a second purpose of the current study is to give an introduction of canonical correlation analysis to dream researchers both intuitively and more formally.

Tools for Dream Analysis

The HVdC system is one of the most widely used methods for dream content analysis. It defines coding rules for characters, social interactions (aggression, friendliness, sexuality), activities, success and failure, misfortune and good fortune, emotions, settings, objects, and other elements. As Domhoff (1996) said, still true, the Hall/Van de Castle coding system “is the most comprehensive and detailed system for the study of dream content developed to date.” One can find more information at https://dreams.ucsc.edu/Coding/ and Hall and Van de Castle (1966).

Some automatic text analysis software was developed specifically for dream studies. One of the earliest is word search technology (Bulkeley, 2009, 2014; Bulkeley & Domhoff, 2010; Domhoff & Schneider, 2008). It was developed for searching for word usage in dream reports on DreamBank.net. A version is now available on another website, the Sleep and Dreams Database (Bulkeley, 2009). Once one chooses a set of dreams and a word or word string to search for, the websites will show their frequencies, and which dream reports in the set contain the words. Word strings of special interest for dreams, regarding perception, cognition, emotion, social interactions, common culture, nature, and so on, were developed for the word search technique. Results of word search technology have been shown to be compatible with traditional approaches (Bulkeley, 2014). LIWC is software developed for analyzing texts in general. The current version is LIWC 2015 (Pennebaker, Boyd, Jordan, & Blackburn, 2015). When text is put into LIWC, it automatically tabulates percents of word usage in approximately 100 different categories, about emotion, grammar and vocabulary, social processes, and so on. As a powerful automatic content analysis tool, LIWC has been used widely, including in psychological research (e.g., Kahn, Tobin, Massey, & Anderson, 2007; Vaughn, 2018), and occasionally in dream studies (e.g., Hawkins & Boyd, 2017; McNamara, Pae, Teed, Tripodis, & Sebastian, 2016; Wong, Amini, & De Koninckand, 2016; Zheng, Schweickert & Song, 2021).

Bulkeley (2014) has found a good consistency between outputs by HVdC and the word search technique by means of univariate analysis. It would be useful to compare the HVdC coding system with the word search technique by canonical correlation analysis. But because both were developed specifically for dreams, one would expect results to correspond well. It is not so clear how well results of HVdC coding and LIWC would correspond, so we compare them here using canonical correlation analysis. As Bulkeley and Graves (2018) argued, results from LIWC should be tested before one can say that LIWC is suitable for dream studies.

Therefore, we examine compatibility of results of LIWC and the HVdC coding system. The HVdC system produces codes for various elements of dreams, such as settings, characters and social interactions. And LIWC produces percents of words in various categories. It might seem at first that results of the two systems must be different because the things they measure are different. The HVdC system treats a dream report as a description of a play performance, with characters in action. LIWC treats the dream report as a script, with words and punctuation marks. Nonetheless, each action in a play, a swordfight, for example, has a corresponding passage in the script and vice versa, so perhaps separate analyses of performance and script may be put in alignment.

In fact, some HVdC classes can be meaningfully compared with LIWC categories. For example, the category “See” in LIWC and the class “Visual” in HVdC are both about seeing activities. Table 1 lists comparable items in HVdC coding and LIWC found by inspection. Obviously, the count in an HVdC class will correspond to some extent with the percent in a LIWC category if they have similar labels, but will they align well? If one is investigating a research question that could be posed in terms of certain HVdC classes, is it reasonable to conduct the study using LIWC categories that have similar labels?

Some classes in HVdC, e.g., distorted settings, have no counterpart in LIWC. A researcher pursuing a question about such a class is not likely to consider using LIWC. So we examine here the comparable items in HVdC and LIWC. We now turn to explaining how canonical correlation, a multivariate analysis of correlation, can be used to examine compatibility of outputs of the two tools. The analysis addresses two questions. Do outputs of the two tools systematically align? How much variability in the output of one can be accounted for by variability in the output of the other?

Table 1. Comparable Items in HVdC Classes and LIWC Categories

HVdC Classes	LIWC Categories
Family	Family
Household	Home
Friendliness	Friends
Sexual Interactions	Sexual
Visual	See
Auditory	Hear
Verbal	Hear
Physical	Motion
Movement	Motion
Dead	Death
Success	Achievement
Sadness	Sadness
Anger	Anger
Social Interactions	Social Processes
Thinking	Analytical Thinking
Happiness	Positive Emotion
Thinking	Cognitive Processes
Body Parts	Body
Apprehension	Anxiety
Money	Money
Food	Ingestion
Female	Female References
Male	Male References

Note: LIWC categories Hear and Motion each appear twice.

Canonical Correlation Analysis

Let's introduce canonical correlation analysis with a simple hypothetical example. Figure 1 shows two sets of observed variables and two pairs of unobserved variables constructed by canonical correlation analysis. Set *A* consists of three observed variables, a_1, a_2, and a_3; and set *B* consists of two observed variables b_1 and b_2. All observed variables are measured on the same units. (Units in our case are dream reports.) The canonical correlation analysis has found two pairs of canonical variates, which are not themselves observed, and a canonical correlation coefficient for each pair, which is the Pearson correlation coefficient between the two canonical variates of the pair, to express the relationship between the two sets of observed variables. Observed variables a_1 and a_3 contribute to constructing the canonical variate CV_{a_1}, and observed variables b_1 and b_2 contribute to constructing the canonical variate CV_{b_1}. The CV_{a_1} and the CV_{b_1} are a pair of canonical variates. Similarly, the second pair of canonical variates are constructed from the observed variables a_2 and b_2.

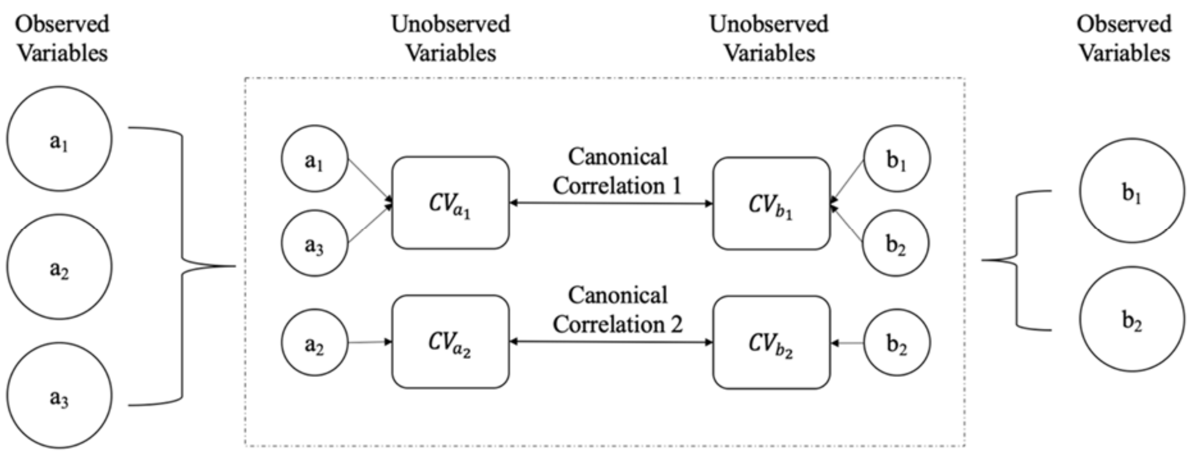

Figure 1. An example of canonical correlation analysis between variable sets *A* and *B*. Circles contain observed variables. Squares contain canonical variates constructed by the analysis.

A canonical correlation analysis is a generalization of a regression analysis. With a linear regression without an intercept, the equation for predicting y from x can be written as

$$\alpha x = y.$$

A multiple regression with two independent variables predicting y can be written as

$$\alpha_1 x_1 + \alpha_2 x_2 = y.$$

With canonical correlation, two or more variables predict two or more other variables, for example,

$$\alpha_1 x_1 + \alpha_2 x_2 = \beta_1 y_1 + \beta_2 y_2.$$

In more complicated cases, when there are more than two variables for x or y, it is possible to have more than one equation of the form above.

Now let us give a more formal introduction of canonical correlation analysis. Suppose there are multivariate random variables $X = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_m \end{pmatrix}$ and $Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$ observed, each measured on the same units. With canonical correlation analysis, we get a first pair of canonical variates (CV_{x_1}, CV_{y_1}). The first canonical variate is a linear combination of variables within X and the second is a linear combination of variables within Y. That is,

$$CV_{x_1} = \alpha_{11}x_1 + \alpha_{12}x_2 + \dots + \alpha_{1m}x_m \text{ and } CV_{y_1} = \beta_{11}y_1 + \beta_{12}y_2 + \dots + \beta_{1n}y_n.$$

The coefficients $\alpha_{11}, \ldots, \alpha_{1m}$ and $\beta_{11}, \ldots, \beta_{1n}$ are chosen to make the correlation over units between CV_{x_1} and CV_{y_1} as large as possible. That correlation is the first canonical correlation. Also, for convenience, the coefficients are adjusted to make the variances of CV_{x_1} and CV_{y_1} each 1. This does not affect the value of the canonical correlation.

We get a second pair of canonical variates (CV_{x_2}, CV_{y_2}), the first a linear combination of variables within X and the second a linear combination of variables within Y. The coefficients are chosen to make the correlation between CV_{x_2} and CV_{y_2} as large as possible, with the restriction that four correlations over units are 0, those between CV_{x_1} and CV_{x_2}, CV_{x_1} and CV_{y_2}, CV_{y_1} and

CV_{x_2}, CV_{y_1} and CV_{y_2}. Further, the variances of CV_{x_2} and CV_{y_2} are each constrained to be 1. The correlation between CV_{x_2} and CV_{y_2} is the second canonical correlation.

We continue getting pairs of canonical variates until the number of canonical variate pairs is the smaller of *m* and *n*. Here, we assume *m* < *n*. Then the canonical variates can be expressed in a list of *m* pairs of equations.

$CV_{x_1} = \alpha_{11}x_1 + \alpha_{12}x_2 + \ldots + \alpha_{1m}x_m$ and $CV_{y_1} = \beta_{11}y_1 + \beta_{12}y_2 + \ldots + \beta_{1n}y_n$

$CV_{x_2} = \alpha_{21}x_1 + \alpha_{22}x_2 + \ldots + \alpha_{2m}x_m$ and $CV_{y_2} = \beta_{21}y_1 + \beta_{22}y_2 + \ldots + \beta_{2n}y_n$

……

$CV_{x_m} = \alpha_{m1}x_1 + \alpha_{m2}x_2 + \ldots + \alpha_{mm}x_m$ and $CV_{y_m} = \beta_{m1}y_1 + \beta_{m2}y_2 + \ldots + \beta_{mn}y_n$

In the equations above, the coefficients α and β are called canonical weights (also known as canonical coefficients). The pair of CV_{x_i} and CV_{y_i}, *i = 1, 2, …, m*, is the *i^th^* pair of canonical variates. As we introduced earlier, for the first pair of canonical variates, coefficients α and β are chosen to maximize the correlation between CV_{x_1} and CV_{y_1}. For the i^{th} pair of canonical variates, coefficients α and β are chosen to maximize the canonical correlation between CV_{x_i} and CV_{y_i}, subject to the restriction that each of CV_{x_i} and CV_{x_i} is uncorrelated with any canonical variate higher in the list of equations. Successive pairs are similarly restricted. The only nonzero correlations between canonical variates are those between the two in the same pair. More explicitly, any combinations (CV_{x_g}, CV_{x_h}), $g \neq h$ and $1 \leq g, h \leq m$, are uncorrelated. Similarly, any combinations (CV_{y_g}, CV_{y_h}), $g \neq h$ and $1 \leq g, h \leq m$, are uncorrelated. Any combinations (CV_{x_g}, CV_{y_h}), $g \neq h$ and $1 \leq g, h \leq m$, are uncorrelated as well. The variance of each canonical variate is 1 (see, e.g., Johnson & Wickern, 1988, p. 590).

By canonical correlation analysis, we measure the strength of the association between the multivariate random variables *X* and *Y*. Canonical correlation coefficients, canonical weights, and canonical loadings are three statistics about the relationship between the canonical variates and the original observed variables. A canonical correlation coefficient measures the strength of association between a pair of canonical variates. To understand the relationship between the original observed variables and the canonical variates, one can look at the canonical weights and the canonical loadings (also known as the structure correlations). A canonical loading is the correlation between a canonical variate (e.g., CV_{x_i}) and one of the original observed variables (e.g., x_j) that the canonical variate is a function of. The absolute value of the canonical loading reflects the contribution of an original observed variable to the canonical variate, with a larger absolute value indicating that the original observed variable is more important for the canonical variate. For more information about canonical correlation analysis, see Tabachnick and Fidell (2013) and Johnson and Wichern (1998).

Method

In the current study, we did canonical correlation analysis using the comparable items in Table 1 to examine the compatibility between outputs of HVdC and LIWC. We analyzed a classic set of norm dreams collected by Hall and Van de Castle (1966). The dream reports were collected from 100 male and 100 female American college students from 1947 to 1950. There were originally 1000 dream reports, with 5 reports from each participant. Unluckily, only 981 (491 from males and 490 from females) are available now on Dreambank.net and it is not known which dream reports are from which participants. The codes of these 981 dreams, based on the HVdC coding system, are available on Dreambank.net. This classic set of norm dreams has been studied many times, and there have been several replications and follow ups in dreams of college students and other adult samples (e.g., Bulkeley, 2009; Domhoff, 1996, 2003).

For more information about HVdC classes and coding rules see Hall and Van de Castle (1966) and https://dreams.ucsc.edu/Coding/. See the LIWC2015 development manual (Pennebaker, Boyd, Jordan, & Blackburn 2015) for more information about the development and psychometric properties of LIWC and its categories.

In LIWC we selected the word categories that have similar meanings with classes in HVdC; see Table 1 (see Table 2 for sources of the HVdC classes in Table 1 in the HVdC coding system). Then, we put the norm dream reports (981 in total) from Dreambank.net into LIWC and got percents of words in each category in each dream report. From DreamBank.net we had the HVdC codes of the norm dreams from Hall and Van de Castle (1966). For each HVdC class in Table 1, we counted the occurrence of the codes for the class in each dream report.

Table 2. HVdC Classes in the Current Study and the Corresponding Class Source in the HVdC System

HVdC Class (Current Study)	HVdC Coding System Source
Female	Characters – Gender – Female
Male	Characters – Gender – Male
Family	Characters – Identity – Immediate family members + relatives
Dead	Characters – Additional coding rules
Friendliness	Social Interactions – Friendliness
Sexual interactions	Social Interactions – Sexuality
Social interactions	Social Interactions (sum of aggression, friendliness, and sexual interactions)
Physical	Activities – Physical
Movement	Activities – Movement
Verbal	Activities – Verbal
Vision	Activities – Vision
Hearing	Activities – Hearing
Thinking	Activities – Thinking
Success	Success and Failure Success
Anger	Emotions – Anger
Apprehension	Emotions – Apprehension
Sadness	Emotions – Sadness
Happiness	Emotions – Happiness
Household	Objects – Household
Food	Objects – Food
Body parts	Objects – Body parts
Money	Objects – Money

To give an example, here is Female Dream Report #001.

> I dreamed it was next summer and that I was going to be married to my boyfriend at home. Mother advised me not to, but said she would not stand in my way. He had very little money and George, after finally convincing me not to finish school, said maybe mother was right. I had made up my mind, however, and would not hear of a postponement. The next part is after we are married. Although I did not see the ceremony, I knew it was very simple. Someone asked where we should go on our honeymoon and I can't remember where I said, but they thought it was the place in Europe named the same. I was embarrassed when I said it was not. Then I thought maybe I should have married another fellow I know who would have taken me to Europe, but I immediately dismissed this from my mind. Mother said then that the place we were going was very nice, although neither of us had previously been there. This made me feel much better. Next we were sitting in a large dining room. George was drinking a beer and I had a coke. All of a sudden it struck me that this was my wedding night and I got nervous and sort of afraid. I then asked George to order me a double shot which I never did get to drink because I woke up.

See Table 3 for the counts of HVdC classes and word percents in LIWC categories of interest in this dream report. For instance, in the HVdC coding system, there are two occurrences of male characters – my boyfriend and George, and one occurrence of a female character – Mother. Thus, the count of male characters in our dataset is 2 and the count of female characters is 1. In Table 3, “word percent” means the percent of words of a given category in a text. The percents of words counted by LIWC about male and female are 1.26% and 1.67% respectively.

Overall, we got a dataset with percents of relevant LIWC categories and counts of HVdC classes in the 981 dream reports. Using these data, we did canonical correlation analyses to compare the consistency between content analysis results by HVdC and LIWC.

Results

The correlation between two variables may be low if one or both are highly skewed (e.g., Kirk, 1999). For the purpose of learning the relation between HVdC and LIWC variables, it is not informative to find that for highly skewed variables correlations are low. It is of more interest to find out whether correlations are low after skewness has been reduced. Hence, we did preliminary tests and did logarithmic transformations on variables that were highly skewed (Tabachnick &

Table 3. Example: Counts of HVdC Classes and Word Percents of LIWC Categories in Female Dream Report #0001

HVdC Classes	Counts	LIWC categories	Word Percents
Family	1	Family	3.35
Household	0	Home	0.84
Friendliness	2	Friends	0.84
Sexual interactions	0	Sexual	0.0001
Visual	0	See	0.42
Auditory	0	Hear	2.51
Verbal	6	Hear	2.51
Physical	2	Motion	1.26
Movement	0	Motion	1.26
Dead	0	Death	0.0001
Success	0	Achievement	0.42
Sadness	0	Sadness	0.0001
Anger	0	Anger	0.0001
Social Interactions	4	Social processes	12.97
Thinking	2	Analytical thinking	11.03
Happiness	1	Positive emotion	0.84
Thinking	2	Cognitive processes	14.64
Body parts	0	Body	0.0001
Apprehension	2	Anxiety	1.26
Money	1	Money	0.42
Food	3	Ingestion	2.09
Female	1	Female references	1.67
Male	2	Male references	1.26

Fidell, 2013; For discussion, see Ayodeji & Obilade, 2018). If the absolute value of the calculated skewness is bigger than 1, then the variable is considered to be highly skewed (following Tabachnick & Fidell, 2013). In our study, all 23 HVdC variables were found to be significantly positively skewed and were log transformed. Seventeen of the 21 LIWC variables were significantly positively skewed and were log transformed too. The exceptions were LIWC variables of social processes and cognitive processes, which were nonsignificantly positively skewed and the LIWC variable of analytical thinking which was nonsignificantly negatively skewed.

We did canonical correlation analysis using SAS 9.4. The raw HVdC variables are logs of counts and the raw LIWC variables are logs of percents or percents. These are on different scales. Standardizing makes them more comparable. Each variable was transformed by subtracting its mean and dividing by its standard deviation. The canonical correlation analysis was done on these standardized values.

See Table 4 for results. Although the output is voluminous and rather complicated, there are indications that similar HVdC classes and LIWC categories align very well. The table lists pairs of canonical variates having significant canonical correlations. Before looking at details, one can immediately see that many HVdC classes and LIWC categories line up well. For each canonical variate pair, the HVdC classes and LIWC categories are listed in order of their contributions. It is obvious that a label on the left hand side of the table often lines up with a similar label in the same row on the right hand side. For example, for the first pair, the labels on the left hand side are “Female” and “Male” and labels on the right hand side are “Female references” and “Male references.” For the fourth pair, the left hand labels are “Family”, “Body Parts” and “Apprehension” while the right hand labels are “Family,” “Body,” and “Anxiety.” Such alignment is not forced to happen by the algorithm. There are sometimes misalignments between the HVdC classes and LIWC categories. For example, for the 11th pair of canonical variates, the second HVdC class is “Verbal” while the second LIWC category is “Positive emotion.” Despite some misalignments, individual HVdC classes and similar LIWC categories line up very well and this indicates good correspondence between the two systems.

An HVdC variable is a class such as “Family.” An HVdC canonical variate is a linear combination of HVdC variables. In Table 4, the correlation between an HVdC variable and an HVdC canonical variate is in the column r_X and is called a canonical loading. Likewise, on the right side of Table 4, the correlation between a LIWC variable and a LIWC canonical variate is in column r_Y. On the left side, for example, the canonical loading between "Female" and V1 is .953 and that between "Male" and V1 is - .366. For each canonical variate on the left side of the table, the respective variables are listed in order of absolute value of canonical loading; the same is true on the right side. Every one of the 23 HVdC classes under consideration contributes something in the complete expression for V1, although some contribute little. With the criteria of Cohen (1988) a correlation of .5 is considered large and one of .3 is considered medium. To make Table 4 more comprehensible, only variables with canonical loadings whose absolute values are greater than .3 are listed, following Tabachnick and Fidell (2013).

We note that the canonical weight (coefficient) of an observed variable in the expression of its respective canonical variate is another way of evaluating importance of the observed variable to the canonical variate. This way can be misleading because of multicollinearity (Dattalo, 2014), so we use canonical loading as a gauge of importance.

Every HVdC class and every LIWC category on the list of similar items in Table 1 appears in Table 4 as contributing with a large or medium correlation to at least one canonical variate, with the sole exception of the HVdC class Physical. So the HVdC classes and LIWC categories with similar labels pretty well cover each other. This suggests that an investigator using LIWC categories instead of similar HVdC classes would not be ignoring something important.

Pairs of canonical variates are listed in Table 4 in order of their canonical correlations. These are in the column headed r_{XY}. Many are high. For example, the first pair of canonical variates (V1 and W1) are highly correlated, $r_{XY} = 0.887$. There are 21 pairs of canonical variates. Wilks' lambda test is one of the most used methods to assess overall model fit. In canonical correlation analysis, the Wilks' lambda test is used for testing the null hypothesis that the given canonical correlation and all smaller ones are equal to zero in the population. It is commonly reported for two things: first, to test whether any linear relation is significant between the two sets of variables; second, to test how many of the canonical variate pairs are significantly linearly related. Wilks' lambda is converted to an F ratio to test for significance. As shown in Table 5, the first 18 of the 21 canonical correlations are significant at the 0.05 level, and they range from 0.887 to 0.169. To be more specific, the first test in Table 5 is of whether all 21 canonical correlations are equal to zero and the result leads to rejection of this hypothesis (F = 12.260, numerator df = 462,

Table 4. Canonical Correlation Analysis Results

HVdC Classes	α	r_X	CV_X	r_{XY}	CV_Y	β	r_Y	LIWC Categories
Female	0.901	0.953	V1	0.887	W1	0.909	0.956	Female References
Male	-0.305	-0.366				-0.271	-0.375	Male References
Male	0.853	0.885	V2	0.816	W2	0.863	0.894	Male References
Family	0.191	0.393				0.163	0.482	Family
Social Interactions	0.094	0.334				-0.010	0.466	Social Processes
Body Parts	0.687	0.764	V3	0.746	W3	0.637	0.750	Body
Apprehension	0.434	0.590				0.467	0.559	Anxiety
						-0.141	-0.314	Social Processes
Family	-0.747	-0.710	V4	0.699	W4	-0.765	-0.660	Family
Body Parts	0.420	0.419				0.439	0.424	Body
Apprehension	-0.396	-0.410				-0.393	-0.363	Anxiety
Apprehension	0.691	0.627	V5	0.656	W5	0.675	0.662	Anxiety
Family	-0.619	-0.478				-0.590	-0.475	Family
Sexual Interactions	0.578	0.540	V6	0.562	W6	0.542	0.535	Sexual
Visual	-0.339	-0.402				-0.347	-0.400	See
Food	-0.359	-0.397				-0.294	-0.394	Ingestion
						0.343	0.353	Anger
						0.393	0.352	Social Processes
						-0.061	-0.304	Analytical Thinking
Sadness	0.388	0.354	V7	0.544	W7	0.470	0.437	Sadness
Movement	0.388	0.353						
						0.374	0.370	Money
						0.395	0.350	Ingestion
						0.326	0.349	Home
Movement	-0.426	-0.518	V8	0.511	W8	-0.345	-0.469	Motion
						-0.212	-0.416	Analytical Thinking
						0.270	0.390	Cognitive Processes
Food	0.509	0.496				0.460	0.371	Ingestion

Table 4 continues

HVdC Classes	α	r_X	CV_X	r_{XY}	CV_Y	β	r_Y	LIWC Categories
Visual	-0.329	-0.364				-0.411	-0.369	See
						-0.398	-0.362	Sexual
Sadness	-0.683	-0.708	V9	0.490	W9	-0.665	-0.693	Sadness
						0.575	0.601	Ingestion
Visual	0.646	0.579	V10	0.445	W10	0.631	0.614	See
Movement	-0.496	-0.440				-0.447	-0.591	Motion
Thinking	0.341	0.403				0.363	0.468	Cognitive Processes
Auditory	-0.436	-0.460	V11	0.424	W11	-0.659	-0.625	Hear
Verbal	-0.466	-0.409						
Success	0.438	0.419						
Happiness	0.259	0.363				0.534	0.497	Positive Emotion
Sexual Interactions	0.425	0.387				0.366	0.350	Sexual
Anger	-0.611	-0.646	V12	0.379	W12	-0.726	-0.625	Anger
Dead	0.366	0.341				0.567	0.407	Death
Money	0.721	0.713	V13	0.361	W13	0.796	0.783	Money
						-0.278	-0.313	Ingestion
Success	0.252	0.333						
Sadness	-0.281	-0.325						
Happiness	-0.691	-0.651	V14	0.298	W14	-0.580	-0.604	Positive Emotion
Thinking	0.354	0.354						
Friendliness	-0.346	-0.315						
Success	0.666	0.663	V15	0.261	W15	0.456	0.505	Achievement
						0.366	0.435	Analytical Thinking
						-0.335	-0.314	Home
Auditory	0.298	0.358						
Dead	0.638	0.580	V16	0.227	W16	0.522	0.505	Death
Social Interactions	0.731	0.381						

Table 4 continues

HVdC Classes	α	r_X	CV_X	r_{XY}	CV_Y	β	r_Y	LIWC Categories
						0.582	0.406	Motion
						-0.375	-0.401	Achievement
						0.388	0.383	Anger
Thinking	0.679	0.577	V17	0.204	W17	0.582	0.400	Cognitive Process
Household	-0.547	-0.433						
Friendliness	0.496	0.338						
						0.483	0.370	Hear
						0.603	0.367	Motion
Thinking	0.408	0.444						
Household	0.457	0.442	V18	0.169	W18	0.584	0.535	Home
Happiness	0.452	0.412				0.493	0.394	Positive Emotion
Friendliness	-0.347	-0.305						

Note. The standardized canonical coefficients (weights) for the HVdC classes are in column α. The correlation between an HVdC class and an HVdC canonical variate is in column as r_X. The canonical correlation between a pair of canonical variates is in column r_{XY}. The standardized canonical coefficients (weights) for the LIWC categories are in column β; The correlation between a LIWC category and a LIWC canonical variate is in column r_Y.

denominator *df* = 9670.5, $p < 0.001$). Then, the second test is about whether the canonical correlations 2 to 21 are all 0 and the result leads to rejection of this hypothesis ($F = 10.060$, numerator *df* = 420, denominator *df* = 9232.8, $p < 0.001$). Finally, results of the 19th test show the canonical correlations 19, 20, and 21 are not significantly different from 0 ($F = 0.380$, numerator *df* = 2, denominator *df* = 650.0 $p = 0.681$). Hence, except canonical correlations 19, 20, and 21, all others are significantly different from 0. In applications it is unusual for many pairs of canonical variates to have significant canonical correlations (Tabachnick & Fidell, 2013). The canonical correlations indicate good correspondence between similar HVdC classes and LIWC categories.

An HVdC canonical variate is a linear combination of HVdC variables. The weights (coefficients) are in column α of Table 4. Because the variables are standardized, the coefficients are sometimes called standardized canonical coefficients. The first two rows give the first HVdC canonical variate, V1, as

$$\text{V1} = .901\text{Female} - .305\text{Male} + \text{lesser terms.}$$

Likewise, weights of LIWC canonical variates are in column β of Table 4. They give the first LIWC canonical variate, W1, as

$$\text{W1} = .956\text{Female references} - .271\text{Male references} + \text{lesser terms.}$$

The weights of Female and Female references are quite close, as are those of Male and Male references. For other canonical variates, numerical values of weights are often approximately equal for similar HVdC classes and LIWC categories. A good example is the fourth canonical variate pair. If we round to the first decimal and ignore lesser terms,

$$\text{V4} = -.7\text{Family} + .4\text{Body parts} - .4\text{Apprehension}$$

$$\text{W4} = -.8\text{Family} + .4\text{Body} - .4\text{Anxiety.}$$

And for the fifth canonical variate pair,

$$\text{V5} = .7\text{Apprehension} - .6\text{Family}$$

$$\text{W5} = .7\text{Anxiety} - .6\text{Family.}$$

Table 5. Canonical Correlations and Significance Levels for All Sets of Canonical Correlations

		Test of H0: The canonical correlations in the current row and all that follow are zero				
No.	r_{XY}	Likelihood Ratio	Approx. *F* Value	Num *df*	Den *df*	*p*
1	0.887	0.001	12.260	462	9670.5	<.0001
2	0.816	0.004	10.060	420	9232.8	<.0001
3	0.746	0.013	8.630	380	8794.2	<.0001
4	0.699	0.028	7.630	342	8354.7	<.0001
5	0.656	0.055	6.790	306	7914.3	<.0001
6	0.562	0.097	6.060	272	7473.1	<.0001
7	0.544	0.142	5.690	240	7031.0	<.0001
8	0.511	0.201	5.280	210	6588.2	<.0001
9	0.490	0.273	4.900	182	6144.6	<.0001
10	0.445	0.359	4.460	156	5700.3	<.0001
11	0.424	0.448	4.110	132	5255.5	<.0001
12	0.379	0.546	3.690	110	4810.2	<.0001
13	0.361	0.637	3.330	90	4364.5	<.0001
14	0.298	0.732	2.860	72	3918.8	<.0001
15	0.261	0.804	2.570	56	3473.4	<.0001
16	0.227	0.863	2.300	42	3028.8	<.0001
17	0.204	0.910	2.060	30	2586.0	0.001
18	0.169	0.949	1.690	20	2146.8	0.028
19	0.121	0.977	1.240	12	1714.7	0.252
20	0.083	0.992	0.870	6	1298.0	0.515
21	0.034	0.999	0.380	2	650.0	0.681

Note. Canonical correlation coefficient is denoted r_{XY}; Num *df* stands for *F* numerator's degree of freedom; Den *df* stands for *F* denominator's degree of freedom.

The reader can easily find others. Close numerical values are not forced by the canonical correlation algorithm. If so they would always be equal, but there are exceptions, as in V11 and W11. Rather, close weight values occur because of good correspondence in the data between similar HVdC classes and LIWC categories.

Aspects of results indicate that HVdC classes and LIWC categories with similar labels correspond well qualitatively and quantitatively. It is tempting to conclude that each can be used instead of the other. But do the HVdC classes account for all that the LIWC categories reveal, and vice versa?

The canonical correlation analysis provides answers in the following way. Each HVdC variable has a variance. There are 23 of them. The aggregate HVdC variance is the sum of these 23 variances. Likewise, the aggregate LIWC variance is the sum of the variances of the 21 LIWC variables. (Because the variables were standardized, the aggregate HVdC variance is 23 and the aggregate LIWC variance is 21.) The proportion of the aggregate HVdC variance explained by an HVdC canonical variate can be calculated. Of more interest is calculating the redundancy, which is the proportion of aggregate LIWC variance explained by an HVdC canonical variate. A high redundancy indicates a high prediction ability of HVdC variates to LIWC variates.

Table 6 and Table 7 show the aggregate variances of the HVdC classes and LIWC categories explained by the canonical variates. (Because the variables were standardized, the aggregate variances are sometimes called standardized variances.) We can see the proportion of the aggregate variance of the variables explained by a single canonical variate under the column "Proportion". For example, in Table 6 the proportion of the aggregate variance of HVdC classes explained by the first HVdC CV_X is 0.055 and by the first LIWC CV_Y is 0.043. Additionally, "Cumulative Proportion" records the proportion of the aggregate variance of the variables explained by one canonical variate together with prior canonical variates. For example, the proportions of aggregate variances of HVdC classes explained by the first two HVdC CV_X s are 0.055 and 0.064 separately, so the cumulative proportion of aggregate HVdC variance explained by the first two HVdC canonical variates is 0.119. The cumulative proportions in the last row of each table are the proportions explained by all 18 significant canonical variates. As shown in Table 6 and Table 7, the cumulative variance explained of the HVdC classes by the 18 HVdC CV_X s is 84.2 %; and that explained of the LIWC categories by the 18 LIWC CV_Y s is 89.8 %. These high proportions indicate that a substantial portion of the variance of the HVdC classes and LIWC categories can be explained by their respective significant canonical variates.

On the other hand, from the bottom rows of Table 6 and 7, we see that only 25.1% of HVdC classes' aggregate variance can be explained by LIWC canonical variates CV_Ys and only 27.6% of LIWC categories' aggregate variance can be explained by HVdC CV_Xs. Overall, results of the canonical correlation analysis show high correspondence between the observations of similar items in the LIWC and HVdC systems, which gives a strong signal of the usability of LIWC in dream research. But large aggregate variance of each system left unexplained by the other indicates that neither system can be replaced by the other.

Discussion

Use of automatic text analysis tools is increasing in dream research. Work is needed to examine whether the results of automatic analysis make sense and are consistent with what is found by traditional methods. The most popular traditional coding system for dream content is that of Hall and Van De Castle (1966), which we refer to as HVdC. A widely used text analysis system, starting to be used in dream studies, is Linguistic Inquiry and Word Count (Pennebaker, Boyd, Jordan & Blackburn, 2015), commonly called LIWC. Many HVdC classes and LIWC categories have similar labels, such as "social interactions" and "social processes". We compared outputs of the two systems for these similar items on the norm dreams of Hall and Van De Castle (1966), using canonical correlation analysis. Results showed a good correspondence between the content analysis outputs from LIWC and the HVdC coding system. When items in the two systems have similar labels, results for them are consistent. On the other hand, only about 25 % of the aggregate variance of items in one system can be explained by the canonical variates in the other system. So one system cannot replace the other, even when considering only their similar items.

A canonical correlation analysis finds composites of variables. Researchers using LIWC and HVdC sometimes form composites. LIWC 2015 includes some, called summary categories, that may be of interest to dream researchers (see, e.g., Bulkeley & Graves, 2018; Zheng & Schweickert, 2021a). Emotional tone is equal to the LIWC scores of positive emotion words minus those of negative emotion words and hence indexes the overall emotional positivity of a text (Cohn, Mehl, & Pennebaker, 2004). The summary category "Analytic" is composed of eight function word categories (article, preposition, personal pronouns, impersonal pronouns, auxiliary verbs,

Table 6. Aggregate Variance of HVdC Classes Explained by Canonical Variates CV_X and CV_Y

	CV_X		CV_Y	
CV No.	Proportion	Cumulative Proportion	Proportion	Cumulative Proportion
1	0.055	0.055	0.043	0.043
2	0.064	0.119	0.042	0.086
3	0.054	0.173	0.030	0.116
4	0.049	0.222	0.024	0.140
5	0.041	0.262	0.018	0.157
6	0.051	0.313	0.016	0.173
7	0.043	0.356	0.013	0.186
8	0.046	0.402	0.012	0.198
9	0.052	0.453	0.012	0.210
10	0.046	0.499	0.009	0.219
11	0.044	0.543	0.008	0.227
12	0.041	0.584	0.006	0.233
13	0.050	0.634	0.007	0.240
14	0.044	0.678	0.004	0.243
15	0.044	0.722	0.003	0.246
16	0.036	0.758	0.002	0.248
17	0.038	0.796	0.002	0.250
18	0.046	0.842	0.001	0.251

Table 7. Aggregate Variance of the LIWC Categories Explained by Canonical Variates CV_Y and CV_X

	CV_X		CV_Y	
CV No.	Proportion	Cumulative Proportion	Proportion	Cumulative Proportion
1	0.061	0.061	0.048	0.048
2	0.074	0.135	0.049	0.097
3	0.062	0.197	0.034	0.132
4	0.048	0.244	0.023	0.155
5	0.048	0.292	0.021	0.175
6	0.058	0.350	0.018	0.194
7	0.045	0.395	0.013	0.207
8	0.054	0.448	0.014	0.221
9	0.051	0.500	0.012	0.233
10	0.055	0.555	0.011	0.244
11	0.045	0.600	0.008	0.252
12	0.039	0.639	0.006	0.258
13	0.047	0.686	0.006	0.264
14	0.042	0.727	0.004	0.268
15	0.051	0.779	0.004	0.271
16	0.043	0.821	0.002	0.273
17	0.039	0.860	0.002	0.275
18	0.038	0.898	0.001	0.276

conjunctions, adverbs, and negation) in LIWC (Jordan, Sterling, Pennebaker, & Boyd, 2019; Pennebaker, Chung, Frazee, Lavergne, & Beaver, 2014). A high score in “Analytic” reflects an emphasis in conveying relationships among concepts, while a low score reflects a narrative experiential style (Jordan et al., 2019; Pennebaker et al., 2014). The summary category "Clout" indexes the relative social status among characters in a text (Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2014).

In the HVdC system, composites are often formed as numerators or denominators of content indicators, see, e. g., Domhoff (2003, Table 24). For example, the Family percent is the sum of counts in classes Family and Relatives divided by the sum of counts of all humans. The Negative emotions percent is the count of negative emotions divided by the count of all emotions; it is akin to emotional tone in LIWC. For more information, see Domhoff (2003), p. 74-75.

The composites found by the canonical correlation analysis are worthy of note. Hall and Domhoff (1963) reported that women dream equally often of men and women, and men dream more often of men than of women. The percent of female and male characters has continued to be of interest to dream researchers (e.g., Hall, 1984; Paul & Schredl, 2012; Schredl, 2001). The first HVdC canonical variate (approximately .9Female - .3Male) is sensitive to the proportions of female and male characters. It correlates highly with the first LIWC canonical variate (approximately .9Female references - .3Male references). This suggests that for a researcher investigating the proportions of female and male characters an automatic analysis by LIWC would be a good approximation to the more direct HVdC analysis.

Linear combinations produced by the canonical correlation analysis suggest comparing dream reports in terms of classes one might not spontaneously think of combining. One example is pair 5, Anxiety minus Family. Another is pair 12, Death minus Anger. Good correspondence between HVdC classes and LIWC categories with similar labels suggests it is reasonable to explore questions about such composites with relatively fast LIWC analysis and then confirm with more time consuming HVdC coding.

The limitations of using LIWC for dream studies have been discussed by Bulkeley and Graves (2018). For example, texts recording dreams are not often “clean”, which means free of typographical errors, associational comments, and idiosyncratic language. The prerequisite of “clean” texts is, however, important for LIWC to do the best word frequency analysis work. The dream reports used in the current study were collected from 1947 to 1950. It is an open question whether LIWC, a tool updated in 2015, is optimal for texts from an older time. Last but not least, the categories of LIWC are predetermined and cannot cover all important topics of dreams. An appropriate combined use of LIWC and HVdC may be able to provide dream researchers with a more comprehensive measurement of dream contents. The word search technique (Domhoff & Schneider, 2008; Bulkeley, 2009, 2014; Bulkeley & Domhoff, 2010) has terms especially selected for relevance to dream studies and is worth considering for automatic dream content analysis.

A limitation of the Hall Van De Castle (1966) system is that when a character is coded the name of the character is not retained. Yet it is of interest to keep track of a character through a series of dreams (e.g., Schweickert, Xi, Viau-Quesnel & Zheng, 2020). For this purpose, a researcher using LIWC can create a custom dictionary (Pennebaker, et al, 2015) and a researcher using the word search technique can directly search for character names (e.g., Han, 2014).

The ability of canonical correlation analysis to detect relationships between two sets of variables fits the nature of many dream research methodology topics. Despite its complexity, canonical correlation analysis is able to reveal rich information about relationships among multivariate variables. In a canonical correlation analysis of HVdC classes and LIWC categories with similar labels, we find that corresponding individual items line up remarkably well in their contributions to canonical variates. A large number of the canonical correlations are significant, 18 of 21, indicating that many composites of HVdC variables correlate well with composites of LIWC variables. However, aggregate variance of one system explained by canonical variates of the other is about .25, not negligible, but not large. The two systems do not overlap completely, but where they do overlap they align well.

Overall, results of Chapter 1 indicate a high potential of LIWC for dream content analysis. We also gave an intuitive introduction of CCA to lower the statistical barrier for dream researchers to use multivariate methods. In Chapter 2, we used word frequency data by LIWC in dream reports to test the continuity hypothesis of dream research regarding gender differences, aging effect, and the difference between the blind and the sighted. What is more, we built support vector machine models to achieve a binary characteristics prediction of people (female or male, blind or sighted) based on the word frequencies of all LIWC categories in dream reports.

CHAPTER 2: TEST OF CONTINUITY HYPOTHESIS BY AUTOMATIC DREAM CONTENT ANALYSIS AND DETECTION OF BINARY CHARACTERISTICS OF DREAMERS

Introduction

People have been curious about the association between dreams and waking life for thousands of years. In the western world, Artemidorus compiled *Oneirocritica* (The Interpretation of Dreams), which provided an encyclopedic explanation for dreams, in the 2nd century AD. In our time, scientific studies have shown that dream content is meaningfully related to the personal concerns in waking life (e.g., Domhoff & Schneider, 2008a, 2008b, 2018). A longstanding hypothesis regarding the relationship between wakefulness and dreaming is the continuity hypothesis (Hall & Nordby, 1972), which is "There is considerable congruence between what a person dreams about at night and what he does or thinks about when he is awake" (Hall & Nordby, 1972, p. 125). This hypothesis has gained much support from dream research (e.g., Domhoff, 1996; Schredl & Hofmann, 2003; Strauch & Meier, 1996) in terms of people's personality features (e.g., Hawkins & Boyd, 2017), events in life (e.g., Proksch & Schredl, 1999), beliefs (e.g., Bulkeley, 2009), emotions (e.g., Hartmann, 2007), and social contents (e.g., Schweickert, Xi, Viau-Quesnel, & Zheng, 2020). However, the link between wakefulness and dreaming is still far from well understood.

Linguistic Inquiry and Word Count (LIWC) is software for text content analysis (Pennebaker, Boyd, Jordan, & Blackburn, 2015). It can automatically tabulate the frequencies of various word categories in a text once the text is imported into LIWC. LIWC is well-validated for measuring various psychological features in language usage (Tausczik & Pennebaker, 2010; Pennebaker, Chung, Frazee, Lavergne, & Beaver, 2014) and could be useful for testing the continuity hypothesis by analyzing word frequencies in dream reports. The first aim of the current study is to test the consistencies between waking life and dreams in terms of gender, age, and blind vs. sighted differences by word frequency analysis using LIWC.

Machine learning (ML) is one of the most rapidly developing techniques in 21st century. ML algorithms, such as the support vector machine (SVM), have benefitted a variety of areas, such as language processing, image and graph analysis, biology, and mental health (e.g., Gaonkar & Davatzikos, 2013; Sacchet, Prasad, Foland-Ross, Thompson, & Gotlib, 2015). Even though ML has occasionally been applied to dream research, it deserves further investigation in dream studies to achieve the automatic analysis of data of huge quantities or unstructured data. Hence, our second aim is to build SVM models to achieve a binary characteristics detection of a person based on the content of the person's dreams.

Appling LIWC to Dream Studies

With the rapid development of science and technologies, dream researchers have developed tools to quantitatively measure the contents of dream, for example, the Hall Van de Castle (HVdC) coding system (Hall & Van de Castle, 1966) and a Word Search Technology based on the search engine on DreamBank.net (Bulkeley, 2009; Bulkeley, 2014; Domhoff & Schneider, 2008a, 2008b; (Schneider & Domhoff, 2019). The HVdC system has unique coding rules for different aspects of dream content, for example, social interactions and characters, and it is one of the most used methods for coding dream contents. However, the HVdC system relies heavily on human hand-coding, which may bring in coding bias and also be slow in coding progress. If automatic coding methods were feasible and useful for dream contents, then we believe that the progress of dream research would be faster, and coding results would be more consistent.

LIWC2015 is software for analyzing texts (Pennebaker, Boyd, Jordan, & Blackburn, 2015). It can automatically tabulate the frequencies of words for about 100 different categories, e.g., grammar, vocabulary, emotion, social processes, time focus, and so on, in a text once the text is imported into LIWC. LIWC has been tested as a well-validated tool to measure various psychological constructs and is widely used in psychological studies (e.g., Park & Conway, 2017; Tov, Ng, Lin, & Qiu, 2013).

LIWC has also been used as in dream research, though not much (e.g., McNamara, Pae, Teed, Tripodis, & Sebastian, 2016; Wong, Amini, & De Koninckand, 2016). Zheng and Schweickert (2021a), Chapter 1 here, showed a high compatibility of outputs by LIWC and HVdC system on coding dream contents, which indicated a good usability of LIWC for dream researchers and provided an impetus for the current study. Accordingly, we applied LIWC to analyze dream contents in terms of word frequencies in the current study.

Study 1: Gender Differences

Introduction

Men and women use words differently in waking life. For example, Pennebaker (2011) found that women use more the first-person singular pronoun (I), cognitive words, social words, personal pronouns, verbs, negative emotion, negations, certainty words, and hedge phrases; men use more articles, big words, nouns, prepositions, numbers, words per sentence, and swear words[1]. Some studies have analyzed the difference of word usage in dream reports by females and males (e.g., Wong, Amini, & De Koninck, 2016). For example, McNamara, Pae, Teed, Tripodis, and Sebastian (2016) analyzed dream reports with LIWC and found that word usage about cognitive processing was significantly associated with words of verbs and function words, the personal pronoun I, social processes, health and emotion, and perceptual processes. In another example, Wong, Amini, and De Koninckand (2016) developed a computer program to differentiate the dreams of females from males based on the gender differences in language features (Mulac, Bradac, & Gibbons, 2001) and emotional tone of dream reports (Amini, Sabourin, & De Koninck, 2011).

As examples of gender differences in waking life, studies using various methods found that males seem to be more analytic, to have better spatial ability, to be more reward-oriented, to take more risks (Byrnes, Miller, & Schafer, 1999; Charness & Gneezy, 2012), to have more sexual fantasies, and to be more easily physically aroused by their sexual thoughts (Ellis & Symons,1990; Knoth, Boyd, & Singer, 1988). By contrast, women tend to have higher levels of distress and to be more likely to have depression and anxiety (World Health Organization, 2002), and to reserve more cognition for family and home issues (Sharma, Chakrabarti, & Grover, 2016). Are these waking life differences reflected in the word usage of dream content?

Based on the above gender differences in waking life and results from Pennebaker (2011), we made hypotheses as follows:

> Hypothesis 1: In males' dreams, the following word categories have higher frequencies than in females' dreams: Articles, Words > 6 letters, Prepositions, Numbers, Words/sentence, Swear words, Analytical thinking, Reward, Risk, Space, Sexual.

[1] Categories of "hedge phrase" and "noun" are not available in LIWC2015.

Hypothesis 2: In females' dreams, the following word categories have higher frequencies than in males' dreams: 1st pers singular (I), Cognitive processes, Social processes, Personal pronouns, Common verbs, Negative emotions, Negations, Certainty, Anxiety, Family, Home.

Method

We used the classic Hall and Van de Castle (1966) norm dreams of female (n = 490) and male (n = 491) students, available on DreamBank.net (Schneider & Domhoff, 2019), then analyzed the word frequencies of the categories mentioned in hypotheses. We used two-sample t tests to examine the differences of the word frequencies in dreams of females and males. For the comparison of each LIWC category, Levene's test was applied to test the equality of variances of two groups. The null hypothesis of Levene's test is that the two population variances are equal. If the calculated F value is significant, then we should reject the null hypothesis, and conclude there is a difference between the variances in the population. Hence, in the two-sample t tests, we should use formulas for unequal variances when the two population variances are shown to be unequal. Otherwise, we should use formulas for equal variances. Additionally, because of multiple comparisons, we applied the controlling procedure of False Discovery Rate at 0.05 to control Type I error.

Results

See Table 8 for the results of the gender difference tests. When saying a difference is significant, we mean it is significant following the False Discovery Rate procedure. Supporting Hypothesis 1, for the categories of "Articles" (mean difference = -6.696, p < .001), "Words > 6 letters" (mean difference = -2.529, p = .012), "Prepositions" (mean difference = -0.730, p < .001), "Analytical thinking" (mean difference = -10.576, p < .001), "Risk" (mean difference = -0.151, p < .001), "Space" (mean difference = -0.662, p = .004), and "Sexual" (mean difference = -0.085, p < .001), the word frequencies in males' dream reports are significantly higher than those in females' dream reports. For the other four categories specified in Hypothesis 1, which are "Reward", "Words/sentence", "Swear", and "Numbers", the results do not support Hypothesis 1. In particular, males used significantly fewer words per sentence, contrary to Hypothesis 1.

Table 8. Results of Gender Difference Tests

	Levene's Test		*t*-test					
Categories	*F*	Sig.	*t*	Mean Diff.	*S.E.*	*d.f.*	Sig.	Sig. after FDR
Analytical thinking	3.086	.079	-7.236	-10.576	1.462	979	.000	YES
Articles	4.801	.029	-6.696	-1.2616	.188	966.885	.000	YES
Common verbs	3.448	.064	6.681	1.524	.228	979	.000	YES
Social processes	.135	.714	5.198	1.733	.333	979	.000	YES
Family	43.384	.000	5.176	.530	.102	826.549	.000	YES
Prepositions	.489	.485	-3.925	-.730	.186	979	.000	YES
Risk	28.701	.000	-3.677	-.151	.041	918.268	.000	YES
Sexual	54.903	.000	-3.684	-.085	.023	659.572	.000	YES
Negations	.013	.910	3.035	.227	.075	979	.002	YES
Words/sentence	3.569	.059	3.011	.872	.290	979	.003	YES
Cognitive processes	.251	.616	2.989	.779	.261	979	.003	YES
Space	.001	.975	-2.861	-.662	.231	979	.004	YES
Anxiety	9.946	.002	2.923	.129	.044	961.934	.004	YES
Certainty	3.980	.046	2.546	.182	.071	970.938	.011	YES
Words > 6 letters	6.558	.011	-2.529	-.679	.269	970.106	.012	YES
Home	3.154	.076	2.284	.250	.109	979	.023	YES
Personal pronouns	7.791	.005	2.133	.570	.267	954.741	.033	YES

Table 8 continues

	Levene's Test		*t*-test					
Categories	*F*	Sig.	*t*	Mean Diff.	*S.E.*	*d.f.*	Sig.	Sig. after FDR
Reward	4.452	.035	-1.805	-.118	.065	967.622	.071	NO
Negative Emotions	1.593	.207	-1.011	-.097	.096	979	.312	NO
Swear	3.539	.060	-.952	-.008	.009	979	.341	NO
1st pers singular	.659	.417	-.371	-.084	.227	979	.711	NO
Numbers	.253	.615	.128	.011	.086	979	.898	NO

Note. FDR stands for False Discovery Rate procedure.

Results largely support Hypothesis 2. For categories of “Cognitive process” (mean difference = 2.989, p = .003), “Social processes” (mean difference = 5.198, p < .001), “Personal pronouns” (mean difference = 2.133, p = .033), “Common verbs” (mean difference = 6.681, p < .001), “Negations” (mean difference = 3.035, p = .002), “Certainty” (mean difference = 2.546, p = .011), “Anxiety” (mean difference = .12, p = .005), “Family” (mean difference = .53, p < .001), and “Home” (mean difference = 2.284, p = .023), the word frequencies in females’ dream reports are significantly higher than those in males’ dream reports. For only two categories specified in Hypothesis 2, “1st pers singular” and “Negative emotions”, the differences are not significant.

Discussion

Previous literature reports that men and women use words differently in daily life. Further, there are gender differences in specific LIWC word category frequencies in texts written while awake (Pennebaker, 2011). If these differences occur in dream reports, for 11 LIWC categories men would have higher frequency than women in dream reports. We found such differences in 7 of the 11. For three of the others, there was no significant difference and for one of the others there was a significant difference in the opposite direction. Also, if the previously found gender differences occur in dream reports, for 11 LIWC categories women would have higher frequency than men in dream reports. We found such differences in 9 of the 11. For the other two there was no significant difference. On the whole, our results indicate that gender differences in waking life and in writing while awake also exist in dreams, providing evidence for the continuity hypothesis.

Study 2: Aging Effect

Introduction

Researchers have discovered that in waking life, aging does not produce decrements in positive emotions (e.g., Carstensen, Pasupathi, Mayr, & Nesselroade, 2000) and negative affect declines across the life span (e.g., Charles, Reynolds, & Gatz, 2001). This is possibly because people tend to become better at regulating their emotions as they grow older (Carstensen, 1995). This effect has been further supported in waking life by analyzing the word usage in social media and diaries, showing that people tend to be less negative with the increase of age (e.g., Mao, Stillwell, Kosinski, & Good, 2017; Pennebaker & Stone, 2003). However, the aging effect has rarely been investigated in dream contents by word frequency analysis.

Some studies tried to find the developmental pattern of dream content (e.g., Dale, Lafrenière, & De Koninck, 2017; Dale, Lortie-Lussier, & De Koninck, 2015) by hand-coding dream reports from people in different age groups using HVdC coding rules. For example, Dale, Lortie-Lussier, and De Koninck (2015) recruited female subjects of five age groups from adolescence to old age. They coded their dreams for given aspects, e.g., characters, activities, interactions and so on, and used trend analysis to explore the change of dream's pattern across time. Several significant results were found. For instance, from the youngest to the oldest, the total number of aggressive interactions went down. Dale, Lafrenière, and De Koninck (2017), applied a similar method to male's dreams and found several patterns of dream imagery that reflected the waking developmental patterns. For instance, aggressive dream imagery predominates in adolescent age group. Dale et al. (2017) provide support for aging differences in dreams but did not investigate the change of dream content across one individual's life span, that is, longitudinally.

In Study 2, we aimed to detect how dream contents change with the increase of age. We analyzed the word frequencies in dream reports of three persons who recorded dreams over long intervals (53 years, 41 years, and 39 years respectively). We considered the following questions. Will the frequency of words about negative emotion in dream reports decrease with the increase of people's age? In LIWC, the category of negative emotion includes three sub-categories: anxiety, sadness, and anger. How would the frequencies of these subcategories change over time? Also, Pennebaker and Stone (2003) found that with increasing age, individuals use fewer self-references, more future-tense and fewer past-tense verbs, and demonstrate a general pattern of increasing usage of cognitive complexity (words longer than 6 letters, total cognitive words, causal, insight, and exclusive words[2]). Will these patterns of word usage appear in dream reports? Based on these previous findings, we made hypotheses and exploratory questions regarding aging effect as follows:

> Hypothesis 3 (aging effect): With the increase of age, the frequency of words about negative emotion decreases, and the frequency of words about positive emotion keeps flat or increases.

[2] The LIWC category for exclusive words is no longer available in LIWC2015.

Hypothesis 4 (comparison with Pennebaker & Stone, 2003): With the increase of age, the frequency of words about self-references and past-tense verbs decreases, and the frequency of words about future-tense and cognitive markers (cognitive mechanisms and causal, insight, and exclusive words) increases.

Exploratory Question:

Which kind of negative emotions account for the decrease in negative emotion, anger, anxiety, or sadness?

Method

On DreamBank, at this time, there are only three dreamers who reported dreams over long durations that can be used to test the aging effect. The three dreamers are: Barb Sanders (total 4504 reports, covering 41 years), Dorothea (total 900 reports, covering 53 years), and Emma (total 1221 reports, covering 39 years). All are female. All names are pseudonyms. More information about the three dreamers and their dream reports can be found on Dreambank. We analyzed the word frequencies of the categories mentioned above in the three dreamers' reports. Then, we built linear regression models using report dates, a representation of age, as the independent variable and word frequencies of a LIWC category as the dependent variable to investigate the variation of the word frequencies in dreams across time.

Results

See Table 9 for results for Barb Sanders. Frequencies of "Negative emotions" words (*B* = -4.623e-05, *S.E.* = 1.555e-05, *t* = -2.97**) were negatively related with time, while frequencies of "Positive Emotions" words (*B* = 4.472e-05, *S.E.* = 1.943e-05, *t* = 2.30*) were positively associated with time. Hypothesis 3 is supported here. For the three kinds of negative emotions, only word frequencies of "Anxiety" (*B* = -1.831e-05, *S.E.* = 6.262e-06, *t* = -2.92**) and "Sadness" (*B* = -1.313e-05, *S.E.* = 6.484e-06, *t* = -2.03*) were significantly negatively associated with time. For categories listed in Hypothesis 4, the word frequencies of "Focus on past" (*B* = -4.161e-04, *S.E.* = 2.320e-05, *t* = -17.94***), "First-person singular" (*B* = -1.847e-04, *S.E.* = 2.796e-05, *t* = -6.61***), and "First-person plural" (*B* = -2.739e-05, *S.E.* = 1.333e-05, *t* = -2.05*) decreased across time. The word frequencies of "Big words" increased across time (*B* = 1.203e-04, *S.E.* = 3.077e-05, *t* =

3.91***). These results are consistent with Hypothesis 4. However, for categories of "Focus on future" (*B* = -3.981e-05, *S.E.* = 1.072e-05, *t* = -3.715***) and "Causation" (*B* = -2.175e-05, *S.E.* = 8.299e-06, *t* = -2.62**), the word frequencies decreased with the increase of time, inconsistent with what is predicted by Hypothesis 4. For word frequencies of "Cognitive processes" and "Insight", the effect of time was not significant.

For all categories, the proportion of variance accounted for as a linear function of age (R^2) is very small. This is understandable because the biggest influence in dreams on, say, positive emotions, is the content of the particular dreams. The age of the dreamer would be expected to have little additional effect.

See Table 10 for results for Dorothea. As was found for Barb Sanders, the frequency of "Negative emotions" words (*B* = -4.632e-05, *S.E.* = 1.233e-05, *t* = -3.76***) was negatively related with time. The effect of time for predicting the frequencies of "Positive Emotions" words was not significant. Hypothesis 3 is supported again. For the three kinds of negative emotions, only word frequency of "Anger" (*B* = -2.449e-05, *S.E.* = 5.162e-06, *t* = -4.74***) was significantly negatively associated with time. This is inconsistent with what we found in Barb Sanders' dreams. For categories listed in Hypothesis 4, the word frequencies of "Focus on past" (*B* = 3.756e-04, *S.E.* = 5.336e-05, *t* = 7.04***) and "First-person singular" (*B* = 1.752e-04, *S.E.* = 3.186e-05, *t* = 5.50***) were positively correlated with time. The word frequencies of "Big words" decreased across time (*B* = -1.909e-04, *S.E.* = 4.214e-05, *t* = -4.53***). These results are inconsistent with Hypothesis 4. However, for the category of "Cognitive processes" (*B* = 7.574e-05, *S.E.* = 3.805e-05, *t* = 1.99*), the word frequencies increased across time, which was consistent with what predicted by Hypothesis 4. For word usage of "First-person plural", "Focus on future", "Insight", and "Causation", the effect of time was not significant.

Again, for all categories, the proportion of variance accounted for as a linear function of age (R^2) is very small, a not unreasonable outcome.

Table 9. Linear Associations Between Word Frequencies of LIWC Categories and Age: Barb Sanders

		Intercept			Slope			
Variable	a	*SE*	*t*	b	*SE*	*t*	R^2	Adjusted R^2
Emotional State								
Positive Emotions	1.172	.656	1.79	4.472e-05	1.943e-05	2.30*	.002	.002
Negative Emotions	3.496	.525	6.65***	-4.623e-05	1.555e-05	-2.97**	.003	.003
Anxiety	1.088	.212	5.15***	-1.831e-05	6.262e-06	-2.92**	.003	.003
Anger	.764	.264	2.89*	-6.719e-06	7.810e-06	-.86	.000	.000
Sadness	.832	.219	3.80***	-1.313e-05	6.484e-06	-2.03*	.002	.001
Self-References								
First-Person Singular	11.520	.945	16.15***	-1.847e-04	2.796e-05	-6.61***	.017	.016
First-Person Plural	1.943	.450	4.31***	-2.739e-05	1.333e-05	-2.05*	.002	.001
Time Orientation								
Focus on Past	16.090	.784	20.53***	-4.161e-04	2.320e-05	-17.94***	.110	.110
Focus on future	2.838	.362	7.84***	-3.981e-05	1.072e-05	-3.715***	.005	.005
Cognitive Complexity								
Big Words (>6 letters)	7.343	1.039	7.15***	1.203e-04	3.077e-05	3.91***	.006	.005
Cognitive Processes	6.468	.914	7.07***	5.037e-05	2.707e-05	1.86	.001	.001
Causation	1.703	.280	6.07***	-2.175e-05	8.299e-06	-2.62**	.003	.002
Insight	1.527	.442	3.45***	1.713e-05	1.309e-05	1.31	.001	.000

Note. $*p \leq .05$. $**p \leq .01$. $***p \leq .001$.

Table 10. Linear Associations Between Word Frequencies of LIWC Categories and Age: Dorothea

Variable	a	SE	t	b	SE	t	R^2	Adjusted R^2
		Intercept			Slope			
Emotional State								
Positive Emotions	1.314	.255	5.51***	-1.587e-05	1.216e-05	-1.31	.003	.001
Negative Emotions	2.035	.259	7.87***	-4.632e-05	1.233e-05	-3.76***	.023	.021
Anxiety	.303	.083	3.64***	-7.256e-06	3.966e-06	-1.83	.006	.004
Anger	.669	.108	6.18***	-2.449e-05	5.162e-06	-4.74***	.036	.035
Sadness	.393	.149	2.64**	-2.662e-07	7.109e-06	-0.04	2.354e-06	-.002
Self-References								
First-Person Singular	3.920	.668	5.87***	1.752e-04	3.186e-05	5.50***	.048	.047
First-Person Plural	.564	.316	1.79	2.189e-05	1.505e-05	1.45	.004	.002
Time Orientation								
Focus on Past	.081	1.119	0.07	3.756e-04	5.336e-05	7.04***	.077	.075
Focus on Future	1.502	.291	5.16***	-3.282e-06	1.388e-05	-0.24	9.386e-05	-.002
Cognitive Complexity								
Big Words (>6 letters)	15.230	.884	17.23***	-1.909e-04	4.214e-05	-4.53***	.033	.032
Cognitive Processes	7.620	.798	9.55***	7.574e-05	3.805e-05	1.99*	.007	.005
Causation	.748	.218	3.44***	8.939e-06	1.037e-05	.86	.001	.000
Insight	2.486	.336	7.394***	-2.909e-05	1.603e-05	-1.81	0.005	0.004

Note. *$p \leq .05$. **$p \leq .01$. ***$p \leq .001$.

See Table 11 for results of Emma. As for Barb Sanders and Dorothea, the frequency of "Negative emotions" words (*B* = -4.440e-05, *S.E.* = 1.351e-05, *t* = -3.29**) was negatively related with time. However, the frequency of "Positive Emotions" words was also negatively associated with time (*B* = -4.635e-05, *S.E.* = 1.704e-05, *t* = -2.72**), which was against Hypothesis 3. For the three kinds of negative emotions, only word frequency of "Anger" (*B* = -1.569e-05, *S.E.* = 6.825e-06, *t* = -2.29*) was significantly negatively associated with time. For categories listed in Hypothesis 4, the word frequency of "First-person plural" (*B* = 6.633e-05, *S.E.* = 1.579e-05, *t* = 4.20***) was positively correlated with time. The word usages of "Focus on future" (*B* = -3.899e-05, *S.E.* = 1.107e-05, *t* = -3.52***) and "Big words" (*B* = -9.057e-05, *S.E.* = 3.586e-05, *t* = -2.53*) decreased across time. These were inconsistent with Hypothesis 4. For word usages of "First-person singular", "Focus on past", "Cognitive processes", "Insight", and "Causation", the effect of time was not significant.

As with the first two dreamers, for all categories, the proportion of variance accounted for as a linear function of age (R^2) is very small, a not unreasonable outcome.

Discussion

Study 2 aimed to detect the relationship between aging and dream contents by word frequency analysis. We used dream reports of three persons who recorded dreams for tens of years. One statement in Hypothesis 3 was well supported by our results: For all three dreamers, the frequency of negative emotions words was significantly negatively associated with time. However, the other statement in Hypothesis 3 was not well supported: For one dreamer, the frequency of positive emotions was significantly positively associated with time, as the hypothesis states, but for the other two the association was significantly negative, contrary to the hypothesis. Hypothesis 4, about frequencies of several LIWC categories, was generally not supported. For some of the relevant LIWC categories, the effect of time on frequencies in dream reports is significant, but the signs of the linear coefficients are different from what found by Pennebaker and Stone (2003). However, the magnitudes of the linear coefficients are very small, near to zero in our results, so one could argue that the evidence against Hypothesis 4 is not strong. For the three kinds of negative emotions, there were different trends of word frequencies across time in the three dreamer's reports, so we cannot conclude that one emotion is more relevant than another for negative emotion decreases with age.

Table 11. Linear Associations Between Word Frequencies of LIWC Categories and Age: Emma

		Intercept			Slope			
Variable	a	*SE*	*t*	*b*	*SE*	*t*	R^2	Adjusted R^2
Emotional State								
Positive Emotions	3.571	.481	7.43***	-4.635e-05	1.704e-05	-2.72**	.008	.007
Negative Emotions	2.764	.381	7.26***	-4.440e-05	1.351e-05	-3.29**	.012	.011
Anxiety	4.839e-01	.172	2.82**	-4.804e-06	6.095e-06	-0.79	.001	.000
Anger	8.036e-01	.193	4.16***	-1.569e-05	6.825e-06	-2.29*	.006	.005
Sadness	5.326e-01	.193	2.76**	-5.583e-06	6.859e-06	-0.81	.001	.000
Self-References								
First-Person Singular	5.803	.751	7.73***	2.274e-05	2.662e-05	.85	.001	.000
First-Person Plural	6.770e-02	.445	0.15	6.633e-05	1.579e-05	4.20***	.019	.018
Time Orientation								
Focus on Past	8.205e-01	.371	2.21*	1.980e-05	1.314e-05	1.51	.003	.001
Focus on Future	2.252	.312	7.21***	-3.899e-05	1.107e-05	-3.52***	.014	.013
Cognitive Complexity								
Big Words (>6 letters)	1.501e+01	1.011	14.84***	-9.057e-05	3.586e-05	-2.53*	.007	.006
Cognitive Processes	5.327	.845	6.30***	3.789e-05	2.998e-05	1.26	.002	.001
Causation	8.935e-01	.261	3.42**	-4.679e-06	9.260e-06	-0.51	.000	.000
Insight	1.217	.377	3.23**	8.985e-06	1.336e-05	0.67	.001	.000

Study 3: Blind-Sighted Differences

Introduction

For sighted people, dreaming is mostly a visual experience, and some portions of dreams have auditory and tactual sensations (e.g., Snyder, 1970; Zadra, Nielsen, & Donderi, 1998). Blind people who were sightless since birth or very early childhood seem to have no visual perceptions in dreams and rely on tactual sensory references a lot, as in real life (Hurovitz, Dunn, Domhoff, & Fiss, 1999; Kennedy, 1993, 1997). Based on these earlier findings, we hypothesized that blind people use fewer visual words, more tactual words, and more auditory words than sighted people in dream reports.

We also hypothesized that the nature of blindness (congenital or adventitious blindness) and degree of blindness (total blindness, partial blindness, or perceiving only very bright light) would make differences in the frequencies of perception words (“See”, “Feel”, and “Hear”). Here, we combined “partial blindness” and “perceiving only very bright light” into “not total blindness”. Blind people’s dreams have been analyzed, even though not much, by quantitative approaches. For example, Bulkeley (2015, June) used the Word Search Technique to study a blind female dreamer’s dreams and made 21 inferences about her waking life based on word usage. For example, he inferred that hearing was an especially important sense for her. It turned out 15 of the 21 inferences were confirmed by the dreamer as accurate. The present study used the automatic content analysis by LIWC to conduct a systematic quantitative analysis on blind people’s dreams.

We made hypotheses regarding blind-sighted differences as follows. Considering the possible effect of gender differences, in this analysis, we also included gender as a variable.

> Hypothesis 5: Blind people use more words of tactus (LIWC category “Feel”) and audition (LIWC category “Hear”), and fewer words of vision (LIWC category See”) in dream reports than sighted people.
>
> Hypothesis 6: The nature of blindness (congenital or adventitious) and degree of blindness (total or not total) make differences in the frequencies of perception words in blind people’s dreams. Specifically, people who are congenitally blind or who are totally blind report fewer words of vision (LIWC category “See”) and more

words of tactus (LIWC category “Feel”) and audition (LIWC category “Hear”), than blind people who are blind adventitiously or who are not totally blind.

Method

We used dream reports of blind dreamers (in report sets of “Blind dreamers (F) [*n* = 238]”, “Blind dreamers (M) [*n* = 143]”, and “Edna: a blind woman [*n* = 19]”) and Hall Van de Castle norm dreams (female: *n* = 490; male: *n* = 491) on Dream Bank. The number of dream reports from female blind dreamers is n = 257, from male is *n* = 143. Information about blind dreamers and their reports is in Table 12. More demographic information about the blind dreamers can be found in Hurovitz, Dunn, Domhoff, and Fiss (1999).

We first analyzed the word frequencies of categories “See”, “Hear”, and “Feel” in the dream reports with LIWC. Then, we conducted two-way ANOVA analyses using the frequencies of word categories “See”, “Hear”, and “Feel” as dependent variables. To test Hypothesis 5, we considered the fixed effects of blindness (blind or sighted) and gender (male or female) in the ANOVA models; for Hypothesis 6, we considered the fixed effects of blind nature by gender and blind degree by gender in the ANOVA models.

Results

See Table 13, Table 14, and Table 15 for results of two-way ANOVA analyses with the dependent variables of frequencies of word categories “See”, “Hear”, and “Feel” and the independent variables of gender and blindness. The effect of blindness was significant in all of the three ANOVA models. As shown in Table 13, for the category of “See”, the frequency in dream reports by the blind people (mean estimate = 0.855) was significantly lower than that by the sighted dreamers (mean estimate = 1.605). For the category of “Hear” (Table 14), the frequency in dream reports by the blind people (mean estimate = 1.543) was significantly higher than that by the sighted dreamers (mean estimate = 0.783). Similarly, as shown in Table 15, words of the category “Feel” appeared significantly more frequently in blind dreamers’ reports (mean estimate = 0.834) than in sighted dreamers’ reports (mean estimate = 0.634). Results greatly supported Hypothesis 5.

Table 12. Information of Blind Dreamers and Reports

Dreamer	Gender	Age	Nature of Blindness	Degree of Blindness	# of Reports
1	F	15	A	T	10
2	F	18	C	T	10
3	F	12	A	N	39
4	F	19	C	N	59
5	F	12	C	T	37
6	F	14	A	T	6
7	F	18	C	T	32
8	F	18	A	T	10
9	F	12	A	T	24
10	F	13	C	T	9
11	F	Missing	C	Missing	19
12	M	16	C	T	61
13	M	18	A	N	21
14	M	14.5	A	T	20
15	M	13	A	T	22
16	M	12	C	T	12

Note. C = congenital blindness; A = adventitious blindness; T = total blindness; N = not total blindness.
The reports were retried on December 20th, 2020. There are minor errors in the information displayed on Dreambank.net regarding the number of dream reports by some blind dreamers.

Table 13. Results of Blind-Sighted Difference: Category of “See”

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Intercept	1617.121	1	1617.121	733.168	.000
Gender	22.919	1	22.919	10.391	.001
Blind	150.276	1	150.276	68.132	.000
Gender * Blind	2.820	1	2.820	1.278	.258
Error	3037.199	1377	2.206		
Total	5915.069	1381			

Table 14. Results of Blind-Sighted Difference: Category of “Hear”

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Intercept	1446.159	1	1446.159	727.665	.000
Gender	3.778	1	3.778	1.901	.168
Blind	154.337	1	154.337	77.658	.000
Gender * Blind	1.467	1	1.467	.738	.390
Error	2736.645	1377	1.987		
Total	4328.253	1381			

Table 15. Results of Blind-Sighted Difference: Category of "Feel"

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Intercept	575.943	1	575.943	479.559	.000
Gender	3.744	1	3.744	3.117	.078
Blind	10.749	1	10.749	8.950	.003
Gender * Blind	.110	1	.110	.091	.762
Error	1653.754	1377	1.201		
Total	2317.056	1381			

As for the effect of the nature of blindness on perceptual word usage in dreams, results are in Table 16, Table 17, and Table 18. Congenitally blind people used fewer words of "See" (mean estimate of frequency = 0.578) than adventitious blind people (mean estimate of frequency = 1.277), and the effect is significant, as shown in Table 16. For the category of "Hear", the effect of blindness nature is also significant (see Table 17). More specifically, congenitally blind people used more words of "Hear" (mean estimate of frequency = 1.824) than adventitious blind people (mean estimate of frequency = 1.119). However, for the category "Feel", the word frequency difference between reports by the blind and the sighted is not significant (see Table 18).

See Tables 19, 20 and 21 for the effect of the degree of blindness on perceptual word usage in dreams. Totally blind people used fewer words of "See" (mean estimate of frequency = 0.649) and "Feel" (mean estimate of frequency = 0.810) in dream reports than not totally blind people (mean estimate of frequency of "See" words = 1.515; mean estimate of frequency of "Feel" words = 1.189), and the differences are significant. However, for the category of "Hear", the difference of word frequency is not significant. The direction of the difference of "Feel" words is in contrast to our hypothesis. One possible reason can be that totally blind people just have less perceptual information than not totally blind people because of more restricted actions in life. There is a LIWC category "Perceptual processes", which is the sum of categories "See", "Hear", and "Feel". Hence, we tested the effect of blindness degree and nature on the word usage of "Perceptual processes" as well. As shown in Table 22, totally blind people did use fewer words regarding "Perceptual processes" (mean estimate of frequency = 3.385) than not totally blind people (mean estimate of frequency = 4.098). By contrast, blindness nature has no significant effect on the word usage of "Perceptual processes" (see Table 23). Overall, Hypothesis 6 is well supported by our results too.

Discussion

By comparing the word frequencies of categories about perception, we found that blind people make significantly fewer references to vision but rely more on audition and tactus in dreams than sighted people. What is more, the nature and degree of blindness are found to make a difference in blind people's dream contents. People who are congenitally blind or who are totally blind report fewer words of vision than people who are blind adventitiously or who are partially blind or who can perceive very bright light. Besides, congenitally blind people dreamed

Table 16. The Influence of Blindness Nature on the Usage of Words About “See”

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Intercept	305.635	1	305.635	197.736	.000
Nature	43.426	1	43.426	28.095	.000
Gender	9.044	1	9.044	5.851	.016
Nature * Gender	5.104	1	5.104	3.302	.070
Error	612.085	396	1.546		
Total	985.862	400			

Table 17. The Influence of Blindness Nature on the Usage of Words About “Hear”

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Intercept	769.271	1	769.271	225.126	.000
Nature	44.146	1	44.146	12.919	.000
Gender	.398	1	.398	.116	.733
Nature * Gender	4.201	1	4.201	1.229	.268
Error	1353.160	396	3.417		
Total	2401.372	400			

Table 18. The Influence of Blindness Nature on the Usage of Words About “Feel”

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Intercept	243.975	1	243.975	147.583	.000
Nature	4.558	1	4.558	2.757	.098
Gender	2.592	1	2.592	1.568	.211
Nature * Gender	12.506	1	12.506	7.565	.006
Error	654.641	396	1.653		
Total	935.914	400			

Table 19. The Influence of Blindness Degree on the Usage of Words About “See”

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Intercept	263.884	1	263.884	162.841	.000
Degree	42.286	1	42.286	26.095	.000
Gender	.545	1	.545	.336	.562
Degree * Gender	2.416	1	2.416	1.491	.223
Error	610.930	377	1.621		
Total	967.551	381			

Table 20. The Influence of Blindness Degree on the Usage of Words About “Hear”

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Intercept	517.743	1	517.743	146.856	.000
Degree	5.776	1	5.776	1.638	.201
Gender	.096	1	.096	.027	.869
Degree * Gender	12.002	1	12.002	3.404	.066
Error	1329.117	377	3.526		
Total	2274.962	381			

Table 21. The Influence of Blindness Degree on the Usage of Words About “Feel”

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Intercept	225.021	1	225.021	141.912	.000
Degree	8.104	1	8.104	5.111	.024
Gender	19.486	1	19.486	12.289	.001
Degree * Gender	31.468	1	31.468	19.845	.000
Error	597.787	377	1.586		
Total	874.970	381			

Table 22. The Influence of Blindness Degree on the Usage of Words About "Perceptual Information"

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Intercept	3154.269	1	3154.269	454.425	.000
Degree	28.605	1	28.605	4.121	.043
Gender	19.191	1	19.191	2.765	.097
Degree * Gender	110.893	1	110.893	15.976	.000
Error	2616.844	377	6.941		
Total	7214.601	381			

Table 23. The Influence of Blindness Nature on the Usage of Words About "Perceptual Information"

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Intercept	4132.687	1	4132.687	589.995	.000
Nature	.007	1	.007	.001	.975
Gender	3.875	1	3.875	.553	.457
Nature * Gender	5.444	1	5.444	.777	.379
Error	2773.828	396	7.005		
Total	7575.344	400			

about more auditory references than adventitiously blind people. However, for completely blind people, the word frequency of “Feel” is significantly lower than that for people who are not completely blind. This may be because the degree of blindness is associated with the total words about perceptual processes. Our results demonstrated that totally blind people did report fewer perception words than not completely blind people. Yet, the nature of blindness (congenital vs. adventitious) has nothing to do with the total counts of perceptual words. Our results are greatly consistent with former studies (e.g., Hurovitz, Dunn, Domhoff, & Fiss, 1999) and provide important evidence for the continuity hypothesis of dreams.

Study 4: Automatic Detection of Binary Characteristics of Dreams by SVM

Introduction to Support Vector Machines

Machine learning (ML) techniques can benefit dream studies but have rarely been used in dream research (an example of dream research using ML is Wong, Amini, & De Koninck, 2016). In ML, support-vector machines (SVMs) are models with learning algorithms to analyze data for classification (Ben-Hur, Horn, Siegelmann, & Vapnik, 2001; Cortes & Vapnik, 1995). Basically, when one has a set of training examples, each of which has been labeled with one of a given number of categories, and puts the samples into an SVM training algorithm, then an SVM can predict categorical labels for new testing examples. For an intuitive introduction of the working mechanism of SVM, see Zheng and Schweickert (2021b). See Figure 2 for a simple example of SVM.

As shown in the left part of Figure 2, suppose we have 6 solid-circle dots of Category 1 and 5 dashed-circle dots of Category 2. There are two variables, x1 and x2, that we can use to distinguish the two categories. Then by training an SVM, we can get an optimal hyperplane, as shown in the right part of Figure 2, which has the maximized margin for both Category 1 and Category 2. The dots in black are support vectors. (They are vectors when considered as lists of coordinate values.) Once we have new dots without knowing their category, we can use the trained model to categorize them.

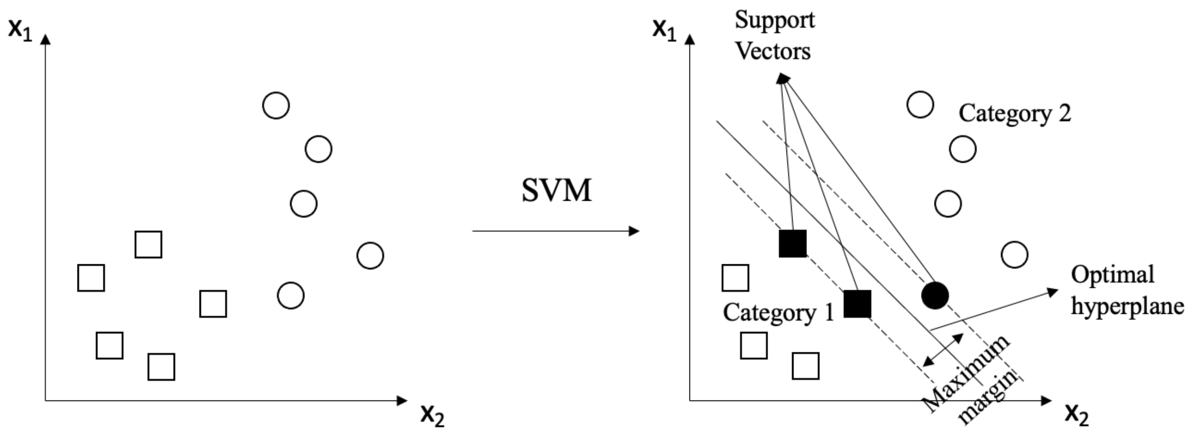

Figure 2. An example of SVM.

The SVM algorithm has been widely applied in biology, language processing, image and graph analysis, and other sciences (e.g., Gaonkar & Davatzikos, 2013). For example, Sacchet, Prasad, Foland-Ross, Thompson, and Gotlib (2015) differentiated major depressive disorder participants from healthy controls using MRI graphs data measured by structural graph metrics together with SVM. Despite the powerful detecting and predicting binary characteristics of SVM, to our knowledge, SVM has rarely been used in dream studies (one example is Zheng & Schweickert, 2021b).

Some studies tried to automatically detect gender from dream reports, using an automatic analysis technique which combines the outputs of LIWC and the transformations of the data (Wong, Amini, & De Koninck, 2016). This group mainly considered the language features of females and males and used simple logistic regression for classification (also see Matwin, Razavi, De Koninck, & Amini, 2010). With a logistic regression model, a real number value can be transformed to a new value in [0, 1]. Then, the new value will be labeled as 1 if it is larger than a critical value, else it will be assigned a label 0. Both SVM and logistic regression have their own merits. Specifically, SVM can build an optimal hyperplane between categories by maximizing the margins. Hence, SVM is less sensitive to outliers than logistic regression models. This sensitivity may allow us to detect binary characteristics of dreamers using dream reports in an improved way. In the current study, we built SVM models to predict the binary characteristics of a dreamer, whether a dreamer is female or male (Analysis 1), and whether a dreamer is blind or sighted (Analysis 2), by word frequencies of all LIWC categories.

Method

For Analysis 1, we used Hall Van de Castle norm dreams of female (*n* = 490) and male (*n* = 491) to train SVM models for detecting the gender of a dreamer. For Analysis 2, we used the norm dreams and reports from blind dreamers (female: *n* = 257; male: *n* = 143) to predict whether a dreamer is blind or sighted. To keep the number of reports by sighted dreamers and blind dreamers the same, we randomly chose 400 norm dreams from the 981 reports of norm dreams for Analysis 2. Our data are frequencies of all LIWC categories in the reports. When training a SVM model, we used 66.7% of the reports and applied to them the 10-fold cross-validation procedure, which is often used in SVM modeling, to tune the model and get best performance parameters. After building a SVM model using the best performance parameters, we used the remaining 33.3% of the reports to test the performance of the SVM model. For more information of overfitting and 10-fold cross-validation, see Zheng and Schweickert (2021b). We used the package "e1071" in RStudio 1.1.383 to build the SVM model and chose nu-classification and radial kernel for the SVM model.

Results

Analysis 1**.** If a dreamer is male and also predicted to be male by SVM using the word frequency data of dream reports, then this is one count of the true positive (TP); conversely, if the dreamer is predicted to be a female by SVM, then this is one count of the false negative (FN). Meanwhile, if a dreamer is female and also predicted to be a female, then it is true negative (TN); otherwise, if the dreamer is predicted to be male, then it is false positive (FP). To assess the performance and prediction ability of an SVM model, four metrics are usually calculated:

Accuracy = (TP + TN) / (TP + FP + FN + TN),

Precision = TP / (TP + FP),

Recall = TP / (TP + FN),

F1 Score = 2 * (Recall * Precision) / (Recall + Precision).

For all these four metrics, a bigger number between zero and one indicates a better performance of a model. Additionally, we used Cohen's kappa to measure the agreement between the predicted machine score and the gender of a dreamer (female is coded as 0 while male is coded as 1). Kappa scores indicate 0 to 0.20 as slight, 0.21 to 0.40 as fair, 0.41 to 0.60 as moderate, 0.61 to 0.80 as substantial, and 0.81 to 1 as almost perfect (McHugh, 2012).

To avoid the possible problem caused by randomly separating reports into training and testing sets, we ran the whole procedure of SVM model building for 1000 times independently. It turned out that the average accuracy of our SVM models for gender detection was 0.852, the average precision was 0.854, the average recall was 0.848, and the average F1 score was 0.851. These indicated that the SVM model was able to precisely predict the gender of a person based on word frequencies of all LIWC categories in dream reports. The average Kappa score of our SVM models was 0.703, which indicated an excellent agreement between the predicted machine score and the gender of a dreamer.

Analysis 2**.** Similar to Analysis 1, we calculated accuracy, precision, recall, and F1 score to assess the performance of the SVM model for blindness prediction. Here, if a dreamer is blind and also predicted to be blind by SVM, then this is one count of the true positive (TP); by contrast, if the dreamer is predicted to be a sighted by SVM, then this is one count of the false negative (FN). Meanwhile, if a dreamer is sighted and also predicted to be sighted by SVM, then it is true negative (TN); otherwise, if the dreamer is predicted to be blind, then it is false positive (FP). We also used Cohen's kappa to measure the agreement between the predicted machine score and the blindness of a dreamer (the sighted is coded as 0 while the blind is coded as 1).

Again, we ran the whole procedure of SVM model building for 1000 times independently to avoid the possible problem caused by randomly separating reports into training and testing sets. It turned out that the average accuracy of our SVM models was 0.869, the average precision was 0.833, the average recall was 0.945, and the average F1 score was 0.882. These indicated that the SVM model was able to precisely predict whether a person was blind or sighted, just based on the word usage in dream reports. The average Kappa score of our SVM models was 0.738, which also demonstrated a substantial agreement between the predicted machine score and blindness of a dreamer.

Discussion

We built SVM models to predict binary characteristics of dreamers based on the word frequencies of LIWC categories in dream reports. Results of Analysis 1 showed a good performance of SVM for predicting the gender of dreamers using their dream reports. The accuracy of the prediction of gender by the automatic analysis of Wong, Amini, and De Koninck (2016) was 74.5% (Cohen's kappa = 0.492). Our SVM models performed better with a correct gender prediction accuracy (predict gender) of 86.9% and the diagnostic reliability (Cohen's kappa) of 0.738. In Analysis 2, the SVM was also able to well detect whether a person is blind or sighted based on dream reports. Overall, the highly precise results of our models show a high potential for SVM models to detect binary characteristics of dreamers based on their dream contents.

Overall Discussion

The continuity hypothesis (Hall & Nordby, 1972) says that dream contents are considerably congruent with people's waking life. To consider the relationship, we used LIWC, a well-established text content analysis tool, to analyze dream reports. We found differences in word usage by men and women in dreams largely consistent with differences previously found in waking life. By analyzing the dream reports of three dreamers covering 40 to 50 years, we found a longitudinal pattern of fewer and fewer negative emotion word usage in dream reports, consistent with previous findings about negative emotion word usage in waking life. Furthermore, we applied automatic content analysis to investigate differences between dreams of the blind and the sighted. We found that blind people included more words about audition and tactus and fewer about vision in dreams than sighted people. Also, the effect of the nature and degree of blindness were found to have significant effects on blind people's dream contents. Last but not least, we made Support Vector Machine models and predicted binary characteristics of dreamers, here gender and blindness, using word frequencies of LIWC categories in dream reports. With these results, our study brings further evidence to support continuity between dreams and waking life. More importantly, the present study shows a high potential of binary characteristics detection of dreamers based on dream contents using machine learning techniques.

The continuity hypothesis of dreaming (Hall & Nordby, 1972) has been widely explored in dream science since 1970s, holding that dreams reflect waking-life experience. A generous amount of literature has contributed to generally support the general formulation of continuity hypothesis (Schredl, 2003; Schredl & Reinhard, 2012). Dream contents have been shown significant relationships in important aspects of waking life, including psychological well-being (Pesant & Zadra, 2006), traumatizing events (Valli, Revonsuo, Pälkäs, & Punamäki, 2006), and personality (Schredl, 2007) and so on. However, critiques and drawbacks have been entailed on continuity hypothesis in dream studies too. For instance, Schredl and Hofmann (2003) stated that small numbers of dreams per participant and measurement methods might cause anincrease of error variance.

Despite the accumulated support for the continuity hypothesis of dreaming (e.g., Domhoff, 1996), limited studies contributed to link between wakefulness and dreaming by the use of word frequencies via LIWC. Our current studies on the consistency between dreams and waking life in terms of gender, aging effect, and blindness difference positively support the continuity hypothesis, which enriches the literature of the continuity hypothesis.

A traditional goal of dream content research is to discover aspects of a dreamer, such as problems in relationships. But before tackling such nebulous aspects, a preliminary step is to verify methods by testing them on conspicuous aspects such as gender, age and sightedness. Study 4 of the current paper demonstrates a high potential of ML techniques for dream research. We used SVM modeling to precisely detect binary characteristics of people based on dream contents. Future studies can use SVM and other ML techniques for more complicated and abstract topics of dream.

Limitations of using LIWC for dream report analysis have been discussed by Bulkeley and Graves (2018) and Zheng and Schweickert (2021a). We note an additional limitation here, the data from only three dreamers, all female, for examining the relationship between aging and dreaming, making results hard to generalize. To summarize, the study enriches the literature of the continuity hypothesis of dream research in terms of gender, age, and blindness by word frequency analysis. We also made a methodological contribution to dream research by well predicting binary characteristics of dreamers based on dreams contents using machine learning techniques.

In the next chapter, we built doubly measured mixed effect models to explore the inconsistencies between wakefulness and dreaming using the word frequencies in dream reports and waking diaries. We also built SVM models to achieve another kind of binary characteristics detection of dreams, that is, whether a report is about dreaming or waking.

CHAPTER 3: DIFFERENTIATING DREAMS FROM WAKEFULNESS BY AUTOMATIC CONTENT ANALYSIS AND SUPPORT VECTOR MACHINES

Introduction

Even though dreams and waking life are consistent in some aspects, they are not always correspondent. Although a few studies tried to explore the differences between waking life texts and dream reports by quantitative analysis (e.g., Bulkeley & Graves, 2018), the discrepancies between dreaming and waking still need and deserve to be discovered. In this study, we aimed to detect the discontinuities between dreams and waking life from the aspects of social contents and cognitive functions by building doubly measured mixed effect models. Furthermore, we built machine learning models to precisely detect binary characteristics of dreams, and here it is whether a text describes waking life or dreams based on the word frequencies of various LIWC categories.

LIWC as a Tool for Dream Studies

Methods and tools have been developed and used to quantitatively measure dream contents, for example, the Hall Van de Castle (HVdC) coding system (Hall & Van de Castle, 1966) and a Word Search Technology based on the search engine on DreamBank.net (Bulkeley, 2009; Bulkeley, 2014; Domhoff & Schneider, 2008a, 2008b). Hand-coding is labor-intensive and may bring in coding bias and be slow in progress. Progress in dream research would be faster, and results would be more statistically consistent if automatic coding methods were feasible.

LIWC is a software for automatically analyzing texts. The most updated version is LIWC 2015 (Pennebaker, Boyd, Jordan, & Blackburn, 2015). Once one puts text into LIWC, it can automatically tabulate the frequencies of word usage for approximately 100 different categories, about emotion, grammar and vocabulary, social processes, and so on. There are also summary categories in LIWC. Of our interest, the category “Tone”, which means emotional tone, is equal to the frequencies of positive emotion words minus the frequencies of negative emotion words and hence indexes the verbal positivity of a text (Cohn, Mehl, & Pennebaker, 2004); the category “Analytic” is composited by eight function word categories in LIWC. The formula of “Analytic” by the LIWC team (Jordan, Sterling, Pennebaker, & Boyd 2019; Pennebaker et al.,

2014) is: Analytical language composite = 30 + article + prep – personal pronouns – impersonal pronouns - auxiliary verbs – conjunctions – adverbs – negation. A high frequency in "Analytic" reflects an emphasis in conveying relationships among concepts, while a low frequency reflects a narrative experiential style. The category "Clout" measures the relative social status among characters in a text (Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2014).

Despite LIWC's usefulness as a powerful automatic content analysis tool, it has not been used in dream studies as often as in some other disciplines (e.g., McNamara, Pae, Teed, Tripodis, & Sebastian, 2016; Wong, Amini, & De Koninckand, 2016). The current study will use LIWC to analyze waking activity diaries and dream reports to figure out the discontinuities between waking and dreaming.

Difference in Social Contents Between Waking Life and Dreams

Social contents are common in waking life, but surprisingly dreams seem to be more social than waking life (e.g., McNamara, McLaren, Smith, Brown, & Stickgold, 2005). Studies have already found a massive presence of social interactions and networks in dreams (e.g., Hall & Van de Castle, 1966; Han, Schweickert, Xi & Viau-Quesnel, 2016; Schweickert, Xi, Viau-Quesnel, & Zheng, 2020). For example, in the coding system developed by Hall and Van de Castle (1966) (one can find more information at https://dreams.ucsc.edu/Coding/), social interactions in dreams are mainly coded into three categories: friendly, aggressive, and sexual. When they coded 1000 dream reports from college students, they found that approximately half of dream reports contain at least one aggressive interaction, and in around 40% dreams there is at least one friendly interaction. Additionally, among various social scenarios, some researchers found that threatening situations seem to appear more frequently in dreams than in waking life (e.g., Valli, Strandholm, Sillanmäki, & Revonsuo, 2008).

Social Simulation Theory (SST) (Revonsuo et al., 2016; Tuominen, Revonsuo & Valli, 2019), which considers dreaming as a simulation for the world of real life, may provide an answer for why dreams are more social than waking life. It claims that dream contents are a kind of specialized simulation of social contents in waking life, which can be beneficial to people in an adaptive sense, e.g., making better preparation for wakefulness. This theory has theoretical backup by previous studies (e.g., Brereton, 2000; McNamara et al., 2005). Another theory, the Threat

Simulation Theory (TST), claims a function of dreams similar to that of SST, but with emphasis on simulating threatening events (Revonsuo, 2000; Valli & Revonsuo, 2009).

Sociality Bias hypothesis, a testable hypothesis based on SST, states that dreams should be biased to overrepresent social content of waking life because of dreams' function of simulating social perceptions and interactions (Revonsuo et al., 2016). This may answer the question mentioned above: why do dreams seem to be more social than waking life? Another testable hypothesis is, derived from TST, dreams contain more negative emotions and fewer positive emotions than waking life. In the current study, we will test the two hypotheses based on word frequencies of dream reports and waking activity diaries. Specifically, we hypothesize that more words about social contents and negative emotions are used in dream reports than in waking diaries, while fewer words about positive emotions are used in dreams than in waking diaries.

Difference in Cognitive Functions Between Waking Life and Dreams

Compared with events in waking life, dream contents are sometimes more bizarre and unpredictable. Such different features as bizarreness of dreaming may demonstrate a lack of unified organizing rules of conscious state in dreams (Churchland, 1988). Dreams are different from wakefulness in terms of some aspects of cognitive functions. For instance, Blagrove (1996, p1103) stated that "Delusional, ad hoc reasoning, attentional and mnemonic deficiencies and the lack of metacognitive insight into the dream state suggest that important aspects of the cognitive, agentive, and autobiographical or narrative self are lacking in many dreams."

Empirical studies have found some patterns of the difference about cognition and consciousness between dreaming and wakefulness. For example, in Study 1 of Kahan and LaBerge (2011), participants consistently reported more reflective awareness of their feelings and behaviors in waking life than in dreams. For another, dreams are found to include fewer perceptions of smell, touch, and taste than occur in waking life (Kahan & LaBerge, 2011; Snyder, 1970; Zadra, Nielsen & Donderi, 1998). In LIWC, there is a category "Cognitive processes" for words about cognitive activities. Subcategories of "Cognitive processes" include "Insight", "Causation", "Discrepancy", "Tentative", "Certainty", and "Differentiation" (Pennebaker, Boyd, Jordan, & Blackburn, 2015). LIWC also has a category "Perceptual processes" calculating the frequency of all perception words. Below this category, there are categories "See", "Hear", and "Feel". In the current study, we will compare the word usage of these categories in dream reports and waking diaries to explore the cognitive difference between dreaming and waking.

Executive function is a broad construct of higher order cognitive functions for performing sophisticated goal-directed tasks, for example, shift of mental sets and inhibition of dominant responses (e.g., Miyake, Friedman, Emerson, Witziki, Howerter, & Wager, 2000; Rabbitt, 1997; Vaughan & Giovanello, 2010). One important executive function is to monitor and update working memory representations (e.g., Miyake et al., 2000). Working memory is defined by Miyake and Shah (1999) as "a basic cognitive mechanism (or a set of mechanisms) that is responsible for keeping track of multiple task-related goals and subgoals, or integrating multiple sources of information." A person with high working memory capacity usually can better deal with new information and relevant old information. Some studies have tried to catalog the language markers associated with executive functioning and working memory. For example, Polsinelli et al. (2020) analyzed the word usage of older adults and found that higher overall executive function levels and better working memory capacities are related with higher frequencies of some LIWC categories, e.g., "Analytic" and "numbers".

Dream studies suggested that people's executive functions and working memory capacities are attenuated during dreaming (e.g., Pace-Schott, 2003; Stickgold, 2005). Prefrontal cortex, especially the dorsolateral portion, is thought to be associated with executive functioning and working memory (e.g., Goldman-Rakic, 1996; Smith & Jonides, 1999). The diminished activation of the dorsolateral prefrontal cortex is recognized as one characteristics of rapid eye movement sleep (e.g., Muzur, Pace-Schott, & Hobson, 2002), the sleep stage from which dreams are most often reported. Since executive functions and working memory capacities are mitigated in dreams, comparing dream reports and waking diaries might reveal similar patterns of word frequencies as those found in Polsinelli et al. (2020). Here, we will use Polsinelli et al. (2020) as a guide, and make an exploratory analysis to see how much the language markers of better executive functioning and working memory capacity found by Polsinelli et al. (2020) can also be found in our study.

Comparing Waking Activities and Dreams for Other Aspects

There are pioneering studies trying to figure out the difference between waking life and dreams by counting word frequencies in reports. For example, Bulkeley and Graves (2018) compared dream reports with six genres of text: Personal blogs, Expressive writings by college students, English novels written between 1660 and 2008, Natural speech recorded during people's daily lives, New York Times articles between January and July of 2014, and Twitter posts (data originally collected by Pennebaker, Boyd, Jordan, & Blackburn, 2015), and found that dream reports have higher frequencies of word usage in the following LIWC categories: focus on the past, first-person singular words, personal pronouns, authenticity, dictionary words, motion, space, and home, compared with other kinds of texts. What is more, dream reports have lower frequencies of words in the following LIWC categories: informal language, focus on the present, assent, positive emotions, clout, second-person references, affective processes, and quotation marks.

However, the comparison of these six genres of texts, such as newspaper articles, with dream reports does not necessarily reflect the difference between waking life and dreams. Here, to better explore the differences, we will analyze the word frequencies in a corpus of dream reports and waking reports collected by Kahan and Sullivan (2012). In this collection, most participants reported a pair of one report of a dream experience and one report of a waking life experience and then reported another pair of reports about one week later[3]. More information about the data will be given later in Method.

Automatic Detection of Binary Characteristics of Dreams by SVM

Machine learning (ML) techniques can well benefit dream studies, even though only a handful applications of ML can be seen in dream research (for an example dream study using machine learning techniques, see Wong, Amini, & De Koninck, 2016). Support-vector machines (SVMs) are models using machine learning algorithms to analyze data for classification (Ben-Hur, Horn, Siegelmann, & Vapnik, 2001; Cortes & Vapnik, 1995). An SVM constructs a (or a set of) high-dimensional plane(s) to robustly separate groups in high-dimensional feature space, based on the number of features. When one has a set of training examples, each of which has been labeled with one of a given number of categories, an SVM can predict categorical labels for new testing examples. In SVM, the distance of a separation hyperplane to the closest dot is called a *margin*. A good SVM model is supposed to have the maximized margin for all categories. The data points that are closest to the separation hyperplane in the high dimensional space are called

[3] Two participants made 8 reports for unknown reasons. For the sake of consistency, we eliminated the 4 extra reports from the two participants in our dataset.

support vectors because each such point can be expressed in the form of a vector. See Figure 1 for a simple example of SVM.

As shown in the left part of Figure 3, suppose we have 6 solid-circle dots of Category 1 and 5 dashed-circle dots of Category 2. A dot, considered as a list of coordinates, is a vector. There are two variables, x1 and x2, that we can use to distinguish the two categories. Then by training an SVM, we can get an optimal hyperplane, as shown in the right part of Figure 1, which has the maximized margin for both Category 1 and Category 2. The dots in grey are support vectors. Once we have new dots without knowing their category, we can use this model to categorize them.

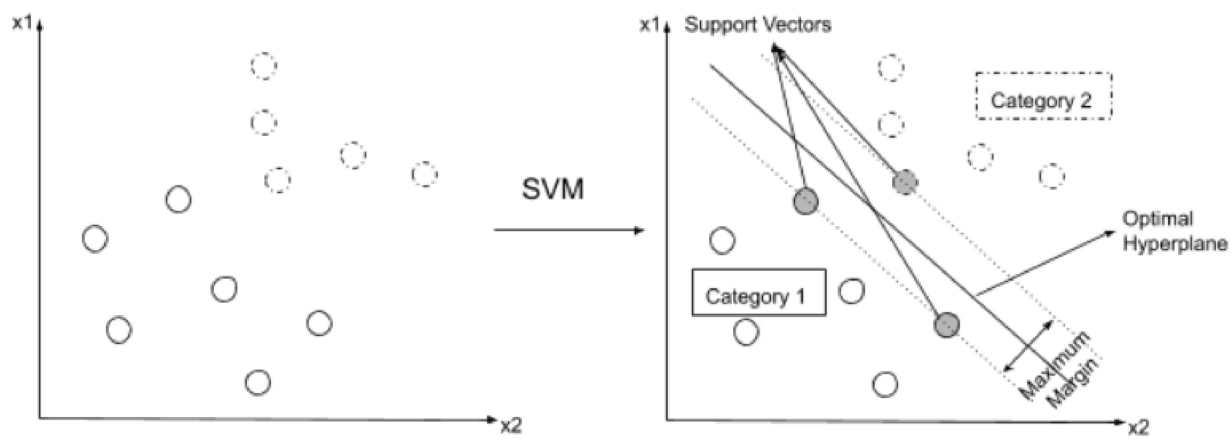

Figure 3. An example of SVM.

The SVM algorithm has been widely applied in biology, language processing, image and graph analysis, and other sciences (e.g., Gaonkar & Davatzikos, 2013). For example, Sacchet, Prasad, Foland-Ross, Thompson, and Gotlib (2015) differentiated major depressive disorder participants from healthy controls using MRI graphs data measured by structural graph metrics together with SVM. Despite the powerful detecting and predicting binary characteristics of SVM, to our knowledge, SVM has not been used in dream studies, not to mention the comparison between waking and dreaming. Here, we will build SVM models to achieve the prediction of binary characteristics of a dream report by word frequencies.

In the current study, we used word frequencies data to build doubly measured mixed effect models. We first looked at whether hypotheses developed based on SST and TST could be supported by comparing word frequencies in dream reports and waking diaries. After that, we compared the frequencies of LIWC categories about cognitive function and perception in dreams and wakefulness. We also compared our findings with those by Polsinelli et al. (2020) to see whether the language markers of better executive function and working memory capacities found by them also exist here. Beyond that, we described and summarized the difference between dream reports and waking diaries for other aspects. Then results in the current study were compared with those in Bulkeley and Graves (2018) to see whether the differences they found between dream reports and waking life materials also exist in our study using people's waking activity reports. Last, based on the frequencies of all categories, we trained SVM models to distinguish dream reports from waking diaries based on word frequencies in texts.

Method

Materials

We used texts collected by Kahan and Sullivan (2012), which consist of 370 waking activity reports and 370 dream reports from 184 participants. The group of participants included 64 males aging from 18 to 31 years old (*M* = 19.97, *SD* = 2.56) and 120 females aging from 18 to 27 years old (*M* = 18.95, *SD* = 1.40). In this collection of reports, a waking activity report records a 15-min waking experiences of a participant in an interval no sooner than two hours after waking and no later than two hours before bedtime. For a dream report, a participant woke up 30-min prior to the usual waking time and recorded the dream experience based on the guidelines for dream reporting provided by researchers. Most participants made two pairs of dream reports and waking activity reports at two time points (week 1 and week 2). For more information about participants and the collection of reports, see Kahan and Sullivan (2012).

Analysis Procedure

In Analysis 1, we built doubly measured mixed effect models using the word frequencies of all LIWC categories as dependent variables to detect the differences between waking activity reports and dream diaries. Here, within-subject effects include two factors: factor 1 is whether the report is about dreaming or waking (D or W), and factor 2 is the time point when reports were collected (1 or 2). This is why the models are *measured*. The effects of factor 1 and 2 are fixed.

Participants were randomly recruited so the between-subject effect was random. Hence, the models considered *mixed effects*. The statistical model can be expressed as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_k + \varepsilon_{ijk},$$

where

> Y_{ijk} : the word frequency of a LIWC category in a text which may be influenced by time i, whether a text is about dreaming or waking j, and subject k ($i = 1, 2$; $j = 1, 2$; $k = 1, 2, 3, \ldots, 184$);
>
> μ: overall mean;
>
> α_i: fixed effect due to time;
>
> β_j: fixed effect due to the factor of whether a report is about dreaming or waking, nested within the factor of time;
>
> δ_k: random effect due to subject;
>
> ε_{ijk}: random errors.

We used SAS 9.4 to build doubly measured mixed effect models. Every LIWC category was used as a dependent variable in a doubly measured mixed effect model, so totally there were 91 models in Analysis 1. We applied the false discovery rate procedure to control Type I error caused by multiple comparisons. Then, we singled out LIWC categories that are associated with social contents (see Table 24) and emotions about threatening events (see Table 25). We tracked the results regarding these categories to see whether the directions of differences were consistent with the Social Simulation Theory (SST) and the Threat Simulation Theory (TST). Based on SST, we hypothesized more words about social contents in dreams than wakefulness. Based on TST, we hypothesized that more words about negative emotions and fewer words about positive emotion were used in dream reports than in waking activity diaries. After that, we looked at results of LIWC categories about cognitive functions and perceptions, and compared results by us and Polsinelli et al. (2020). Then, we summarized the difference between dream reports and waking diaries for other aspects. Finally, as a confirmation check, results in the current study were compared to findings in Bulkeley and Graves (2018).

Table 24. Results Regarding Social Simulation Theory: Differences Between Dream and Wake Reports

LIWC Categories	Mean Difference	Significant After FDR?	Support SST?
Social Processes	3.9892	TRUE	yes
Family	0.5311	TRUE	yes
Female References	0.9379	TRUE	yes
Male References	1.1082	TRUE	yes
Affiliation	1.3179	TRUE	yes
Clout	16.381	TRUE	yes
Friends	0.3404	TRUE	yes
Sexual	0.068	TRUE	yes

Note. For the category "sexual", the original *p* value of the difference is 0.035, which is still significant after FDR for 9 comparisons in Table 24, but is not significant after FDR for 92 comparisons shown in Table 28.

Table 25. Results Regarding Threat Simulation Theory: Differences Between Dream and Wake Reports

LIWC Categories	Mean Difference	Significant After FDR?	Support SST?
Emotional Tone	-8.622	PTRUE	yes
Positive Emotion	-0.712	TRUE	yes
Anger	0.04993	FALSE	no
Sadness	0.03215	FALSE	no
Negative Emotion	0.05465	FALSE	no
Anxiety	0.02909	FALSE	no

In Analysis 2, we built SVM models to predict whether a diary is about dreaming or waking by word frequencies of all LIWC categories. We used 66.7% of the reports to create and train a model. Each word category of LIWC was used as one dimension for classification. An SVM model can be *overfitted* if the noises of the training data are modeled. Then the ability of the model to categorize the testing sample will be impaired. The *k-fold cross-validation* is a common procedure to address this problem. By this method, the entire set is split into *k* sets, of which each set is called a *fold*. A new model is trained using $k-1$ of the folds as training data and the remaining 1-fold as testing data. Then, this method allows us to train a model and test a model *k* times and finally report a model with the best performance. Here, we applied the 10-fold cross-validation procedure, which is often used in SVM modeling, to tune the model and get best performance parameters. After building a SVM model using the best performance parameters, we used the remaining 33.3% of the reports to test the performance of the SVM model. We used the package "e1071" in RStudio 1.1.383 to build the SVM model and chose nu-classification and radial kernel for the SVM models.

Results

Analysis 1

Social contents. Social simulation theory (SST) gained good support from our results, as shown in Table 24. All categories related to social contents appeared significantly more frequently in dream reports than in waking activity diaries. Meanwhile, the results in Table 25 provided some evidence for TST. In results about category "Positive emotion", we see that positive emotion related words appeared less frequently in dreams than in wakefulness. A smaller frequency of "Emotional tone" showed that the difference between the frequencies of positive emotion words and negative emotion words was bigger in dreams than that in waking life. For LIWC categories about negative emotion, "Negative emotion", "Anger", "Sadness", and "Anxiety", the frequencies in dreams were slightly higher than those in waking diaries, even though there was no significant difference.

Cognitive functions. As we introduced earlier, LIWC has a summary category "Cognitive processes" for overall cognitive activity words, and subcategories of this category include

"Insight", "Causation", "Discrepancy", "Tentative", "Certainty", and "Differentiation" (Pennebaker, Boyd, Jordan, & Blackburn, 2015). There seems to be no unitary pattern of the differences about these cognition related categories. We see from Table 26 that for categories "Certainty" and "Differentiation", the frequencies were significantly higher in dreams than those in wakefulness. By contrast, words of the category "Causation" appeared significantly less frequently in dream reports. For the category of "Insight" (mean estimate (dream) = 2.714, mean estimate (wake) = 2.927, $t = -1.41$), "Discrepancy" (mean estimate (dream) = 0.970, mean estimate (wake) = 1.075, $t = -1.08$), and "Tentative" (mean estimate (dream) = 2.070, mean estimate (wake) = 1.998, $t = 0.60$), the differences are not significant. Words of the category "Cognitive processes" appeared slightly more in dreams than in wakefulness, but the difference was not significant either (mean estimate (dream) = 10.312, mean estimate (wake) = 9.880, $t = 1.28$). In short, some cognitive words occur significantly more often in dream than wake reports, some significantly less often, and some not significantly differently, with no obvious pattern.

A pattern of perceptual information displayed in dreams and wakefulness can be found in Table 26. Generally, the frequency of perception words ("Perceptual processes") was significantly lower in dreams than those in wakefulness. In LIWC, there are three categories of perceptions: "See", "Feel", and "Hear". As shown in Table 26, only for the category of "See", the frequency in dreams was significantly higher than that in wakefulness. By contrast, the frequency of "Feel" was significantly lower in dreams. The frequency of "Hear" was lower in dreams too but the difference was not significant (mean estimate (dream) = 0.959, mean estimate (wake) = 1.112, $t = -1.39$). This is consistent with what Kahan and LaBerge (2011) found.

Table 27 shows comparisons between our results and those of Polsinelli, et al. (2020) regarding the language markers of better executive function. They found that reports by people with higher overall execution functions included more words of "Analytical thinking", "Numbers", and "Words > 6 letters" (which means words longer than 6 letters) in daily diaries. They summarize saying people with better overall executive functioning tend to use more analytic, specific, and complex language (Polsinelli et al., 2020). If the executive function is attenuated in dreams, to be consistent with findings by Posinelli's team (2020), the frequencies of "Analytical thinking", "Numbers", and "Words > 6 letters" should be significantly higher in waking activity diaries than those in dreams. And this is confirmed by results here, even though the difference of the usage of

Table 26. Results Regarding Cognitive Functions

	Mean Estimate		Difference of Means				FDR Procedure		
LIWC Category	Dream	Wake	Difference	S.E.	D.F.	*t*	Rank of *p*	Adjusted α	Sig.
Causation	1.174	1.615	-0.440	0.108	184	-4.070***	20	0.011	yes
Certainty	1.140	0.709	0.431	0.076	184	5.680***	21	0.012	yes
Differentiation	2.944	2.164	0.780	0.153	184	5.100***	22	0.012	yes
Feel	0.951	1.907	-0.957	0.115	182	-8.320***	27	0.015	yes
Perceptual Processes	3.652	4.420	-0.769	0.199	187	-3.860***	34	0.019	yes
See	1.614	1.183	0.431	0.127	246	3.400***	38	0.021	yes

Note. "S.E." stands for standard error; "sig" stands for significance.
*$p \leq .05$. **$p \leq .01$. ***$p \leq .001$.

Table 27. Comparisons of Results in the Current Study and Results in Polsineli et al. (2020) Regarding Word Frequencies and Executive Functioning

LIWC Categories	Mean Difference	Significant after FDR?	Polsinelli et al. Sign and Sig.	Do Results Agree?
Positive Emotion	-0.712	TRUE	+, significant	no
Articles	1.342	TRUE	-, significant	no
Emotional Tone	-8.622	TRUE	+, significant	no
Words > 6 letters	-1.937	TRUE	-, significant	yes
Auxiliary Verbs	0.684	TRUE	+, significant	yes
Numbers	-0.249	TRUE	-, significant	yes
Sexual	0.068	FALSE	-, significant	no
Prepositions	-0.333	FALSE	-, significant	no
Swear Words	0.007	FALSE	-, significant	no
Analytical Thinking	-0.818	FALSE	-, significant	no

“Analytical thinking” is not significant. For other categories that were not explicitly related to executive functioning, we found a significantly higher frequency of “Auxiliary verbs” in dream reports, which is consistent with Posinelli et al. (2020) if the executive functioning is attenuated in dreams.

Posinelli et al. (2020) found that word frequencies of 12 LIWC categories are significantly different in writings by subjects with high or low working memory, which they distinguish from high or low executive function. As shown in Table 28, our results of 6 categories of the 12 have significant differences with the same sign as Posinelli et al. (2020). Results here provide some supports for previous findings of Posinelli et al. (2020), although not very strongly.

Other aspects. See Table 29 for comparisons of word usage in dream reports and waking diaries for other aspects. First, for LIWC categories of personal pronouns, for only the category of “1st pers singular” (1st personal singular pronoun), is the frequency in dream reports lower than that in waking diaries, and the difference was significant. For the categories of “1st pers plural” (1st personal plural pronoun), “3rd pers singular” (3rd personal singular pronoun), and “3rd pers plural” (3rd personal plural pronoun), the frequencies in dream reports were significantly higher than those in waking diaries. For the category of “2nd person” (2nd personal pronoun), the frequency in dream reports was higher than that in waking diaries, but the difference was not significant (mean estimate (dream) = 0.098, mean estimate (wake) = 0.050, t = 1.80).

Second, there is a difference between the focus of time in wakefulness and dreams. For the categories of “Present focus” and “Future focus”, the frequencies in dream reports were significantly lower than those in waking diaries, while the frequency of “Past focus” was significantly higher in dreams than that in waking activity reports. Dream reports are about events that at the time of writing would be in the past, and the same is true for wake reports. Hence, the difference in tense is not due to difference in temporal viewpoint but results from the different contents. Dreams would contain more elements about the past than the present or the future, if dreams were a kind of reorganization of knowledge and assimilation of old materials in memory. It seems results here could provide some evidence for this.

Table 28. Comparisons of Results in the Current Study and Results in Polsinelli, et al. (2020) Regarding Word Frequencies and Working Memory

LIWC Categories	Mean Difference	Significant after FDR?	Polsinelli et al. Sign and Sig.	Do Results Agree?
Present Focus	-1.662	TRUE	+, significant	no
Words > 6 letters	-1.937	TRUE	-, significant	yes
Health	-0.305	TRUE	-, significant	yes
Time	-1.971	TRUE	-, significant	yes
Dictionary Words	-0.902	TRUE	+, significant	no
Auxiliary Verbs	0.684	TRUE	+, significant	yes
Home	-0.337	TRUE	-, significant	yes
Numbers	-0.249	TRUE	-, significant	yes
Sexual	0.068	FALSE	-, significant	no
Common Verbs	-0.571	FALSE	+, significant	no
Personal Pronouns	0.327	FALSE	+, significant	no
Analytical Thinking	-0.818	FALSE	-, significant	no

Table 29. Comparisons of Word Frequencies Between Dream Reports and Waking Activities for Aspects not in Previous Tables

	Mean Estimate		Difference of Means				FDR Procedure		
LIWC Category	Dream	Wake	Difference	S.E.	D.F.	*t*	Rank of *p*	Adjusted α	Sig.
1st Pers Singular	10.195	12.279	-2.084	0.308	180	-6.770***	10	0.005	yes
1st Pers Plural	1.300	0.604	0.697	0.123	187	5.660***	11	0.006	yes
3rd Pers Singular	2.313	0.994	1.320	0.161	187	8.180***	12	0.007	yes
3rd Pers Plural	0.686	0.332	0.354	0.064	188	5.480***	13	0.007	yes
Present Focus	4.821	6.483	-1.662	0.325	182	-5.120***	1	0.001	yes
Health	0.341	0.646	-0.305	0.072	252	-4.230***	6	0.003	yes
Ingestion	0.646	1.408	-0.762	0.141	187	-5.410***	25	0.014	yes
Biological Processes	1.843	3.016	-1.173	0.203	265	-5.770***	28	0.015	yes
Past Focus	10.712	9.153	1.559	0.301	182	5.180***	30	0.016	yes
Future Focus	1.101	1.351	-0.250	0.091	187	-2.740**	48	0.026	yes

Note. Sig means significance with False Discovery Rate.
*p < .05. **p < .01. ***p < .001.

Third, results showed that less biological information was presented in dreams. In LIWC, words about biological information were counted for the category "Biological processes", which consisted of four subcategories - "Health", "Ingestion", "Body", and "Sexual" (Pennebaker, Boyd, Jordan, & Blackburn, 2015). As shown in Table 29, dreams contained significantly fewer words about "Health", "Biological processes", and "Ingestion". The frequency of "Sexual" was slightly higher in dreams than those in waking diaries but the difference was not significant after the false discovery rate procedure (mean estimate (dream) = 0.082, mean estimate (wake) = 0.014, $t = 2.12$).

Last, one can find the comparison of our results with findings of Bulkeley and Graves (2018) about word frequencies from all LIWC categories in dream reports and waking writings such as novels and newspaper articles in Table 30. For only 5 categories, which are "Present focus", "Past focus", "Space", "Motion", and "Auxiliary verbs", the differences of word frequencies between Dream and Wake found in the current study have the same sign as those in Bulkeley and Graves (2018). Both studies found that dream reports contain significantly fewer words about focusing on the present and significantly more words about focusing on the past than waking life materials. Meanwhile, more words about motion and space were found to be in dreams by both of us. The differences in sign of some results may reveal the difference between waking life materials about personal activities and other genres of texts such as newspaper articles and novels.

Analysis 2

We built an SVM model to achieve the automatic detection of a binary characteristic, which is the nature of a report (dreaming or waking) in the current study, based on the word frequencies. If a report is about wakefulness and also predicted to be a waking report by SVM, then this is one count of the true positive (TP); by contract, if it is predicted to be a dream report by SVM, then this is one count of the false negative (FN). Meanwhile, if a report is about dreaming and also predicted to be a dream report, then it is true negative (TN); otherwise, if it is predicted to be a waking diary, then it is false positive (FP). To assess the performance and prediction ability of an SVM model, four metrics are usually calculated:

Accuracy = (TP + TN) / (TP + FP + FN + TN),

Precision = TP / (TP + FP),

Recall = TP / (TP + FN),

F1 Score = 2 * (Recall * Precision) / (Recall + Precision).

For all these four metrics, a bigger number between zero and one indicates a better performance of a model. Additionally, we used Cohen's kappa to measure the agreement between the predicted machine score and the nature of a report (wakefulness or dreaming). Kappa scores indicate 0 to 0.20 as slight, 0.21 to 0.40 as fair, 0.41 to 0.60 as moderate, 0.61 to 0.80 as substantial, and 0.81 to 1 as almost perfect (McHugh, 2012).

To avoid the possible problem caused by randomly separating reports into training and testing sets, we ran the whole procedure of SVM model building for 1000 times independently. It turned out that the average accuracy of our SVM models was 0.884, the average precision was 0.894, the average recall was 0.878, and the average F1 score was 0.886. These indicated that the SVM model was able to precisely predict whether a text was about dreaming or wakefulness. The average Kappa score of our SVM model is 0.768, which indicates an excellent agreement between the predicted machine score and the nature of a report. Overall, the highly precise results of our SVM model revealed a high potential for detecting binary characteristics of dream reports by SVM.

Discussion

The first aim of the current study is to detect differences between wakefulness and dreaming by the automatic content analysis technique of LIWC. By analyzing word frequencies, we found several differences. We found that word frequencies of social content related categories in dreams were significantly higher than those in waking diaries, support for Social Simulation Theory (Revonsuo et al., 2016; Tuominen, Revonsuo & Valli, 2019). We also found that frequencies of positive emotion categories are significantly higher in wakefulness than those in dreams, while frequencies of negative emotion categories are lower in wakefulness than those in dreams, even though not significantly. These results partly support Threat Simulation Theory (Revonsuo, 2000; Valli & Revonsuo, 2009).

Table 30. Comparisons of Results in the Current Study and Results in Bulkeley & Graves (2018)

LIWC Categories	Mean Difference	Significant after FDR?	Bulkeley & Graves sign	Do Results Agree?
Present Focus	-1.662	TRUE	-	yes
Positive Emotion	-0.712	TRUE	+	no
Clout	16.381	TRUE	-	no
1st Pers Singular	-2.084	TRUE	+	no
Past Focus	1.559	TRUE	+	yes
Space	1.519	TRUE	+	yes
Authentic	-5.762	TRUE	+	no
Motion	0.658	TRUE	+	yes
Affective Processes	-0.703	TRUE	-	no
Dictionary Words	-0.902	TRUE	+	no
Question Marks	0.068	TRUE	-	no
Auxiliary Verbs	0.684	TRUE	+	yes
Home	-0.337	TRUE	+	no
2nd person	0.048	FALSE	-	no
Assent	-0.009	FALSE	-	no
Informal Language	-0.001	FALSE	-	no

Our results show good compatibility with earlier results of Posinelli et al. (2020) about word frequencies and executive functioning and working memory. Our findings provide some evidence for that executive functioning and working memory capacity are attenuated in dreams. Additionally, to some extent, our results are consistent with Bulkeley and Graves (2018) for comparing dream reports and waking life writings such as newspaper articles and novels in the aspects of time focus and motion/space.

Some higher-order cognitive processes, such as logical thinking and reflective awareness, seem to be deficient in dreams (e.g., James, 1890; Kahan & LaBerge, 2011). Some researchers also argued that self-consciousness is lacking in dreams (e.g., Windt & Metzinger, 2007). Here, we found that only for the category of 1st personal singular pronoun, the word frequency in dream reports was significantly lower than that in waking diaries. For all the other categories of personal pronouns, words frequencies in dream reports were higher than those in waking diaries. This pattern of personal pronouns usage may indicate a lack of self-consciousness in dreams, which deserves further investigation.

Materials, such as surveys, reports, and diaries, are increasingly available for dream researchers. While such materials contain useful information, it is difficult and cumbersome for researchers to extract in traditional hand-coding ways. It now is possible to do automatic quantitative analysis of dream contents with new techniques, such as LIWC. Progress can be better because automatic techniques are likely to be faster and lower in biases caused by human coding.

Further, we applied the machine learning algorithm of Support Vector Machines (SVM) to dream studies to detect binary characteristics of dream reports. Here, we built an SVM model that can achieve a good detection of whether a text is about dreaming or waking. In Zheng and Schweickert (2021b), Chapter 2 here, we applied SVM to dream reports to detect the gender or blindness of dreamers based on word frequencies in dream reports and achieved good prediction accuracy as well. Good detection of gender and blindness suggests it will be possible to detect more subtle binary characteristics through dreams. Dream contents, especially of nightmares, are different between ordinary people and people with mental illness. For example, Levin and Nielsen (2007) found that nightmares are related to anxiety and depressive symptoms (also see Nadorff, Porter, Rhoades, Kunik, Greisinger, & Stanley, 2014). Achieving binary characteristics detection of a dreamer based on dream reports has a potential to contribute to the diagnostics of mental disorders.

There are limitations in the current study. Limitations of using LIWC for dream studies have previously been discussed by Bulkeley and Graves (2018). One example is that dream reports are not often free of typographical errors, associational comments, and idiosyncratic language, which may influence the performance of LIWC. Another limitation is the unknown validity of post-awakening dream reports. Dream reports used here were subjective and collected by subjects themselves. Even with honest subjects, there are still problems of rapid deterioration of dream content immediately after waking up, known as dream amnesia (e.g., Windt & Noreika, 2010). As argued by Hobson, Pace-Schott, and Stickgold (2000), both mnemonic and attentional functioning can be lower in dreams than those during wakefulness. All these may reduce the validity of data from dream reports.

Overall, our studies in the three chapters contribute to understanding the consistencies and inconsistencies between dreams and waking life. In Chapter 1, we discovered a high potential of using LIWC for quantitatively analyze dream contents by canonical correlation analysis, which ensured the suitability of a faster and more accurate tool for coding dreams. Chapter 2 extended the literature of continuity hypothesis of dream research in terms of gender difference, aging effect, the sighted-blind comparison. Besides, we introduced such machine learning techniques as support vector machines to dream studies and precisely detected the binary characteristics of people based on the word frequencies in dream reports. Last but not least, in Chapter 3, the automatic dream content analysis was extended to another use, comparing dream and wake reports for the aspects of social contents and cognitive functions, and results of Analysis 1 fill up the gap of scientific studies in the discontinuity between waking and dreaming. What is more, support vector machines were applied for accurately discriminating dream reports from waking diaries, which is another kind of binary characteristics detection.

REFERENCES

Amini, R., Sabourin, C., & Koninck, J. D. (2011). Word associations contribute to machine learning in automatic scoring of degree of emotional tones in dream reports. *Consciousness and Cognition*, *20*(4), 1570–1576. https://doi.org/10.1016/j.concog.2011.08.003

Ayodeji, I. O., & Obilade, T. O. (2018). Development of a canonical correlation model involving non linearity and asymmetric variables. *Communications in Statistics-Theory and Methods*, *47*(2), 399-414. https://doi.org/10.1080/03610926.2017.1303731

Ben-Hur, A., Horn, D., Siegelmann, H. T., & Vapnik, V. (2001). Support vector clustering. *Journal of Machine Learning Research*, *2*(Dec), 125-137.

Bernstein, D. M., & Roberts, B. (1995). Assessing dreams through self-report questionnaires: Relations with past research and personality. *Dreaming*, *5*(1), 13-27.

Blagrove, M. (1996). Problems with the cognitive psychological modeling of dreaming. *The Journal of Mind and Behavior*, *17*(2), 99-134.

Breger, L., Hunter, I., & Lane, R. W. (1971). *The effect of stress on dreams.* Psychological Issues, 7(3), Monograph 27.

Brereton, D. P. (2000). Dreaming, adaptation, and consciousness. The social mapping hypothesis. *Ethos*, *28*(3), 379-409.

Bruce J. Ellis M. A. & Symons, D. (1990). Sex differences in sexual fantasy: An evolutionary psychological approach. *The Journal of Sex Research*, *27*(4), 527-555. https://doi.org/10.1080/00224499009551579

Bulkeley, K. (2009). Seeking patterns in dream content: A systematic approach to word searches. *Consciousness and Cognition*, *18*(4), 905-916. https://doi.org/10.1016/j.concog.2009.08.005

Bulkeley, K. (2014). Digital dream analysis: A revised method. *Consciousness and Cognition, 29*, 159-170. https://doi.org/10.1016/j.concog.2014.08.015

Bulkeley, K. (2015). The digital revolution in dream research.In M. Kramer & M. Glucksman (Eds.), *Dream research: Contributions to clinical practice* (pp. 95-105). New York, NY: Routledge/Taylor & Francis Group.

Bulkeley, K., & Domhoff, G. W. (2010). Detecting meaning in dream reports: An extension of a word search approach. *Dreaming*, *20*(2), 77-95. https://doi.org/10.1037/a0019773

Bulkeley, K., & Graves, M. (2018). Using the LIWC program to study dreams. *Dreaming*, *28*(1), 43-58. http://dx.doi.org/10.1037/drm0000071

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, *125*(3), 367-383. https://doi.org/10.1037/0033-2909.125.3.367

Carstensen, L. (1995). Evidence for a life-span theory of socioemotional selectivity. *Current Directions in Psychological Science, 4*(5), 151-156.

Carstensen, L. L., Pasupathi, M., Mayr, U., & Nesselroade, J. R. (2000). Emotional experience in everyday life across the adult life span. *Journal of Personality and Social Psychology*, *79*(4), 644-655. https://doi.org/10.1037/0022-3514.79.4.644

Charles, S. T., Reynolds, C. A., & Gatz, M. (2001). Age-related differences and change in positive and negative affect over 23 years. *Journal of Personality and Social Psychology, 80*(1), 136-151. https://doi.org/10.1037/0022-3514.80.1.136

Charness, G., & Gneezy, U. (2012). Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization*, *83*(1), 50-58. https://doi.org/10.1016/j.jebo.2011.06.007

Churchland, P. S. (1988). Reduction and the neurobiological basis of consciousness. In A. J. Marcel & E. Bisiach (Eds.), *Consciousness in contemporary science* (pp. 273-304). Oxford England: Clarendon Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Oxfordshire, England: Routledge.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science*, *15*(10), 687-693. https://doi.org/10.1111/j.0956-7976.2004.00741.x

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297.

Dale, A., Lafrenière, A., & Koninck, J. D. (2017). Dream content of Canadian males from adolescence to old age: An exploration of ontogenetic patterns. *Consciousness and Cognition, 49*, 145-156. https://doi.org/10.1016/j.concog.2017.01.008

Dale, A., Lortie-Lussier, M., & Koninck, J. D. (2015). Ontogenetic patterns in the dreams of women across the lifespan. *Consciousness and Cognition*, *37*, 214-224. https://doi.org/10.1016/j.concog.2015.09.008

Dattalo, P. (2014). *A demonstration of canonical correlation analysis with orthogonal rotation to facilitate interpretation*. Unpublished manuscript, School of Social Work, Virginia Commonwealth University, Richmond, Virginia.

Domhoff, G. W. (1996). *Finding meaning in dreams: A quantitative approach*. New York, NY: Plenum.

Domhoff, G. W. (2003). *The scientific study of dreams: Neural networks, cognitive development, and content analysis*. Washington, DC: American Psychological Association.

Domhoff, G. W. (2017). Now an invasion by a Freudian concept-snatcher: Reply to Erdelyi. *Dreaming, 27*(4), 345-350. https://doi.org/10.1037/drm0000068

Domhoff, G. W., & Schneider, A. (2008a). Studying dream content using the archive and search engine on DreamBank.net. *Consciousness and Cognition, 17*(4), 1238-1247.

Domhoff, G. W., & Schneider, A. (2008b). Similarities and differences in dream content at the cross-cultural, gender, and individual levels. *Consciousness and Cognition, 17*(4), 1257-1265. https://doi.org/10.1016/j.concog.2008.08.005

Domhoff, G. W., & Schneider, A. (2018). Are dreams social simulations? Or are they enactments of conceptions and personal concerns? An empirical and theoretical comparison of two dream theories. *Dreaming, 28,* 1-23. doi:10.1037/drm0000080

Ellman, S. J., & Weinstein, L. N. (1991). REM sleep and dream formation: A theoretical integration. In S. J. Ellman & J. S. Antrobus (Eds.), *Wiley series on personality processes. The mind in sleep: Psychology and psychophysiology* (pp. 466-488). Hoboken, NJ: John Wiley & Sons.

Erdelyi, M. H. (2017). The continuity hypothesis. *Dreaming, 27*(4), 334-344.

Fogli, A., Aiello, L. M. , & Quercia, D. (2020). Our dreams, our selves: automatic analysis of dream reports. *Royal Society Open Science, 7*(8), 192080. https://doi.org/10.1098/rsos.192080

Gaonkar, B., & Davatzikos, C. (2013). Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. *Neuroimage, 78,* 270-283. https://doi.org/10.1016/j.neuroimage.2013.03.066

Hall, C. S. (1953). *The meaning of dreams.* New York, NY: Harper.

Hall, C. S. (1984). "A ubiquitous sex difference in dreams" revisited. *Journal of Personality and Social Psychology, 46*, 1109-1117.

Hall, C. S., & Domhoff, B. J. (1963). A ubiquitous sex difference in dreams. *Journal of Abnormal and Social Psychology, 66*, 278-280.

Hall, C. S., & Nordby, V. J. (1972). *The individual and his dreams*. New York, NY: Signet Book.

Hall, C. S., & Van De Castle, R. L. (1966). *The content analysis of dreams.* New York, NY: Appleton-Century-Crofts.

Han, H. J. (2014). *Structural and longitudinal analysis of cognitive social networks in dreams.* Unpublished doctoral dissertation, Purdue University, West Lafayette, IN.

Han, H. J., Schweickert, R., Xi, Z., & Viau-Quesnel, C. (2016). The cognitive social network in dreams: Transitivity, assortativity, and giant component proportion are monotonic. *Cognitive Science, 40*(3), 671-696. https://doi.org/10.1111/cogs.12244

Hartmann, E. (2007). The nature and functions of dreaming. In *The new science of dreaming: Content, recall and personality correlates* (pp. 171-192). Oxford, England: Oxford University Press.

Hartmann, E., Elkin, R., & Garg, M. (1991). Personality and dreaming: The dreams of people with very thick or very thin boundaries. *Dreaming, 1*(4), 311-324. https://doi.org/10.1037/h0094342

Hawkins, II, R. C., Boyd, R. (2017). Such stuff as dreams are made on: Dream language, LIWC norms and personality correlates. *Dreaming, 27*(2), 102-121.

Henson, R. K. (2000). Demystifying parametric analyses: Illustrating canonical correlation analysis as the multivariate general linear model. *Multiple Linear Regression Viewpoints*, *26*(1), 11-19.

Hobson, J. A., Pace-Schott, E. F., & Stickgold, R. (2000). Dreaming and the brain: Toward a cognitive neuroscience of conscious states. *Behavioral and Brain Sciences, 23*(6), 793-842. https://doi.org/10.1017/S0140525X00003976

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika, 28*(3/4), 321-77.

Hurovitz, C. S., Dunn, S., Domhoff, G. W., & Fiss, H. (1999). The dreams of blind men and women: A replication and extension of previous findings. *Dreaming*, *9*(2-3), 183-193. https://doi.org/10.1023/A:1021397817164

James, W. (1890). *The principles of psychology* (Vols. 1 & 2). New York, NY: Henry Holt and Company.

Johnson, R. A., & Wichern, D. W. (1998). *Applied multivariate analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Jordan, K. N., Sterling, J., Pennebaker, J. W., & Boyd, R. L. (2019). Examining long-term trends in politics and culture through language of political leaders and cultural institutions. *Proceedings of the National Academy of Sciences*, *116*(9), 3476-3481. https://doi.org/10.1073/pnas.1811987116

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2014). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, *33*(2), 125-143. https://doi.org/10.1177/0261927x13502654

Kahan, T. L., & LaBerge, S. P. (2011). Dreaming and waking: similarities and differences revisited. *Consciousness and Cognition,* *20*(3), 494-514. https://doi.org/10.1016/j.concog.2010.09.002

Kahan, T. L., & Sullivan, K. T. (2012). Assessing metacognitive skills in waking and sleep: A psychometric analysis of the Metacognitive, Affective, Cognitive Experience (MACE) questionnaire. *Consciousness and Cognition,* *21*(1), 340-352. https://doi.org/10.1016/j.concog.2011.11.005

Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American Journal of Psychology*, *120*(2), 263-286.

Kennedy, J. M. (1993). *Drawing & the blind: Pictures to touch*. London, England: Yale University Press.

Kennedy, J. (1997). How the blind draw. *Scientific American, 276*(1), 76-81. Retrieved January 5, 2021, from http://www.jstor.org/stable/24993567

Kirk, R. E. (1999). *Statistics: An introduction* (4th ed.). Fort Worth, TX: Harcourt Brace.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance-testing system. *Psychological Bulletin*, *85*(2), 410-416. https://doi.org/10.1037/0033-2909.85.2.410

Knoth, R., Boyd, K., & Singer, B. (1988). Empirical tests of sexual selection theory: Predictions of sex differences in onset, intensity, and time course of sexual arousal. *The Journal of Sex Research*, *24*(1), 73-89. https://doi.org/10.1080/00224498809551399

Koivisto, M., Salminen-Vaparanta, N., Grassini, S., & Revonsuo, A. (2016). Subjective visual awareness emerges prior to P3. *European Journal of Neuroscience, 43*(12), 1601-1611. https://doi.org/10.1111/ejn.13264

Levin, R., & Nielsen, T. A. (2007). Disturbed dreaming, posttraumatic stress disorder, and affect distress: A review and neurocognitive model. *Psychological Bulletin, 133*(3), 482-528. https://doi.org/10.1037/0033-2909.133.3.482

Mao, M., Stillwell, D., Kosinski, M., & Good, D. (2017). Testing ageing theory among later middle-aged and older users using social media. *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 247-250. https://doi.org/10.1145/3022198.3026352

Matwin, S., De Koninck, J., Razavi, A. H., & Amini, R. R. (2010). Classification of dreams using machine learning. *Proceedings of the 2010 Conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, 169-174.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica, 22*(3), 276-282. https://doi.org/10.11613/bm.2012.031

McNamara, P., McLaren, D., Smith, D., Brown, A., & Stickgold, R. (2005). A “Jekyll and Hyde” within: Aggressive versus friendly interactions in REM and non-REM dreams. *Psychological Science, 16*(2), 130-136. https://doi.org/10.1111/j.0956-7976.2005.00793.x

McNamara, P., Pae, V., Teed, B., Tripodis, Y., & Sebastian, A. (2016). Longitudinal studies of gender differences in cognitional process in dream content. *Journal of Dream Research*, *9*(1). https://doi.org/10.11588/ijodr.2016.1.26552

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49-100. https://doi.org/10.1006/cogp.1999.0734

Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge, United Kingdom: Cambridge University Press.

Mulac, A., Bradac, J. J., & Gibbons, P. (2001). Empirical support for the gender-as-culture hypothesis: An intercultural analysis of male/female language differences. *Human Communication Research*, *27*(1), 121-152. https://doi.org/10.1111/j.1468-2958.2001.tb00778.x

Muzur, A., Pace-Schott, E. F., & Hobson, J. A. (2002). The prefrontal cortex in sleep. *Trends in Cognitive Sciences*, *6*(11), 475-481. https://doi.org/10.1016/s1364-6613(02)01992-7

Nadorff, M. R., Porter, B., Rhoades, H. M., Greisinger, A. J., Kunik, M. E., & Stanley, M. A. (2014). Bad dream frequency in older adults with generalized anxiety disorder: prevalence, correlates, and effect of cognitive behavioral treatment for anxiety. *Behavioral Sleep Medicine*, *12*(1), 28-40. https://doi.org/10.1080/15402002.2012.755125

Newen, A., & Vogeley, K. (2003). Self-representation: Searching for a neural signature of self-consciousness. *Consciousness and Cognition*, *12*(4), 529-543. https://doi.org/10.1016/S1053-8100(03)00080-1

Pace-Schott, E. F. (2003). Postscript: Recent findings on the neurobiology of sleep and dreaming. In E. F. Pace-Schott, M. Solms, M. Blagrove, & S. Harnad (Eds.), *Sleep and dreaming: Scientific advances and reconsiderations* (pp. 335-350). Cambridge, United Kingdom: Cambridge University Press. https://doi.org/10.1017/CBO9780511615511.012

Park, A., & Conway, M. (2017). Longitudinal changes in psychological states in online health community members: Understanding the long-term effects of participating in an online depression community. *Journal of Medical Internet Research*, *19*(3), e71. https://doi.org/10.2196/jmir.6826

Paul, F., & Schredl, M. (2012). Male-female ratio in waking-life contacts and dream characters. *International Journal of Dream Research*, *5*, 119-124.

Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist*, *211*(2828), 42-45. https://doi.org/10.1016/S0262-4079(11)62167-2

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015.* Austin, TX: University of Texas at Austin.

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PloS One*, *9*(12), e115844. https://doi.org/10.1371/journal.pone.0115844

Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology, 85*(2), 291-301. https://doi.org/10.1037/0022-3514.85.2.291

Polsinelli, A. J., Moseley, S. A., Grilli, M. D., Glisky, E. L., & Mehl, M. R. (2020). Natural, everyday language use provides a window into the integrity of older adults' executive functioning. *The Journals of Gerontology: Series B.* https://doi.org/10.1093/geronb/gbaa055

Proksch, K., & Schredl, M. (1999). Impact of parental divorce on children's dreams. *Journal of Divorce & Remarriage, 30*(1-2), 71-82. https://doi.org/10.1300/J087v30n01_05

Rabbitt, P. (Ed.). (1997). Introduction: Methodologies and models in the study of executive function. In *Methodology of frontal and executive function* (pp. 1-38). Oxfordshire, England: Routledge.

Revonsuo, A. (2000). The reinterpretation of dreams: An evolutionary hypothesis of the function of dreaming. *Behavioral and Brain Sciences, 23*(6), 877-901. https://doi.org/10.1017/S0140525X00004015

Revonsuo, A., Tuominen, J., & Valli, K. (2015). *The avatars in the machine: Dreaming as a simulation of social reality.* Open MIND. Frankfurt am Main: MIND Group. http://dx.doi.org/10.15502/9783958570375

Sacchet, M. D., Prasad, G., Foland-Ross, L. C., Thompson, P. M., & Gotlib, I. H. (2015). Support vector machine classification of major depressive disorder using diffusion-weighted neuroimaging and graph theory. *Frontiers in Psychiatry, 6*, 21. https://doi.org/10.3389/fpsyt.2015.00021

Schneider, A., & Domhoff, G. W. (2019). DreamBank. Retrieved from www.dreambank.net

Schredl, M. (2001). Dreams of singles: effects of waking-life social contacts on dream content. *Personality and Individual Differences, 31*, 269-275. doi:10.1016/S0191-8869(00)00135-5

Schredl, M., & Hofmann, F. (2003). Continuity between waking activities and dream activities. *Consciousness and Cognition, 12*(2), 298-308. https://doi.org/10.1016/s1053-8100(02)00072-7

Schredl, M., Schäfer, G., Hofmann, F., & Jacob, S. (1999). Dream content and personality: Thick vs. thin boundaries. *Dreaming, 9*(4), 257-263. https://doi.org/10.1023/A:1021336103535

Schweickert, R., Xi, Z., Viau-Quesnel, C., & Zheng, X. (2020). Power law distribution of frequencies of characters in dreams explained by random walk on semantic network. *International Journal of Dream Research*, *13*(2), 192-201. https://doi.org/10.11588/ijodr.2020.2.71370

Sharma, N., Chakrabarti, S., & Grover, S. (2016). Gender differences in caregiving among family - caregivers of people with mental illnesses. *World Journal of Psychiatry*, *6*(1), 7-17. https://doi.org/10.5498/wjp.v6.i1.7

Sherry, A., & Henson, R. K. (2005). Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer. *Journal of Personality Assessment*, *84*(1), 37-48. https://doi.org/10.1207/s15327752jpa8401_09

Snyder, F. (1970) The phenomenology of REM dreaming. In H. Madow & L. H. Snow (Eds.), *The psychodynamic implications of physiological studies on dreams* (pp. 124-151). Springfield, IL: Charles C. Thomas.

Stickgold, R. (2005). Sleep-dependent memory consolidation. *Nature, 437*(7063), 1272-1278. https://doi.org/10.1038/nature04286

Strauch, I., & Meier, B. (1996). *In search of dreams: Results of experimental dream research*. Albany, NY: SUNY Press.

Tabachnick, B. G., & Fidell, L. S. (2007). *Experimental designs using ANOVA*. Belmont, CA: Thomson/Brooks/Cole.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). London, England: Pearson.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24-54. https://doi.org/10.1177/0261927X09351676

Thompson, G. (Ed.). (1991). *Markets, hierarchies and networks: the coordination of social life*. Thousand Oaks, CA: Sage.

Tov, W., Ng, K. L., Lin, H., & Qiu, L. (2013). Detecting well-being via computerized content analysis of brief diary entries. *Psychological Assessment*, *25*(4), 1069-1078. https://doi.org/10.1037/a0033007

Tuominen, J., Stenberg, T., Revonsuo, A., & Valli, K. (2019). Social contents in dreams: An empirical test of the social simulation theory. *Consciousness and Cognition*, *69*, 133-145. https://doi.org/10.1016/j.concog.2019.01.017

Valli, K., & Revonsuo, A. (2009). The threat simulation theory in light of recent empirical evidence: A review. *The American Journal of Psychology, 122*(1), 17-38.

Valli, K., Strandholm, T., Sillanmäki, L., & Revonsuo, A. (2008). Dreams are more negative than real life: Implications for the function of dreaming. *Cognition and Emotion, 22*(5), 833-861.

Vaughn, L. A. (2018). Contents of hopes and duties: A linguistic analysis. *Frontiers in Psychology*, *9*, 757. https://doi.org/10.3389/fpsyg.2018.00757

Vaughan, L., & Giovanello, K. (2010). Executive function in daily life: Age-related influences of executive processes on instrumental activities of daily living. *Psychology and Aging, 25*(2), 343-355. https://doi.org/10.1037/a0017729

Voss, U., Holzmann, R., Tuin, I., & Hobson, A. J. (2009). Lucid dreaming: a state of consciousness with features of both waking and non-lucid dreaming. *Sleep, 32*(9), 1191-1200.

Windt, J. M., & Metzinger, T. (2007). The philosophy of dreaming and self-consciousness: What happens to the experiential subject during the dream state? In D. Barrett & P. McNamara (Eds.), *Praeger perspectives. The new science of dreaming: Vol. 3. Cultural and theoretical perspectives* (pp. 193-247). Westport, CT: Praeger Publishers/Greenwood Publishing Group.

Windt, J. M., & Noreika, V. (2011). How to integrate dreaming into a general theory of consciousness—a critical review of existing positions and suggestions for future research. *Consciousness and Cognition, 20*(4), 1091-1107. https://doi.org/10.1016/j.concog.2010.09.010

Wong, C., Amini, R., & De Koninck, J. (2016). Automatic gender detection of dream reports: A promising approach. *Consciousness and Cognition, 44*, 20-28. https://doi.org/10.1016/j.concog.2016.06.004

World Health Organization. (2002). *The world health report 2002: Reducing risks, promoting healthy life*. Geneva, Switzerland: World Health Organization.

Zadra, A. L., Nielsen, T. A., & Donderi, D. C. (1998). Prevalence of auditory, olfactory, and gustatory experiences in home dreams. *Perceptual and Motor Skills, 87*(3), 819-826. https://doi.org/10.2466/pms.1998.87.3.819

Zheng, X., & Schweickert, R. (2021a). *Comparing Hall Van de Castle coding and linguistic inquiry and word count using canonical correlation analysis*. Manuscript submitted for publication.

Zheng, X., & Schweickert, R. (2021b). *Differentiating dreams from wake reports by automatic content analysis and support vector machine*s. Manuscript in preparation.

Zheng, X., Schweickert, R., & Song, M. (2021). *Test continuity theory by automatic content analysis and achieve binary characteristics detection*. Manuscript in preparation.