Automated Text Analysis in Psychology:
Methods, Applications, and Future Developments

Rumen Iliev, University of Michigan

Morteza Dehghani, University of Southern California

Eyal Sagi, Northwestern University

Author Note

Correspondence should be addressed to the first author at riliev@umich.edu. This research has been supported in part by an AFOSR Young Investigator award to MD, and ARTIS research grant to RI. We are thankful to Jeremy Ginges, Sid Horton, Antonio Damasio, Jonas Kaplan, Sarah Gimbel, Kate Johnson, Lisa Aziz-Zadeh, Jesse Graham, Peter Khooshabeh, Peter Carnevale, and Derek Harmon for their helpful comments and suggestions.

Abstract

Recent years have seen rapid developments in automated text analysis methods focused on measuring psychological and demographic properties. While this development has mainly been driven by computer scientists and computational linguists, such methods can be of great value for social scientists in general, and for psychologists in particular. In this paper, we review some of the most popular approaches to automated text analysis from the perspective of social scientists, and give examples of their applications in different theoretical domains. After describing some of the pros and cons of these methods, we speculate about future methodological developments, and how they might change social sciences. We conclude that despite the fact that current methods have many disadvantages and pitfalls compared to more traditional methods of data collection, the constant increase of computational power and the wide availability of textual data will inevitably make automated text analysis a common tool for psychologists.

Keywords: automated text analysis, psychological variables, demographics, technology, big data, psycho-informatics

Automated Text Analysis in Psychology: Methods, Applications, and Future Developments

Technological innovations that allow scientists to collect qualitatively different types of data have facilitated some of the most important theoretical advances in psychological science. A few examples of such novel tools for data collection include the measurement of precise reaction time (Helmholtz, 1850), mechanical control of stimuli exposure (Mueller & Schumann, 1894), measurement of galvanic skin response (Vigouroux, 1879; Jung, 1906), and electro- encephalograms (Berger, 1929). More contemporary examples include fMRI (Ogawa, Lee, Kay, & Tank, 1990) and optical brain imaging (Villriger & Chance, 1997). Regardless of the concrete theoretical questions being asked, the access to different types of data has been central for the success of social sciences. Recently, however, social scientists have been facing not a qualitative, but a quantitative change in technology. This change can be summarized in two main points: 1. the availability of vast amounts of human-related data, and 2. constantly increasing computational power. Some of this data is already in analysis-friendly form, such as social network information (Lewis, Kaufman, Gonzalez, Wimmer, & Christakis, 2008; Lerman & Ghosh, 2010), diurnal activity patterns (Krishnamurthy, Gill, & Arlitt, 2008), reputation (Standifird, 2001), or Facebook “likes” (Kosinski, Stillwell, & Graepel, 2013). An enormous amount of data, however, is in the form of human generated text, and that is not something that can be directly analyzed. Despite the difficulties of using computer algorithms for analyzing written text, the field is quickly developing. Different companies offer specialized software for automated text analysis, and more recently tools for text analysis have become part of standard statistical packages (e.g., SAS Text Miner, SPSS Text Analytics, R). Given the growing importance of such methods for social scientists, in this paper we review some of the main approaches that have been used to derive measures of subjective properties of individuals or groups based on the texts they produce.

Our primary goal here is to describe the most popular methods for inferring authors' characteristics in large bodies of text and to describe how such methods can be useful for social scientists. Since automated text analysis can be used for collecting many different types of psychologically relevant data, our focus will be on the methods themselves, rather than on the particular domains of application. However, to illustrate different methods, we will use a broad set of examples, including some from clinical psychology, personality and individual differences, intelligence, knowledge assessment, lie detection, political attitudes, group dynamics, and cultural change. The rest of the paper is organized as follows: First, we briefly introduce the idea of using language in general and text in particular as a source of information about the author. Next, we discuss three popular approaches for automating such tasks: 1. User-defined dictionaries; 2. Extraction of language features that maximize predictive accuracy; 3. Patterns of word co-occurrence in a semantic space. After the outline of the three major approaches, we briefly describe some less popular, but promising, recent developments in automated text analysis. We conclude with a discussion of the pros and cons of the different methods and speculate about the future of automated text analysis in social sciences.

The Gold Standard in Natural Language Processing: Human Coders

At first glance, using existing written text or speech transcripts for inferring properties of a person is a straightforward idea. Not only is language often used by psychologists to make inferences about properties of the human mind (Freud, 1901; Rorschach, 1921; Murray, 1943; Van Dijk & Kintsch, 1977; Weber, Hsee, & Sokolowska, 1998; Braun & Clarke, 2006), but it is also our primary mode of communication, and frequently our source of information about others. Based on what someone says, we make judgments about personality, general knowledge, past, and, quite often, about the value of future interactions with the speaker. We are similarly good at interpreting written text:

when we read a note, email, letter, or article, we can often tell if the author was happy or sad, polite or rude, expert or novice, and sometimes we can even infer gender, religion, or political orientation. This ability has allowed social scientists to collect data using human coders as interpreters of spoken or written interviews. Using human coders exclusively, however, quickly becomes impractical as the amount of text increases. In today's standards, with millions of new tweets, blog posts, comments, and reviews generated daily, traditional methods that rely on human coders can easily limit the scope of research projects.

If we need to deal with large volumes of text, automated text analysis quickly becomes the most plausible option. However, in doing so we face the problem of extracting meaning from text, and, while humans are strikingly good at this, computer algorithms find it particularly challenging. Despite decades of research on natural language processing by computer scientists, computational linguists, and cognitive psychologists, computers are still a long way away from matching human performance when it comes to identifying meaning. To illustrate the challenges that a computer program faces when trying to extract meaning, we will use a relatively simple example from the field of sentiment analysis. Imagine that we want to understand whether or not a person is happy with a particular camera. By using a parsing algorithm, it is not difficult to discover if the noun "camera" is linked to a value-laden adjective. Some reviewer might write: "This is an awful camera", while another one can write: "This is an awesome camera". After consulting a dictionary to evaluate the meaning of words (e.g., Esuli & Sebastiani, 2006), we can easily conclude that the first reviewer likes the camera while the second one does not. Unfortunately, easy sentences like the examples above are not that frequent, and in most cases the semantics of an adjective changes based on context. For example, in a blog post an expert might be comparing two cameras, and if she says: "The battery life of this Nikon is really *long*", she probably has a positive attitude, but if she says: "The focusing time of the Pentax is really *long*", the

reviewer is probably expressing a negative attitude (after Liu, 2010). The same word used for the same product might have very different meanings, depending on the particular feature being described.

Even though a fully functioning, automated extraction of meaning from text is not yet possible, researchers have made progress in using large-scale bodies of text as data sources. Avoiding the direct challenge that semantics presents to automated text analysis, most methods rely on the fact that computers can deal with large numbers of relatively simple features. Even if each feature captures a very small proportion of the meaning of a text, when many features are taken into account, the accuracy of predictions can become surprisingly high. For presentation purposes, we split the methods into three major groups, depending on the properties of the features they use for analysis. In the first group of methods, which we call User-Defined Dictionaries (UDD), researchers generate the features themselves. In the second group, which we call Feature Extraction, researchers use computer algorithms to find the features that are the strongest predictors for some variables of interest. In the third group, which we call Word Co-occurrence the focus is on the relationship between features. Since this three prong-distinction is for presentation purposes only, it leaves out a number of other methods, which we will briefly cover under Other Methods. Before we continue with the descriptions of the methods, it should be noted that this review is very broadly aimed at social scientists. Hence, we will minimize our discussion of the technical details and differences between the various methods and focus more on their applications.

User-Defined Dictionaries

Probably the most straightforward way to explore how language is linked to the properties of the speaker is to look for particular themes in his or her speech. Since many of the problems relevant to psychologists are often reflected in language, one can predefine sets of words associated with particular topics. If a researcher is interested in the general mood of a person, the focus can be placed on the emotional value of the words in text. For example, one can predefine a dictionary with negative words, such as *sad, depressed, gloomy, pain*, etc. Similarly, if a researcher cares about personality, the focus can be on adjectives describing a person, such as *fun, cool, social, easygoing*, etc. Then, the text is searched for the words from a particular dictionary, and the relative number of hits can be used as an indicator of the degree to which the text is related to a specific theoretical construct. This procedure is similar to content analysis of text using human coders, with the main difference being that in dictionary-based methods, the categories of interest are represented by single words, so a computer algorithm can automatically search through large bodies of text.

The most popular example of a dictionary-based approach in recent years[1] is the Linguistic Inquiry Word Count (LIWC[2]) developed by James Pennebaker and his collaborators (Pennebaker, 2011). LIWC has been extensively applied to various psychological domains (for a more detailed review, the reader should refer to Tausczik & Pennebaker, 2010). LIWC performs word counts and catalogs words into psychologically meaningful categories. The default LIWC2007 dictionary includes 76 different language categories containing 4,500 words and word stems. LIWC assigns each word to a specific linguistic category and reports the total number of words in each category normalized by the total number of words in the document. Some of these categories are related to specific contents, such as leisure, religion, money, or psychological processes. As a simple example, when people write about pleasant events, they are more likely to use words from the dictionary representing the positivity

category (Kahn, Tobin, Massey, & Anderson, 2007). Similarly, depressed individuals score higher on words associated with negative emotions (Rude, Gortner, & Pennebaker, 2004)[3]. Some content-based categories have shown more surprising associations. For instance, when subjects are trying to write deceptive texts, they are more likely to use words from the motion category (Newman, Pennebaker, Berry, & Richards, 2003), and extroverted subjects are less likely to use causality related words (Pennebaker & King, 1999).

In addition to content-based categories, LIWC also analyzes some broader language categories, such as word count, long words, tense and function words (articles, pronouns, conjunctions). Somewhat surprisingly, several of the most interesting findings come from these categories. For example, first-person singular pronouns have been associated with negative experiences (Rude et al., 2004). Suicidal poets are more likely to use first-person singular pronouns than matched non-suicidal poets (Stirman & Pennebaker, 2001). Similarly, depressed people are also more likely to use first-person singular pronouns (Rude et al., 2004), and the same is true for people in lower power positions (Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2013). Further, individuals under stress are also more likely to use first-person singulars; however, when a whole community copes with a tragedy, such as the terrorist attack on September 11, 2001, the usage of first-person plural pronouns decreases (Cohn, Mehl, & Pennebaker, 2004). Other language features have also been linked to psychological variables: Extroverted authors, for example, tend to write longer texts but prefer shorter words and less complex language (Pennebaker & King, 1999; Mehl, Gosling, & Pennebaker, 2006).

One of the main advantages of user-defined dictionaries (UDDs) is that researchers have the freedom to create sets of words that can target any theoretical construct of interest. While the categories of LIWC can be applied to multiple domains, and have the advantage of being empirically validated in many studies, sometimes scientists have to build their own dictionaries. For example, Graham, Haidt and Nosek (2009) used a specialized dictionary as part of an extensive empirical test of the Moral

Foundation Theory (Haidt & Joseph, 2004; Graham et al., 2013). One implication of this theory is that liberals and conservatives differ in what they consider to be part of the moral domain. Graham, Haidt, and Nosek (2009) tested this prediction in three psychological studies, finding that liberals were concerned mainly with harm and fairness, while conservatives in addition were also concerned with loyalty to ingroup, authority, and purity. In their last study, the authors tested whether such differences can be measured in text corpora. They created a dictionary with words corresponding to each of the five moral foundations, and compared liberal and conservative sermons. The analysis of the relative frequencies of the words from the different subdictionaries largely replicated the questionnaire-based results: Liberal sermons had a higher frequency of words related to harm and fairness, while conservative sermons were higher in words related to authority and purity[4].

A recent domain where UDD has shown to be particularly useful is studying historical trends and cultural change. While historical analysis of text has been a common practice among psychologists interested in cultural change (e.g., Wolf, Medin, & Pankratz, 1999), the current availability of large-scale time-stamped text (Michel, et al., 2011) has made such studies particularly detailed and easy to conduct. For example, cultural researchers often focus on East-West cross-cultural differences that can be traced back to ancient philosophical texts (Nisbett, 2004). Such focus on static comparisons, however, ignores the possibility that some of the characteristics associated with western cultures might be a recent development. One way to assess the degree to which values and attitudes have changed over time is to create specific UDDs and see the temporal pattern associated with particular words and expressions. Grienfiled (2013) found that, over the last two centuries, words associated with individualism and independence have become more frequent. Similarly, Kessibir and Kessibir (2012) have found that words expressing concern about others and words indicating moral virtue have decreased in frequency over the last century. Similar pattern have been found by Twenge, Campbell and Gentile (2012) using a narrower time window and participant-generated dictionaries.

The UDD approach has proven applicable to a broad range of questions, including gender differences, personality, clinical diagnosis and treatment, morality, deception, motivation (e.g., Gill, Nowson, & Oberlander, 2009), knowledge assessment (Williams & Dmello, 2010), and cultural epistemological orientations (Dehghani, Bang, Medin, Marin, Leddon, & Waxman, 2013). Nevertheless, we need to mention some of the challenges that the method faces. Firstly, in its basic form (as counts of words appearing in particular user-defined categories), it is blind to the context in which words appear. For example, if we have a dictionary of positive words, we will treat the sentences "I have never been happy in my life" and "I have never been this happy my life" very similarly since they both include the word happy. Moreover, such dictionaries cannot easily capture sarcasm, metaphors, or idiomatic expressions. Consequently, while the statistical tests of studies using dictionaries are typically highly significant, the effect sizes are often quite small[5]. Yet, the simplicity and theoretical flexibility of this method makes it a very useful tool for working with large bodies of texts, and its popularity will probably keep increasing.

Feature Extraction

While user-defined dictionaries have the advantage of high face validity, the small effect sizes associated with the different categories often make them less suitable for precise predictions, especially when dealing with large, noisy datasets. For example, researchers might be interested in gender differences in diurnal activity, and their data might consist of a large number of blog posts. Further, suppose that all blog posts are time stamped, but only for a small proportion is the gender known. How can they make use of the ones with missing data? As we saw in the previous section, they can rely on some LIWC categories, or they can construct their own list of features using their intuition or expertise, and then assign gender to the missing values depending on the overlap between their lists and given

text. However, small effect sizes imply that they will not be much better than chance, which will result in lots of noise in their subsequent findings.

An alternative approach to UDD is to start with texts that differ in some dimension of interest, and then in a bottom-up manner find the features that maximize such differences. In theory, these features can have any property, but in most cases they are character n-grams, single words, short expressions, or tagged parts-of-speech. Typically, in algorithms used to build text classifiers, documents are represented as sets of features and then the algorithm searches for those features that are common in one type of document, but rare or absent in the other types[6]. In the example above, the texts whose gender is known can be used as input to the classification algorithm. The algorithm then tries to extract the features that are more likely to be present in texts written by females only (or by males only). Such algorithms are trained on a subset of the texts, and then the predictive validity of the extracted features is tested on the remainder of the texts (a common method in machine learning). Depending on the particular goals of the project, the training subset might be much smaller than the remainder, or it can be larger, but the procedure can be randomly repeated many times (known as cross-validation). The degree to which the extracted features from the training set correctly classify the remaining texts is taken as an indicator of its reliability/accuracy[7].

Feature extraction methods have shown impressive accuracy in predicting a wide range of properties of the speaker. For example, Dave, Lawrence, and Pennock (2003) used these methods to distinguish between positive and negative reviews with relatively high accuracy (88% for products and 82% for movies). Similar applications have also been able to identify the political party affiliations of U.S. senators based on their speeches in the senate with 92% accuracy (Diermeier, Godbout, Yu, & Kaufmann, 2011), as well as political orientations of bloggers with 91.8% accuracy (Dehghani, Sagae, Sachdeva, & Gratch, 2014). Likewise, feature extraction algorithms have been found to perform well at attributing gender (Mukherjee & Liu, 2010), age and native language (Argamon,, Koppel, Pennebaker,

& Schler, 2009), personality dimensions (Oberlander & Nowson, 2006), sentiments (Dave, Lawrance, & Pennock, 2003), mental disorders (Strous, Koppel., Fine, Nachliel, Shaked, & Zivotofsky, 2009), and identity of the author (Diederich, Kindermann, Leopold, & Paass , 2003; Lewis, Kaufman, Gonzalez, Wimmer, & Christakis, 2005) with similarly high accuracy levels (see also Koppel, Scheler, & Argamon, 2009, for a detailed treatment of the topic of author identification).

It is important for the reader to keep in mind that these impressive results are mostly due to the computational power of current algorithms and the availability and quality of the training sets. Since building the initial feature list is usually automated, these methods might consider a very large number of features. For example, when content words are used as features, the size of the list is often in the thousands, and lists of tens of thousands of features are not uncommon. The correspondence between a particular feature and variable of interest is not always transparent. For instance, even before we run any analysis, we might guess that the word “homosexual” in senators’ speeches will predict conservative ideology, while the word “gay” will predict liberal ideology. However, we could hardly guess in advance that “catfish” and “grazing” are also strong predictors for conservatism, while “fishery” and “lakes” predict liberal orientation (Diermeier et al., 2011). Weaker predictors are frequently even less intuitive to comprehend, yet when combined in large numbers, they can boost the predictive power of the model can be surprisingly high.

While machine learning algorithms used in feature extraction methods face many challenges, two of such shortcomings seem particularly important for the current review. One is practical and the other is theoretical. On the practical level, algorithms that are trained in one domain often perform much worse on other domains even when the variable of interest is the same. For example, Finn and Kushmerick (2006) compared the performance of algorithms that predict the valence of movie and restaurant reviews, and found that even though the algorithms performed well in the domain they were trained in, they did poorly in predicting reviews from the unfamiliar domain. This means that

algorithms might have to be retrained every time the topic of the document set is changed, which can be a significant limitation of their applicability.

The theoretical challenge for machine learning algorithms, from the perspective of social scientists, is that using thousands of features might lead to good practical results for text classification, but it might not be very informative for theoretical purposes. One way to overcome this problem is to analyze the results from a machine learning algorithm using more traditional forms of content analysis. One example comes from Diermeier et al. (2011) who extracted the words that maximized the difference between Republican and Democratic senators. The authors compared their results with common theories that the liberal-conservative divide in the Senate is mainly driven by economic concerns. Contrary to the economic divide hypotheses, the most predictive features were words related to cultural values and believes (e.g. abortion, same sex marriage, stem cell research). Not only their model was useful for assigning political orientation to text entries, but they were also able to make inferences that argue for or against a particular hypothesis. This demonstrates that the integration of machine learning results with classical scientific expertise can lead to findings with significant theoretical implications.

To a large degree, psychologists have been reluctant to use feature extraction methods in their work, yet this reluctance might change in the near future. One potential domain where such methods can be quite helpful to researchers is in the analysis of open-ended questions. Sometimes such questions are part of the dependent variable of interest or part of larger interviews with multiple items. However, due to the time and costs of using human coders to analyze these types of questions, often these answers do not get analyzed thoroughly, or the coders are instructed to look only for predefined keywords or expressions. Yet, such predefined coding schemes might miss genuine language differences between the groups of participants, which could otherwise be captured by a supervised learning algorithm. Another issue with using human coders in this context is that sometimes manual

coding schemes are developed simultaneously with the coding process, and, as such, they can maximize between-group differences by selecting the most distinctive features in the corpus. Such practice can lead to issues with reliability and generalizability, because we might not know if the same coding scheme will lead to the same results with a different sample of participants. The cross-validation methods used in supervised learning algorithms can help circumvent this problem by dividing the corpus into separate learning and testing parts and measuring the predictive accuracy of the extracted features. Lastly, in many cases open-ended questions are used as filler tasks, included in surveys to reduce carry-over and order effects or to cover the true purpose of the study. Typically such questions are not analyzed at all, yet they might be affected by the experimental manipulation, or they might interact with the target question. Supervised learning algorithms might be a quick and inexpensive way to test for the presence of such valuable information. For example, in an experimental study in social psychology, a researcher might choose to use a filler task to separate two measures of relevant dependent variables. In the filler task, participants might be asked to write a short essay on how they have spent the weekend, a topic unassociated with the main goal of the study. In most cases, such essays will not be analyzed, particularly in large sample experiments. Yet, applying a supervised learning algorithm might reveal that the experimental and the control group used very different words and expressions in their essays, and these differences might be theoretically meaningful. Further, the prevalence of the features extracted by the algorithm might become a useful variable in mediation analysis. While such studies are yet to be seen, we believe that the potential of feature extracting methods for application in different types of psychological studies is very high.

Word Co-occurrences

The methods mentioned so far have all focused on analyzing documents mainly by looking at individual words. In the case of user-defined dictionaries, each word that is found in the dictionary

contributes to the documents' overall score in one (or more) dimensions. Supervised machine learning algorithms usually consider words (but also n-grams or syntactic structures) to be features that might be of use in determining the documents classification. However, words are rarely used in isolation. Words in a text often build on each other in order to convey a meaning that each word on its own cannot provide. In that sense, a text is more than the sum of the words it contains. One way to take this into account is to focus on which groups of words tend to occur together in particular contexts. For example, if the words lion, tiger, and zebra appear in text, it is more likely that such documents refer to animals, than a text containing the words stop, yield, and zebra. Capturing the relations between words is the goal of a family of methods that are referred to as Latent Semantic Analysis (LSA[5]).

At the core of the LSA family of methods is the assumption that words are not randomly distributed (Firth, 1957). Consequently, it is possible to reduce the conceptual vocabulary to a smaller number of independent dimensions based on word co-occurrences. In traditional LSA, this set of dimensions forms a semantic space. The method of choice for constructing such a space is singular value decomposition, which is closely related to principal component analysis. The algorithm uses a matrix as an input to describe the frequency of word occurrences within documents. This matrix is then decomposed into matrices that describe both the documents and the words as vectors in a multidimensional space. Words that tend to appear in the same documents are closer in this space, and, similarly, documents that use similar words are also close to each other. Since the space is shared, one can also compute the distance between a word and a document. For example, the word *carburetor* will be closer in semantic space to documents related to car maintenance than to documents about culinary arts. While this method has strict mathematical basis, it is also flexible enough to be adjusted for different practical purposes. For example, one can choose the length of a document, where the document can be a single sentence or it can be a whole book. Similarly, one can choose the number of

dimensions, with numbers between 100 and 400 typically showing best trade-off between simplicity and informativeness (Dumais & Landauer, 1997).

As a side benefit of this process, it should be noted that no human intervention is required to create a semantic space. That is, while the methods described earlier all required an external source of information (either a set of norms to assign words to categories or a set of preclassified documents to train the algorithm), LSA generates a space based solely on the content of the documents in the corpus. In a sense, the corpus itself forms the training set from which LSA learns. Choosing different training corpora, however, can result in different semantic spaces, since word frequencies, and word co-occurrences depend on genre, style and topic. In this sense, how the reduced semantic space will be extracted is independent of the researcher, yet the training corpus needs to be carefully selected. For example, a training set from a Biology textbook will position the term "horse" in a very different semantic space than a training set in History.

It is also important to note that the development of LSA methods, which have their roots in document retrieval, has been originally focused on what is shared across people rather than on what is different. Specifically, LSA was originally developed to address a problem in information retrieval, namely that the same information is often described differently in very different terms by different people. From this perspective, idiosyncrasies, both in documents and in search queries, were seen as an obstacle for computerized search rather than as useful information. Similarly, when the method was applied to cognitive modeling, the interest was still largely in the shared semantic representation (Landauer & Dumais, 1997). Yet if we are interested in individual differences or in between-group differences, the method itself does not prevent us to ask the opposite question, namely, how authors differ between each other.

The most straightforward application of LSA methods for studying differences between authors is simply to check the degree to which authorship is a predictor of distance between texts in the

semantic space. While there is evidence that authors tend to cluster together in the semantic space, the effects are not particularly strong and are sensitive to the number of dimensions chosen (Nakov, 2001). Such results should not be surprising, since LSA tends to cluster documents based on semantic similarity, so that documents from different authors on the same topic are more likely to be close in the semantic spaces, than documents on different topics by the same author. Notice, however, that these results came mainly from analysis of texts by established writers and poets, and it is still an open question if broadening the sample of authors using online entries will lead to stronger within-author similarities.

Another way to use LSA methods for assessing properties of the author is to measure the semantic distance of text to some target document. Such an approach has been particularly important for education researchers, with the hope of reducing manual labor in grading and providing quick online feedback for various types of classes. Typically, a set of pregraded texts are decomposed in a multidimensional space, after which the text to be graded is also decomposed in the same space. The proximity between the new text and the pregraded exemplars is used as a criterion, where the new grade is based on the grade of closest exemplar. Alternatively, it is also possible to start only with a set of "ideal" texts, without having lower-grade exemplars, and the grade to be assigned will depend only on the distance to the "ideal". These methods have been applied to different domains of learning, often leading to impressive results. Foltz et al. (1999) report correlations between human graders and LSA-based grade around .80, which was virtually the same as the correlation between two human raters.

Sometimes it is not enough to grade an essay, but it is also important to distinguish the particular "mental model" that a student relies upon. In research on naïve physics (McCloskey, 1983; DiSessa, 1993), for example, researchers found that while there is one normatively correct model, there are several categories of non-normative models, based on varieties of misconceptions. In terms of grading, different answers based on such non-normative models will all be far from the "ideal," but an

educator might further want to know which are the most widespread misconceptions so they can be addressed. A promising example in this regard comes from Dam and Kaufmann (2008), who applied LSA to interview transcripts of seventh graders who were asked about the cause of seasons on the Earth. Previous research has isolated three main types of models: close/far distance to the Sun; facing/not facing the Sun; tilted axis of rotation of Earth (normatively correct). The researchers trained their algorithm on texts from geology and astronomy, after which they measured the similarity of sets of interviews to three comparison documents, each representing one of the three mental models. The authors' LSA-based application reached 90% accuracy of classification when compared to human raters (see also Sherin, in press).

LSA-based methods can also be useful for studying change over time. One example comes from Campbell and Pennebaker (2003), who were interested in the relationship between writing and health. The authors used data from previous studies, where different groups of subjects had to write short essays on emotional topics for three days. The similarity between the essays from the same author was correlated with health outcomes. When conventional LSA, based on content words, was applied, the authors did not find any meaningful pattern. However, when they adjusted the LSA procedure to account for style, rather than content, the authors found that greater similarity between the essays written on different days was significantly correlated to subsequent medical visits. In other words, people who showed less diversity in their writing styles were more likely to have negative health outcomes. The results held for all three groups, with correlations between essays' similarity and doctor visits in the range of .34 to .51. Subsequent analysis showed that changes in the use of pronouns and particles were the strongest predictors. While the true mechanism behind such strong effects is not fully understood, one explanation suggested by the authors is that change of the context in which pronouns are used can reflect flexibility in perspective taking and thus reevaluation of emotional experiences.

A closely related approach to LSA is a generative method called Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003). LDA[9] assumes that each document in the corpus under analysis is composed of a mixture of topics, and that each topic is a distribution over words. According to this view, documents are generated by repeatedly choosing a topic from a distribution over topics, then choosing a word from a distribution over words which correspond to the chosen topic. The goal of the method then is to find the most likely topic structure that generates the given collection of documents. Chen (2012) illustrates the intuition behind LDA using the following example: Let's assume we have the following sentences in our corpus of analysis: "I like to eat broccoli and bananas. I ate a banana and spinach smoothie for breakfast. Chinchillas and kittens are cute. My sister adopted a kitten yesterday. Look at this cute hamster munching on a piece of broccoli." (Chen, 2012). Given this corpus, and asked for two topics, LDA would try to find the set of two topics that cover the corpus. For instance, LDA could very well discover that the first two sentences are about one topic, sentences three and four about another topic, and the last sentence is a mixture of the first two topics. Given these topics, LDA would also tell us the distribution of the words that compose them. For example, the first topic could include words such as "broccoli," "bananas," "spinach," and "munching" (with probabilities associated with each), and the second topic "kittens," "chinchillas," and "hamster"[10].

Whereas the model underlying LSA is that the meaning of a word can be described as a position in a multidimensional semantic space, LDA makes a weaker assumption and focuses only on the statistical dependence among words. Consequently, the two methods choose very different tools–LSA relies on factorization methods while LDA is rooted in a statistical model of language. Nevertheless, both methods function similarly in general cases, such as those discussed above. However, LDA has an advantage when the researcher is interested in identifying the topics addressed by a corpus. While the dimensions of LSA are abstract and meaningless, the dimensions in LDA are the latent topics that emerge from the corpus. An interesting use of LDA includes the Expressed Agenda model (Grimmer,

2010), which is used to examine authors' priorities through the topics addressed in their language output. Another application comes from Yano, Cohen and Smith (2009), who modeled different characteristics of a collection of political blogs using topic modeling.

LDA can also be used in a supervised, or semi-supervised, manner. As example of the former, Schwartz et al. (2013) used 2000 topics provided by previous LDA analysis, and computed the topic content of Twitter data by county. Using 10-fold cross-validation, they found that LDA topics predicted well-being by county beyond the variance accounted by demographic variables. As an example of semi-supervised application, LDA with topic-in-set knowledge (Andrzejewski & Zhu, 2009) is used to seed small sets of words in a subset of the topics and thereby adding a level of supervision to the process. This semisupervised approach combines the advantages of unsupervised topic modeling using LDA with the ability of encouraging the emergence of certain topics in the model through small sets of words selected from the outset as prior knowledge. However, instead of simply searching for the most probable set of latent topics, a subset of the topics can be initialized to contain specific words. For example, Dehghani, Sagae, Sachdeva, and Gratch (2014) use small sets of words selected from the Moral Foundations Dictionary (Graham, Haidt, & Nosek, 2009) as seeds to encourage the emergence of topics related to different moral concerns, and examined similarities and differences in how such concerns are expressed between liberals and conservatives.

Other Methods

The three-prong classification of methods that we have used so far is rather crude and schematic, and it inevitably leaves out many useful techniques. While we cannot cover all recent developments, we will mention three other types of methods that might be of use for social scientists.

Semantic Role Labeling

As could be seen from our review so far, the methods that have become popular rest on rather straightforward ideas and are relatively easy to implement. Future methods, however, will most likely increasingly rely on syntactic and semantic information, going far beyond simple features and word co-occurrences. One promising development in this direction is the semantic role labeling approach. The main idea behind this approach is that a typical sentence consists of basic information about *who* did *what* to *whom,* and this information about actions, agents, and patients becomes available after a sentence is parsed. For example, encountering the sentence "Mary greeted John," we can easily assign agent-hood to Mary and patient-hood to John. Some sentences might also have more specific information, about *how*, *when,* and *where* the event has happened, and although such information might be idiosyncratic to particular actions or events, computational linguists and computer scientists have been working on systems that will encode both basic and idiosyncratic semantic information. One popular example is FrameNet[11], which is a hand-annotated system that divides events into frames, with each frame being associated with different elements which correspond to different semantic roles (Baker, Fillmore, & Lowe, 1998). Another similar system that has gained popularity is PropBank[12], which is centered around verbs rather than events (Kingsbury & Palmer, 2002).

While current semantic-role labeling does not seem to outperform simple, semantically-blind methods in ordinary text classification tasks (Houen, 2011), we believe that this approach might be of particular interest to social scientists. One reason is that semantic-role labeling methods are focused on causal relationships between entities, and as such can gather information about the set of beliefs that a person has based on simple claims. Instead of splitting people into groups of Republicans or Democrats, or those who like Starbucks versus those who do not, gathering information about causal beliefs could allow researchers to focus on knowledge representation of individuals. For example, it might not be enough to know if a person thinks that climate change is happening, but it might be more

important to know what this person thinks the particular causes behind the process are and how these beliefs are aligned or not with other beliefs that person holds.

Another way in which semantic-role labeling might be of interest to social scientists is that it considers information about the author of a text separately from the semantic agents and patients in the text. The other methods we have discussed are exclusively focused on the psychological and demographic properties of the author. Using semantic-roles, however, provides researchers with the opportunity to distinguish between the author and the opinion holder (Kim & Hovy, 2006). For example, an author might write this about his friend John: "John likes Mary," but the author might not like Mary. Further, the author might be wrong, and in fact John might not like Mary at all. For scientists interested in social relations, group dynamics, or conflict resolution, such information might be very valuable, since a researcher might learn not only what the author thinks on the topic, but also what the author thinks other people think on the same topic.

Cohesion

While the previous methods we discussed in this paper focus on extracting information found in the semantic content of texts, it is also possible to learn about an author by examining *how* they write. One concept that helps us distinguish between different types of writing is to look at the cohesion of a text, which in broad terms can be defined as how structural and lexical properties of language are combined together to convey meaning. Coh-Metrix[13] (Graesser et al., 2004) is a recent development in the field of automated-text analysis that is focused on cohesion as a central property of text and discourse. This approach combines multiple linguistic features, such as lexical diversity, semantic overlap between different parts of the text, connections between propositions, causal links, and syntactic complexity. While initially this approach was focused on readability, coherence, and complexity of text, particularly in the domain of education, it has been successfully applied to detecting

different properties of the author in other contexts. The method has been used for author identification (McCarthy, Lewis, Dufty, & McNamara, 2006), analysis of political speeches (Venegas, 2012), inferring affective states from transcripts (D'Mello., Dowell, & Graesser, 2009; D'Mello & Graesser, 2012), essay grading (McNamara, Crossley, & McCarthy, 2010), and evaluation of social skills (Xu, Murray, Park, & Smith, 2013)[14].

Hybrid Methods

As the discussion above indicates, different methods of text analysis rely on different features of the text and use different statistical techniques for analyzing these features. As a result, they each provide complementary advantages and shed light on different aspects of the corpus. A common practice among computer scientists and computational linguists has been to compare multiple methods on the same task, looking for the most effective tool in terms of accuracy, speed, and computational cost. With the advancement of the field, however, it becomes clear that some method might be better for one aspect of a problem, while others for another.

Consequently, hybrid methods have recently emerged as a promising new approach by taking advantage of the power and flexibility that different techniques provide. For example, Gill, French, Gergle, and Oberlander. (2008) studied language correlates of emotional content in blogs. They found that UDD categories correlate well with joy and anger, but word co-occurrences methods were also able to detect fear. Use of LDA along with the words from the Moral Foundations Theory by Dehghani et al. (2014), discussed above, is another example of using a hybrid method.

Another form of hybrid methods is the combination between manual work and automated algorithms. For example, typically manual work might be automated, or alternatively, manual work might be used as a model for an automated algorithm. An instance for the first case is automated UDD methods. Recall that UDD methods stemmed from categories constructed by humans, usually with the

explicit goal of coding text. Yet with the increasing availability of large databases of semantic relations, such as WordNet (Miller, 1995), it is possible to use automated algorithms for building dictionaries based on one or another semantic relation (Kim & Hovy, 2004; Mishne, 2005) with minimum human input. Second, it is also possible to have typically automated methods learn from human input. For example, you can have a method that uses manual annotations in conjunction with supervised machine learning techniques. In such applications, human coders are asked to manually code and classify particular features in the text, and then machine learning algorithms are used to build models based on these annotated features and classify other sections of the corpus. For example, Sagea et al. (2013) use hand-coded annotations of different narrative levels to train a text classification algorithm for classifying a corpus of narratives and achieved an accuracy of 81%.

One last type of hybrid method that we need to mention before moving to the discussion is network text analysis, which combines properties of word co-occurrence methods, semantic role labeling, and social network analysis. By treating text as a network of inter-related concepts, such methods have been used to analyze what knowledge is shared between different authors and what is unique. Although such methods stem from the idea that word co-occurrences in text reflect cognitive organization of authors' concepts or thoughts, somewhat surprisingly they have been a less popular tool for inferring psychologically relevant characteristics of the author compared to the approaches described above. Nevertheless, since these methods can easily account for different types of contextual information, such as location, time period, or social networks, their popularity among behavioral researchers might increase (Carley, 1997[15]; Popping, 2003).

Discussion

Empirical sciences are as good as their data are, and social scientists have been particularly creative when looking for new ways to address basic questions about how the mind and the society

work. Among the many types of data that have become widely available in recent years, human generated text is both very common and very hard to analyze. Since full extraction of meaning from text is still not possible, different methods have been developed to make use of textual information. Here we have reviewed three major approaches that can be of use for social scientists. In the UDD approach, the researcher preselects words or expressions that might be of theoretical interest. Alternatively, in feature extraction methods, a computer algorithm looks for words or expressions that are more likely to be found in some types of texts but not in other. In word co-occurrence approaches, the researcher is interested in the semantic context in which words appear. We also discussed semantic roles, cohesion, and hybrid methods that are becoming increasingly important tools. Each of these approaches has pros and cons, and a researcher can choose different tools depending on the particular goal of the project. UDDs are probably the most straightforward to use. They are also very suitable for testing specific hypotheses by developing theory-motivated dictionaries. Feature extraction methods are superior for large-scale text classification tasks, where the researcher wants to infer various attributes of the author. Such methods are usually theory-blind, and the features they extract are not easily generalizable across tasks or populations.

What will be the future of automated text analysis in social sciences? While it will not replace any of the major methods of psychological data collection or analysis, we believe that it will become increasingly important. The current methods will become more refined, and there will be more empirical work comparing the values of different methods. Such comparisons will most likely also result in packages that integrate a variety of methods, leading to increased flexibility of the analyses and accuracy of predictions. More labs are developing UDDs, and the sharing of their work will help in building large libraries that will cover wide range of psychological topics. Further, since text data nowadays is often accompanied by social networks, behavioral, time of day, and geographical location data, these additional dimensions can easily be used in the training of supervised learning algorithms.

We also believe that the near future will bring closer collaboration between different fields, where computational linguists and computer scientists will work more often with psychologists and cognitive scientists.

While here we have been concerned mainly with automated text analysis as a tool for analysis of data on demographic variables and psychological states and preferences, text analysis can also lead to more abstract developments in social sciences. One example is related to the sheer amount of psychologically relevant data that will become available in the future (see King, 2005; Miller, 2012; Yarkoni, 2012). Typically, development in social sciences follows the path of initial observation, theory building, then empirical testing, and the final step is often empirical comparison between different theories. Although not always true in practice, the textbook example of research design in social science suggests deriving a theoretical prediction, which then is translated into a precise hypothesis, which leads to data collection to test this hypothesis. With large scale data collection, however, researchers will have access to variables that they have never been concerned about, which could easily lead to novel and unexpected advances based on accidental discoveries rather than on solid theoretical hypotheses. From this perspective, one potential change that automated text analysis methods might lead to is the increased role of bottom-up built theories.

Even though the main purpose of this review is to encourage psychologist to add automated text analysis in their methodological toolboxes, we also need to raise a word of caution. While psychologists are well aware of the danger of systematic errors in data collection and data analysis, applying automated text analysis to real world data brings its own new risks. Similar to the integration of other novel technological developments, learning about these new risks in some cases will happen through trial and error. We illustrate this point with a recent example from analysis of the emotional content of text messages sent in the aftermath of September 11, 2001 (Back, Kufner, & Egloff, 2010). In this work, one of the most striking findings in the result of analyzing text messages was that the time

line of anger related words showed a strong trend that kept constantly increasing for more than 12 hours after the attack. Subsequent reanalysis, however, discovered that some of the SMS messages were automatically generated by phone servers ("critical" server problem), and although irrelevant to the theoretical question, they were identified as anger-related words by the algorithm (Pury, 2011). Since both data collection and data analysis algorithms can contain numerous small steps, chances for hard-to-detect errors happening drastically increases, and small errors being repeated multiple times can easily lead to wrong conclusions (Back, Kufner, & Egloff, 2011).

Before we conclude, we want to raise one last, yet very important question. This question is not about methodological development or theoretical implications, but about the ethical issues of doing research with text generated by people. While in many cases such texts are easily available, the "participants" have seldom agreed for their texts to be used in research. Typically, Institutional Review Boards treat observation of public behavior, or using publicly available data, more leniently since it presents a very low level of risk unless identifiable information is recorded. This might put online text that is not accompanied by IP addresses, email addresses, usernames, and social network data in an exempt category. Yet, since the applications of automated text analysis by social scientists will be related to inferring preferences, attitudes, beliefs, knowledge, psychological states, and demographic information, applying such methods will increase the chances that text excerpts might be enough to identify the author. While IRBs across universities have already made changes to accommodate using data from online surveys better, large-scale text analysis algorithms will inevitably raise novel ethical questions about balancing risks with societal benefits (for relevant discussions, see Hookway, 2008; Eastham, 2011).

Conclusions

Over the last century, psychologists and other social scientists have meticulously developed number of methods for collecting data. Usually a development of hypotheses, careful design, and construction of stimuli or survey questions all precede the data collection. For example, one of the most expensive and demanding types of studies in psychology are longitudinal designs, where researchers sometimes dedicate their whole career to a single long-running study (Vaillant, 2012). Yet in the last decade, those of us who use computers, and other networked devices, have become a part of an emerging longitudinal, cross-sectional and cross-cultural study where data is already being collected. A large part of this spontaneous data collection is in the form of text, which although hard to analyze, is becoming a focal point for multiple scientific fields. While the methods described in this paper are already impressive for some tasks, they are rather crude and ineffective for other problems. What is clear, however, is that these methods will only get better with time, and most likely the future of social sciences will be closely linked to these new developments.

References

Andrzejewski, D. & Zhu, X. (2009). Latent Dirichlet allocation with topic-in-set knowledge. *Proceedings of the NAACL 2009 Workshop on Semi-supervised Learning for NLP*. Association for Computational Linguistics, Stroudsburg, PA, USA, 43:48.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2), 119:123.

Back, M. D., Küfner, A. C., & Egloff, B. (2010). The emotional timeline of September 11, 2001. *Psychological Science*, 21(10), 1417:1419.

Back, M. D., Küfner, A. C., & Egloff, B. (2011). "Automatic or the people?" anger on September 11, 2001, and lessons learned for the analysis of large digital data sets. *Psychological Science*, 22(6), 837:838.

Baddeley, J. L., Pennebaker, J. W., & Beevers, C. G. (2013). Everyday Social Behavior During a Major Depressive Episode. *Social Psychological and Personality Science*, *4*(4), 445-452.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998, August). The Berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1* (pp. 86:90). Association for Computational Linguistics.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Berger, H. (1929). Uber das Elektrenkephalogramm des Menschen (On the human electroencephalogram). *Archiv f. Psychiatrie u. Nervenkrankheiten* 87:527–70.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993:1022.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2), 77:101.

Campbell, R. S., & Pennebaker, J. W. (2003). The Secret life of pronouns flexibility in writing style and physical health. *Psychological Science*, 14(1), 60:65.

Carley, K. (*1997*). Network text analysis: The network position of concepts. In C. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inference from texts and transcripts* (pp. *79*-100). Mahwah, NJ: Lawrence Erlbaum.

Chen, D. (Aug 2011). Introduction to latent dirichlet allocation. *Edwin Chen's Blog.* Retrieved from http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/

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.

Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, 519:528

D'Mello, S., & Graesser, A. (2012) Language and discourse are powerful signals of student emotions during tutoring. *IEEE Transactions on Learning Technologies*, 5(4), 304:317.

D'Mello, S., Dowell, N., & Graesser, A. (2009). Cohesion relationships in tutorial dialogue as predictors of affective states. In *Proc. 2009 Conf. Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*.

Dam, G., & Kaufmann, S. (2008). Computer assessment of interview data using latent semantic analysis. *Behavior Research Methods*, 40(1), 8:20.

Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, 519:528.

Dehghani, M., Bang, M., Medin, D., Marin, A., Leddon, E., & Waxman, S. (2013). Epistemologies in the text of children's books: native-and non-native-authored books. *International Journal of Science Education*, (ahead-of-print), 1:19.

Dehghani, M., Sagae K., Sachdeva, S. & Gratch, J. (2014). Linguistic analysis of the debate over the construction of the `Ground Zero Mosque'. *Journal of Information Technology & Politics.* 11, 1-14.

Diederich, J., Kindermann, J., Leopold, E., & Paass, G. (2003). Authorship attribution with support vector machines. *Applied Intelligence*, 19(1-2), 109:123.

Diermeier, D., Godbout, J. F., Yu B., & Kaufmann, S. (2011). Language and ideology in congress. *British Journal of Political Science*. 42 (01), 31 – 55.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1).

DiSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, 10(2-3), 105:225.

Dumais, S. T., & Landauer, T. K. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological review*, 104(2), 211:240.

Eastham, L. A. (2011). Research using blogs for data: Public documents or private musings? *Research in Nursing & Health*, *34*(4), 353-361.

Esuli, A., & Sebastiani, F. (2006, May). Sentiwordnet: A publicly available lexical resource for opinion mining. *In Proceedings of LREC* 6, 417:422.

Finn, A., & Kushmerick, N. (2006). Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, 57(11), 1506:1518.

Firth, J. (1957) *Papers in Linguistics (1934:1951).* London, UK: Oxford University Press.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. *In World Conference on Educational Multimedia, Hypermedia and Telecommunications,* 1, 939:944.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2).

Freud, S. (1901). *Psychopathology of everyday life*. New York: Basic Books.

Gill, A. J., French, R. M., Gergle, D., & Oberlander, J. (2008, November). The language of emotion in short blog texts. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*.

Gill, A. J., Nowson, S., & Oberlander, J. (2009, May). What Are They Blogging About? Personality, Topic and Motivation in Blogs. In the *Proceedings of 2009 International AAAI Conference on Weblogs and Social Media*.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2), 193:202.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96, 1029:1046

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S., & Ditto, P. H. (2013). Moral Foundations Theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, 47, 55:130.

Greenfield, P. M. (2013). The Changing Psychology of Culture From 1800 Through 2000. *Psychological Science*, 24(9), 1722:1731.

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, *18*(1), 1-35.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133, 55:66.

Helmholtz, H. (1850). Vorläufiger Bericht über die Fortpflanzungsgeschwindigkeit der Nervenreizung. *Archiv für Anatomie, Physiologie und Wissenschaftliche Medizin.*

Hookway, N. (2008). Entering the blogosphere': some strategies for using blogs in social research. *Qualitative research*, 8(1), 91:113.

Houen, S. (2011). *Opinion Mining with Semantic Analysis*. (retrieved from http://www.diku.dk/forskning/Publikationer/specialer/2011/specialerapport_final_Soren_Houen.pdf/)

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features* (pp. 137:142). Springer Berlin Heidelberg.

Jung, C. G. (1904–1907) *Studies in Word Association*. London: Routledge & K. Paul. (contained in Experimental Researches, Collected Works Vol. 2)

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2013). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*.

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*, 263-286.

Kesebir, P., & Kesebir, S. (2012). The cultural salience of moral character and virtue declined in twentieth century America. *The Journal of Positive Psychology*, 7(6), 471:480.

Kim, S. M., & Hovy, E. (2004, August). Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics.

Kim, S. M., & Hovy, E. (2006, July). Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*. Association for Computational Linguistics.

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, 331(6018), 719:721.

Kingsbury, P., & Palmer, M. (2002, May). From TreeBank to PropBank. In the Proceedings of the *International Conference on Language Resources and Evaluation*.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1), 9:26.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*. 110(15), 5802:5805

Krishnamurthy, B., Gill, P., & Arlitt, M. (2008, August). A few chirps about twitter. In *Proceedings of the first workshop on on-line social networks*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2), 211.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3), 259:284.

Lerman, K., & Ghosh, R. (2010, May). Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media* (ICWSM).

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In *Machine learning*: ECML-98 (pp. 4:15). Springer Berlin Heidelberg.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, 30(4), 330:342.

Liu B. (2010). Sentiment Analysis and Subjectivity. Nitin Indurkhya and Fred J. Damerau (Eds.) *Handbook of natural Language Processing*, Second Edition. Taylor and Francis

Madigan, D., Genkin, A., Lewis, D. D., Argamon, S., Fradkin, D., & Ye, L. (2005, June). Author identification on the large scale. In *Proc. of the Meeting of the Classification Society of North America*.

McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In *AAAI Workshop on Learning for Text Categorization*.

McCarthy, P. M., Lewis, G. A., Dufty, D. F., & McNamara, D. S. (2006). Analyzing writing styles with Coh-Metrix. In *Proceedings of the Florida Artificial Intelligence Research Society International Conference*.

McCloskey, M. (1983). Naive theories of motion. In *D.* Gentner & A. Stevens (Eds.) *Mental Models*. Psychology Press.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, 27(1), 57:86.

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, 90(5), 862.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The_Google_Books_Team, Pickett, J. P., Hoiberg, D., Clancy, D. ,Norvig, P., Orwant, J., Pinker, S., Nowak, M., & Lieberman-Aiden, E. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331, 176:182.

Miller, G. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39:41.

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, 7(3), 221:237.

Mishne, G. (2005, August). Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*.

Mueller, G.E. & Schumann, F. (1894). Experimentelle beitrÃge zur untersuchung des gedÃ¤chtnisses (Experimental contributions on the investigation of memory). *Zeitschrift fuer Psychologie*, 6, 81:190.

Mukherjee, A. & Liu, B. (2010) Improving Gender Classification of Blog Authors. *Proceedings of Conference on Empirical Methods in Natural Language Processing*. MIT, Massachusetts, USA.

Mukherjee, A., & Liu, B. (2010, October). Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.*

Murray, H. A. (1943). *Thematic Apperception Test* (Vol. 1). Cambridge: Harvard University Press.

Nakov, P. (2001). Latent semantic analysis for German literature investigation. In *Computational Intelligence*. *Theory and Applications,* 834:841.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3), 211:236.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, *29*(5), 665-675.

Nisbett, R. (2004). *The geography of thought: How Asians and Westerners think differently... and why.* Simonand and Schuster.

Oberlander, J., & Nowson, S. (2006, July). Whose thumb is it anyway?: Classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main conference poster sessions*.

Ogawa S., Lee T.M., Kay A.K. & Tank D.W., (1990) Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87, 9868:9872

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. New York: Bloomsbury Press.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296.

Popping, R. (2003). Knowledge graphs and network text analysis. *Social Science Information*, *42*(1), 91:106.

Pury, C. L. (2011). Automation can lead to confounds in text analysis Back, Küfner, and Egloff (2010) and the Not-So-Angry Americans. *Psychological science*, 22(6), 835:836.

Rorschach, H. (1921/1964) *Psychodiagnostik: A Diagnostic Test Based on Perception*, Berne, Switzerland.

Rude, S., Gortner, E. M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, *18*(8), 1121-1133.

Sagae, K., Gordon, A. S., Dehghani, M., Metke, M., Kim, J. S., Gimbel, S. I., Tipper, C., Kaplan, J., & Immordino-Yang, M. H. (2013). A data-driven approach for classification of subjectivity in personal narratives. In *Proceedings of the 2013 Workshop on Computational Models of Narrative*, OASIcs XX, Scholss Dagstuhl.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Agrawal, M., Park, G. J., ... & Lucas, R. E. (2013, June). Characterizing geographic variation in well-being using tweets. In *Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*.

Standifird, S. S. (2001). Reputation and e-commerce: eBay auctions and the asymmetrical impact of positive and negative ratings. *Journal of Management*, 27(3), 279:295.

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, *63*(4), 517-522.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). *The General Inquirer: A Computer Approach to Content Analysis*.

Strous, R. D., Koppel, M., Fine, J., Nachliel, S., Shaked, G., & Zivotofsky, A. Z. (2009). Automated characterization and identification of schizophrenia in writing. *The Journal of Nervous and Mental Disease*, 197(8), 585:588.

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.

Twenge, J. M., Campbell, W. K., & Gentile, B. (2012). Increases in Individualistic Words and Phrases in American Books, 1960–2008. *PloS one*, 7(7), e40181.

Vaillant, G. E. (2012). *Triumphs of Experience: The Men of the Harvard Grant Study*. Harvard University Press.

Van Dijk, T. A., & Kintsch, W. (1977). *Cognitive psychology and discourse: Recalling and summarizing stories*.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer ,New York

Venegas, R. (2012). Automatic Coherence Profile in Public Speeches of Three Latin American Heads-of-State. In the *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference*.

Vigouroux, R. (1879). Sur le role de la resistance electrique des tissuesdans le'electrodiagnostic. *Comptes Rendus Societe de Biologie* (Series 6),31, 336:339.

Villringer, A., Chance, B. (1997) Non-invasive optical spectroscopy and imaging of human brain function. *Trends Neuroscience*, 20 (1997), pp. 435–442

Weber, E. U., Hsee, C. K., & Sokolowska, J. (1998). What folklore tells us about risk and risk taking: Cross-cultural comparisons of American, German, and Chinese proverbs. *Organizational Behavior and Human Decision Processes*, 75(2), 170:186.

Williams, C., & D'Mello, S. (2010, January). Predicting student knowledge level from domain-independent function and content words. In *Intelligent Tutoring Systems* (pp. 62:71). Springer Berlin Heidelberg.

Wolff, P., Medin, D. L., & Pankratz, C. (1999). Evolution and devolution of folkbiological knowledge. *Cognition*, 73(2), 177:204.

Xu, X., Murray, T., Smith, D., & Woolf, B. P. (2013). If You Were Me and I Were You Mining Social Deliberation in Online Communication. *Proceedings of EDM-13, Educational Data Mining*.

Yarkoni, T. (2012). Psychoinformatics: New horizons at the interface of the psychological and computing sciences. *Current Directions in Psychological Science*, 21(6), 391:397.

Footnotes

[1] For historical background, see Stone, Dunphy & Smith, 1966; Graesser, McNamara, & Louwerse, 2004; and Tausczik & Pennebaker, 2010

[2] Available at http://www.liwc.net/

[3] There is also a pattern of masking negative language in public (Baddeley, Pennebaker, Beevers, 2013)

[4] One of the categories, loyalty to ingroup, initially showed an unpredicted pattern. However, this was reanalyzed using human coders, and consequently the new analysis confirmed the authors' hypothesis.

[5] For example, in Newman, Groom, Handelman, and Pennebaker (2008), 70% of the effect sizes which are significant at $p < .001$ will be considered small in terms of Cohen's classification.

[6] One particular type of such algorithms are called Support Vector Machines (Vapnik, 1995; Joachims, 1998). Two popular Support Vector Machine libraries are the following: LibSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) and SVMLight (http://svmlight.joachims.org/).

[7] Here we very roughly outline the general idea of such methods. For concrete descriptions of different methods, the reader should check Vapnik (1995) and Joachims (1998) for support vector machine, and Lewis (1998) and McCallum & Nigam (1998) for naive Bayesian classifiers.

[8] A useful source for more information and some applications are available at http://lsa.colorado.edu/

[9] Different software implementations of LDA can be found at http://www.cs.princeton.edu/~blei/topicmodeling.html

[10] More precisely, the model uses word frequencies per document, and number of topic as known variables, and approximates the posterior distribution of the hidden variables: topics given document, topics given document and words (see Blei, 2012 for a review).

[11] Available at https://framenet.icsi.berkeley.edu/fndrupal/

[12] Available at https://verbs.colorado.edu/propbank/

[13] Available at www.http://cohmetrix.memphis.edu

[14] Some of the LIWC's categories cover similar topics, including function words, word length, and tenses, so there is some conceptual overlap with Coh-Metrix.

[15] Software implementation is available here: http://www.casos.cs.cmu.edu/projects/automap/