Correction to: Scaling constructs with semantic networks

James A. Danowski¹ · Kenneth Riopelle²

Published online: 9 January 2020
© Springer Nature B.V. 2020

Abstract

This paper introduces a method for creating scales of constructs based on word bigram cooccurrences in natural language text. Instead of using a stop-word list to drop less useful words, we use a start-word list that enables computing the cooccurrences of only these “smart words.” In this way, we can create scales to measure communication constructs by first listing the key terms in the conceptual definition and then expanding the terms by looking up synonyms in dictionaries such as WordNet. Following this, we compute the cooccurrence network among these words with a sliding window. Next, we extract the first dimension through principal component analysis and identify the words that load most highly on the first factor. For these words, we sum the frequencies, which produces the final index for the construct. This operationalization yields index scales that have high construct validity, which contributes to external validity. Extending the procedures of classic psychometric index construction into the natural language domain avoids the biases of data based on fixed-choice questionnaires. To demonstrate the procedures for construct operationalization, we use a dataset of news stories about the BP Deepwater Horizon Gulf Oil Spill, scaling environmental uncertainty and organizational responses to it including innovation, strategic planning, and changes in organizational structure.

Keywords Sentiment analysis · Semantic networks · Network analysis · Environmental uncertainty · Teams · Strategic planning · Innovation · News content · Content analysis · Machine learning

Correction to: *Quality & Quantity* (2019) 53:2671–2683

<https://doi.org/10.1007/s11135-019-00879-5>

The original article can be found online at <https://doi.org/10.1007/s11135-019-00879-5>.

For reason beyond the control of the authors or the editors, the article titled “Scaling constructs with semantic networks” by James A. Danowski and Kenneth Riopelle (<https://doi.org/10.1007/s11135-019-00879-5>) was published in the regular issue Vol. 53 issue 5 instead of this special issue, where it was originally scheduled to appear. Therefore, the full article is reprinted here.

✉ James A. Danowski
jdandowski@gmail.com

¹ Department of Communication, University of Illinois at Chicago, MC 132, 1040 W. Harrison, Chicago, IL 60607, USA

² Department of Industrial and Systems Engineering, Wayne State University, Detroit, USA

1 Introduction

The main goal of this paper is to illustrate a method for constructing scales to measure constructs from natural language text based on psychometric methods (Kline 2015). We use a corpus of news documents associated with the BP Deepwater Horizon Gulf Oil Spill of 2010 that contained 56 million words. To enable the assessment of the predictive validity of the indices we created, our empirical example tests several hypotheses about organizational responses to environmental uncertainty (Duncan 1972).

Much of the contemporary research on machine learning is focused on classification, assigning documents to categories, such as positive or negative (Bishop 2006; Kotsiantis et al. 2007), a nominal level of measurement of the dependent variable. Sometimes more categories are involved, resulting in at best ordinal measurement. Some machine learning is regression based, such as the prediction of continuous values on a dependent variable (Rasmussen 2003). Our approach is in this category, using models for identifying a set of highly similar word tokens, and combining them to build an index of a construct. This operationalization is a ratio level of measurement. The higher level of measurement has benefits of greater specificity and sensitivity in detecting patterns.

2 Traditional scale construction

In psychometric science (Nunnally et al. 1967; Kline 2015), the standard way of constructing an index for a construct begins with the conceptual definition, writing a set of fixed-choice items in the form of questions to which respondents give quantitative estimates. In processing the data, the items are usually factor analyzed to extract the first dimension and the highest loading items on it. Typically, the scores on these items are summed to create the index. So, the result is a continuous scale with high internal validity.

Most psychometric research uses self-reported measures. These have biases that cannot be precisely estimated, including social desirability, demand characteristics, wording, and response rate biases, among other threats to validity (Fowler 2013). Moreover, it is expensive to collect these data. A better alternative may be to analyze large volumes of unstructured natural language texts from messages that people produce.

Our approach is a hybrid of traditional psychometric scale construction methods and what is called formative index construction, based on aggregate system-level data rather than individual self-reports (Diamantopoulos and Winklhofer 2001). Often, aggregate data are available from public or proprietary sources, which are the basis for creating indices to model processes of interest, rather than exclusively using self-reported data. Nevertheless, some such data may also be used, for example, aggregate consumer confidence ratings, in the case of modeling the macro-level economy. In social science, message content from mainstream media, social media, open-ended comments, or other document sources enables computation of formative indices of constructs.

3 Natural language index construction

As does traditional scaling, we begin with the construct to be operationalized, listing the key terms based on the conceptual definition. Rather than relying at this stage on face validity assessments common to traditional scale construction, we elaborate a set of terms that are linguistically highly similar for use in the natural language analysis. If a fixed-choice instrument for measuring the construct is available from prior studies, then the researcher would examine each item and list the words that it represents, in the term expansion process.

One could further expand these terms from either or both procedures through dictionary and thesaurus lookups. A widely used online dictionary of this sort is WordNet (Miller 1995), a database of linguistic relationships. The synonyms it provides are useful in expanding the starter terms for natural language index construction.

The terms expanded through a collection of synonyms, unfortunately, are not accompanied by any metrics that quantify their degree of similarity. In actual language, we expect some of these candidate synonyms to be highly related or even interchangeable, as alternative words may be similarly embedded in textual contexts, but we do not know to what extent. We need to look at how people use these words as they produce messages.

Once computation generates the list of aggregate word frequencies, we have a matrix whose number of rows and columns are the number of unique words in the corpus. The high dimensionality that occurs with large corpora, where the vocabulary may range from 10,000 to 30,000 words or more, is troublesome for computational purposes due to memory and processing requirements. The goal is to reduce this dimensionality.

In our case of semantic index construction, we want to do an extreme dimensionality reduction, focusing on a single dimension that best represents the construct. The technique for this is principal component analysis, in the family of procedures known as factor analysis. The first factor is the linear dimension that explains the most variance among the words. We examine this eigenvector to see the loadings of the words on it. In traditional scaling, one would typically take the items with loadings above .60 or .70 and sum them to create the scale to measure the construct. We use this standard.

4 Choosing units of analysis

In traditional instrument construction, the unit of analysis is typically the same as the unit of observation, the individual. One typically analyzes data from self-reports on fixed-choice items from a sample of respondents. In text processing, the unit of analysis may be similarly straightforward, as when the documents produced by an individual, for example, social media posts, are the unit of analysis. In other cases, choosing units of analysis may be problematic. For news corpora, for example, the research shows that government actors' initiatives and information typically reciprocally set the agenda for the elite media, the *New York Times*, which mainstream news outlets then copy (McCombs and Shaw 1972; Golan 2006). This imitation process reduces the variance across news outlets. Accordingly, it makes partitioning a corpus by news outlet less effective in analyzing variations among words. Rather, a different way of slicing the data to produce normal variation is desirable.

One solution is to segment the corpus over time intervals and then compare the distributions of words across the time slices. This parsing enables using statistical testing

procedures, whether of differences or associations. Time slicing normalizes data across documents and sources. In choosing the time slice interval, such as hours, days, weeks, and months, the researcher needs to consider whether the volume of text available across the intervals is enough to do a valid word analysis.

Guidance comes from Zipf (1949), who concluded after extensive investigations that 10,000-word minimums per observation are necessary for valid results. At the same time, one needs to select a time interval that has at least 30 slices as observations, the level at which means are generally relatively stable.

Our procedures are a combination of supervised and unsupervised methods. In the unsupervised mode, we began by familiarizing ourselves with the data, by aggregating the text into a single file, identifying all word pairs appearing within three words on either side of each word in the text, and network analyzing them and examining numerical results and visualizations. For the supervised phase, we chose terms to operationalize a construct, expanded the terms with synonyms, and then computed the frequencies of these terms for further processing.

5 Smart words

To do semantic network scaling, we began with “start words” as opposed to “stop words.” The latter are usually dropped from text in preprocessing, because common function words like prepositions and articles appear widely with the range of other words, and therefore render networks that look like hairballs or cooked spaghetti, that are also computationally challenged. Because these start words are created based on conceptual definitions for constructs, we could consider them “smart words,” in which they are chosen based on theoretical reasoning.

One could use a variety of sources for producing smart start words. A first step would consider the conceptual definition for the construct and generate a list of key terms from it. If there were fixed-choice instruments available, one would also extract terms from the questionnaire items. To augment term expansion, the investigator could look up each term in a dictionary or thesaurus and add these to the list. We did this in the current research using the API for the online WordNet lexical database (Miller 1995), extracting its synonyms for each word.

Another source one could use is from prior text mining research that may have produced special lexicons for a concept. For example, we used a list of words reported by Loughran and McDonald (2016), who analyzed natural business language from annual and quarterly financial reports. One of their lists was for “uncertainty.” It contained 275 words they identified in their clustering analysis. Another of their lists also related to uncertainty includes modal verbs: “may, should, depending, would, and probable.” We used these as a second start list to index uncertainty.

If the researcher is dealing with an unfamiliar content domain, subject matter experts could be asked to list or review key terms. Another approach would be to do a semantic network analysis of published literature about the concept in scholarly journals and books and extract key terms from the analysis (Cantoni and Danowski 2015). Because the rest of the process derives from the start words used, it is useful to be comprehensive, detailed, and consider all likely morphological forms of terms in preparing the list for extraction of word frequencies. Stemming the words to root forms or reducing morphological forms through lemmatization is not advised for this kind of analysis, because it dampens and

restricts variance in actual word usage, which would reduce the validity and reliability of the resulting scales produced.

6 Semantic network preprocessing

Before using the start words, we preprocessed the corpus, removing duplicate documents, stop words, and words and word pairs occurring less than three times, selecting these options in the semantic network package WORDij (Danowski 1993b, 2013). We did no stemming or lemmatization, to increase linguistic validity. Punctuation elements were dropped, except sentence-ending periods, question marks, and exclamation points. A sliding window, three words wide, tabulated word pair cooccurrences around each word. Windowing stopped at the end of a sentence and resumed at the beginning of the next one. As the window moved through the text, centering on each word, the program tabulated word pairs appearing in it. The order of words in pairs was retained, based on which word appeared before or after another word. This process generated the aggregate matrix in the form of an adjacency list of directed word pairs and their frequencies of occurrence. In addition to these proximate bigrams, the software produces a matrix of word frequencies by observations. To familiarize ourselves with the network of words, we analyzed the bigrams by clustering them with the Clauset–Newman–Moore group identification algorithm in NodeXL (Smith et al. 2010) and then produced a visualization which we examined.

7 Start-word include list

Consider that natural language analysis typically removes stop words, which are function words with little substantive meaning such as “the, of, and, a, to, etc.” The remaining words are analyzed. In contrast, when constructing scales for constructs, we can think of using the opposite of a stop-word list—a start-word or include list. Instead of dropping common function words and analyzing the rest, we want to drop all but the start words, and networks analyze only their occurrences.

The WORDij program has this functionality, called an “include list,” as opposed to its option of using a “drop list” of stop words. With an include list of start words, we remove everything but these words, and the sliding window then moves through the words that remain, tabulating their cooccurrence frequencies and individual word frequencies. We then factored the latter to yield the principal eigenvector, whose highest loading words’ frequencies we then summed to produce the index.

An alternative would be to do cluster analysis to locate groups of words with high similarity within and low similarity between groups. Then, the most central words would be identified to interpret the meaning of each group. One could locate the group within which the key terms appear and examine the centrality of words within it, representing their weights (Danowski and Park 2014). Although this network linkage approach is likely to yield similar results to the distance-based modeling of similarities among word vectors, it is not as extensible into machine learning algorithms. On the other hand, if the goal were to identify topics in the corpus, it would be possible to do so with the linkage-based group identification of clusters of words with a high degree of shared links.

One may view this process as similar to topic modeling (Blei et al. 2003). Topic modeling parallels traditional approaches such as factor analysis in its reducing of the

dimensionality of a matrix. Once specifying the desired number of topics, the researcher examines the top-loading words on a topic vector and attempts to assign a label to the topic. The results from topic modeling using linear Dirichlet analysis (LDA) of word distributions are generally rather crude, although the software packages for it have an advantage over earlier methods such as latent semantic analysis (LSA) (Deerwester et al. 1990), which treats each major eigenvector as a separate topic. Because LDA topic modeling involves arbitrarily setting some reduced number of dimensions, such as 100–300, it lowers computation costs and enables handling of much larger data sets, although with losses in internal and external validity (Danowski 2011), due to the hit or miss nature of choosing the number of topics in advance.

More generally, distance-based vector models applied to text sometimes have serious violations of the triangular inequality relation that the sum of the lengths of any two sides of a triangle is greater than the length of the third side. There is a similar notion in network analysis called closure (Holland and Leinhardt 1976). If A and B are linked and A and C are linked, then the probability that A and C are linked is higher. When all three links exist, this is closure. Nevertheless, many triads in a network are not completely connected, which is the basis for hierarchy (Wasserman and Faust 1994).

In semantic networks, this lack of closure can represent anomalous relations due to differences in levels of abstraction across word networks, in which some levels differentially subsume other levels below them. A second factor is polysemy, where words have different senses depending on context, and these are jumbled in the overall network. Third, the corpus may contain heterogeneous domains. Nevertheless, linkage-based network analysis appears to be more robust in the face of these problems than linear distance models, which can have high levels of violations of the triangular inequality rule, resulting in large negative eigenvalues, a warping of the Euclidean space (Woelfel 1993). For semantic network scaling, however, the distance model is more compatible with index construction goals, which use linear statistical procedures within the bounds of Euclidian space.

This method is equivalent to identifying word embeddings and locating the most similar words to a seed word, as is done in programs packaged in Python, such as word2vec (Mikolov et al. 2013), GloVe (Pennington et al. 2014), and gensim (Rehurek and Sojka 2011). These programs also use the sliding window method we do, as they extract words with similar vectors. Such packages could be used as an alternative to the software we used for this paper. For preprocessing, packages such as NLTK (Perkins 2010) or spaCy (Honnibal and Johnson 2015) could be used.

8 Identifying the first principal component

Once we computed the occurrences of start words, we ran the dimension reduction procedure of principal component analysis to factor the matrix to extract the first linear eigenvector, which explains the most variance among the words. The words that load on this dimension are ones whose vectors lie more fully on it, having relatively small cosines. Because any one of the high-loading words represents the underlying eigenvector, any set of these words above a strength threshold could be used to create multiple indicators of the dimension, but the top-loading words capture the most related, as well as interchangeable words.

The rule we used was to begin at the top with the highest loading word and examine the words ranked below it, selecting those above the weight threshold and excluding those that do not appear to represent the conceptual definition of the construct. The appearance

of such words on the primary dimension may have resulted from differences in abstraction or polysemy. Because there are multiple equivalent indicators of the primary dimension, dropping some of them for reasons of face validity introduces no bias.

9 Empirical example

In choosing the corpus to use in demonstrating the procedures, we decided on news documents as the primary type of text. We had from an earlier study (Danowski 2012), a corpus of news collected associated with the Deepwater Horizon Gulf Oil Spill of 2010, for which BP was found responsible. To have sufficient data for time slicing to create units of analysis, we expanded the size of the original corpus by capturing news documents from a longer period before and after the spill. Because the federal judicial processes did not reach a summary judgment until 2014, we decided to extend the time frame of data collection to encompass possible longer-term effects of the litigation. So, we captured 8 years of news documents after the spill from LexisNexis Academic, using the terms “BP” and “public” to retrieve documents from its collection of “major world publications” in English. For comparability, we then captured documents during the 8 years preceding the spill. The corpus ranged from documents available from 2002 to 2018.

In choosing a construct to scale, we considered that the magnitude of the spill probably produced considerable environmental uncertainty for BP, given the lengthy course of litigation as to the spill’s causes, effects, damages, fines, restitution, and regulatory changes. Accordingly, we decided to use the construct of environmental uncertainty (Duncan 1972) as our primary variable for index construction. Information theory (Shannon and Weaver 1949) conceptualizes information as the reduction of uncertainty. Accordingly, we hypothesized that there would be an increase in the amount of information in the media about BP from before to after the spill.

To provide more than one example of scale construction, we decided to scale possible organizational responses to environmental uncertainty. Also, to use predictive validity to assess the efficacy of our index constructions, as well as to add substantive value to our methodological focus, we tested a set of hypotheses linking increased environmental uncertainty to changes in the internal BP organization, as it may have attempted to reduce some of this uncertainty by strategic planning, internal structural changes, and innovation. Nevertheless, because our primary goal is methodological, we will not extensively explicate the theoretical aspects here.

We hypothesized the following:

H1 As environmental uncertainty increases, the volume of communication increases.

H2 As environmental uncertainty increases, internal communication becomes more integrated.

H3 As environmental uncertainty increases, strategic planning increases.

H4 As environmental uncertainty increases, innovation increases.

Hypothesis 1 As environmental uncertainty increases, the volume of communication increases.

To test H1, we measured the volume of news story content and found that compared to 8 years Before, there was a 40 percent increase in volume 8 years After. Before there were 283.58 MB of text. Stories numbered 20,103, containing 22.5 Million words. After there were 396.5 MB of text containing 64,986 stories and 31 Million words. Each of the Chi-square tests was significant at $p < .00001$. These results supported our hypothesis that an expected increase in environmental uncertainty after the spill would correlate with an increase in the volume of information.

As an initial analysis, we used an include list containing the words: “uncertainty” and “uncertainties,” run on the Before and After texts, which found that both terms nearly doubled in normalized frequencies, as seen in Table 1. By normalized, we mean that we divided the observed frequency value by the total frequencies across all words in the corpus. This standardization of values converted them into a proportion, akin to a probability, ranging from 0 to 1. In some machine learning procedures, normalization has a different meaning, the removal of text case, punctuation, stop words, and lemmatization or stemming of the words to reduce their morphological variation to lower computational memory requirements or processing time.

To prepare an index of uncertainty, we factor analyzed 255 words in the Loughran and McDonald (2016) “uncertainty” list and extracted the first factor for words loading at .70 and above. There were 32 such words. Highest loadings were for “possible, could, depending;” “uncertain” also loaded highly, shown in Table 2. We then ran an include list on the two 96-month series with the 32 words for Before and After. The normalized mean for Before was 33.75 occurrences per month and for After was 81.65 ($t = 2.68$, $df = 173$, $p < .004$).

The fourth test we ran on the 8-year Before/After corpora was an include list of modal words that qualify verbs (Loughran and McDonald 2016). Words indicating contingency statements include: “may, should, depending, would, and probable” (Table 3). The first three significantly increased from Before to After at $p < .05$ and the last two at $p < .10$.

The findings provide support for our hypothesis that the Spill would accompany increased environmental uncertainty for BP. We tested this in three ways: (1) an increase in the words uncertainty and uncertainties in the news content, (2) an increase in the composite uncertainty index, and (3) an increase in contingency words. The finding of increased environmental uncertainty from Before to After led us to test the second hypothesis that environmental uncertainty would link with changes in the internal structure of the organization.

Hypothesis 2 As environmental uncertainty increases, internal communication becomes more integrated.

A key form of internal integration as an organization attempts reducing environmental uncertainty (Duncan 1972; Lawrence and Lorsch 1986) is to create teams from across the organization with a special mix of talent to address a problem. Teaming shifts the structure from a command and control form to a matrix organization (Davis and Lawrence 1978).

Table 1 Uncertainty before and after the Gulf Oil Spill

Word	Frequency		Proportion		Z value
Uncertainty	1630	3792	0.005436	0.009069	−17.535845
Uncertainties	274	1376	0.000914	0.003291	−20.74454

Z values are significant at $p < .00001$

Table 2 Uncertainty words factor

Word	Loading
Possible	.911
Could	.879
Depending	.865
Assume	.842
Suggesting	.839
Risk	.828
Possibility	.821
Varying	.818
Suggest	.817
May	.810
Roughly	.809
Vary	.803
Appear	.800
Differ	.794
Depends	.792
Suggests	.790
Appears	.782
Depend	.771
Dependent	.765
Variations	.758
Approximately	.758
Unclear	.754
Varied	.749
Assumption	.745
Likelihood	.742
Assuming	.712
Vague	.710
Confusion	.708
Might	.708
Believe	.706
Assumes	.704
Risks	.702
Uncertain	.697

Factor 1 of principal component analysis with varimax rotation

Table 3 Modal contingency verb increases from before to after the Gulf Oil Spill

Word	Frequency		Proportion		Z value
	Before	After	Before	After	
May	21696	84,621	0.080825	0.10675	−38.665436*
Should	14,887	54,905	0.055459	0.06926	−24.93761*
Depending	384	1967	0.001431	0.00248	−10.008614*
Would	39,964	124,015	0.148879	0.15645	−9.374719**
Probable	178	807	0.000663	0.00102	−5.219133**

* $p < .05$; ** $p < .10$

Table 4 Team factor loadings

Word	Loading
Teams	.812
Team	.809
Together	.637
Join	.622
Teamed	.562

Factor 1 of principal component analysis with varimax rotation

We took thesaurus entries for “team,” resulting in 33 words. We ran these as an include list on the After corpora, slicing the period into months ($n=96$), and ran a factor analysis on the matrix of these words by these periods. Table 4 shows that “teams, team, together, join, and teamed” loaded above .50 on a factor. Although this threshold was below the standard of .70, we wished to examine the effects of modifying the protocol. We, therefore, used these five terms as a start-word set. The Z-test comparing relative frequencies for the 8 years Before and After found that only the word “together” increased significantly ($Z=6.89$, $p<.0001$). This significance for only one team word is weak evidence.

Hypothesis 3 As environmental uncertainty increases, strategic planning increases.

To index strategic planning, we started with the words “plan, planning, and strategy” and extracted synonyms from WordNet (Miller 1995). This concept expansion produced a list of 17 words. We factor analyzed these and took the words loading above .70 on the first factor: “project, plan, projects, provision, program, and preparation,” shown in Table 5. We then computed the averages of the monthly frequencies for Before and After. The planning factor was significantly higher in the After period with a mean of 428.13 occurrences per month than in the Before corpus’ 38.73 ($t=24.18$, $df=173$, $p<.0001$). Both tests supported the hypothesis that increased environmental uncertainty would be associated with more strategic planning, presumably to reduce uncertainty.

Hypothesis 4 As environmental uncertainty increases, innovation increases.

The fourth hypothesis concerned the innovation response to increased environmental uncertainty. We repeated our index building procedure to test whether innovation may have increased. We took the word “innovation” and extracted 33 synonyms from WordNet (Miller 1995). An include list of these start words was run with the After corpus. The loadings of the words on the largest factor are shown in Table 6. To further refine the terms, we then ran a new include list with the pared-down list of innovation words: “create, creation,

Table 5 Planning words factor

Word	Loading
Project	.899
Plan	.828
Projects	.758
Provision	.708
Program	.706

Factor 1 of principal component analysis with varimax rotation

Table 6 Innovation words factor

Word	Loading
Create	.849
Created	.783
Creation	.770
New	.749
Innovative	.739
Advanced	.668
Innovations	.602

Factor 1 of principal component analysis with varimax rotation

creations, new, innovative” for the monthly Before and After slices. The T-test for the normalized means per month, 118.5 in the Before period and 207.83 in the After ($t=1.74$, $df=175$, $p<.041$) supported the hypothesis that an increase in environmental uncertainty accompanies an increase in innovation.

10 Discussion

This paper has demonstrated how to construct predictively valid scales for constructs using natural language processing techniques guided by psychometric principles. A large corpus of texts from news documents was our test dataset. Our combination of unsupervised and supervised analysis extended the principles of traditional psychometric index construction into the contemporary data science domain. This move increases accuracy, precision, and reliability of measurement, as well as external validity. Several key biases from self-reported data are removed, including social desirability, response rate, instrumentation effects, and question-wording, each of which threatens internal and external validity. Our results have shown that a third kind—predictive validity—is evident from the hypothesis tests linking environmental uncertainty to organizational responses of innovation and strategic planning.

A key feature of our approach to natural language processing is that instead of removing stop words in preprocessing the corpus, we removed all words except our start words derived from construct explication. Our use of an include list of start words inverted the unsupervised semantic network method and turned it into a supervised one. This start-word usage accents the theory-based intelligence of the method, perhaps considering this a “smart word” approach. It enables more theory-driven modeling of the data, in this case, to produce scales for constructs. It is more confirmatory than exploratory, adding more intelligence to the method.

We used two sources for our start-word term expansion. One was the WordNet dictionary; another was the results of prior natural language identification of words measuring uncertainty in financial reports. These provided us with words to use in creating the start list as we expanded the conceptual definition of the construct for semantic network analysis. Both term expansion methods proved useful.

The approach of analyzing word vectors has a solid basis in data science. The developments are coming rapidly at this stage, such that there are some alternative algorithms analysts use for the stages of computation. When researchers initially examine this arena, the

highly technical way data scientists report their evaluation of the alternatives can obscure the higher-level frameworks within which these methods develop. This paper has sought to provide a clear picture of these perspectives and reduce the range of choices based on theoretical and methodological principles.

Our approach has shown how natural language data can be treated with the procedures applied to operationalize theoretical constructs so that not only is validity higher, but research costs are considerably lower by an order of magnitude or more. We demonstrate how the affordances of contemporary computational communication science enable comprehensive hypothesis testing at a relatively low cost of data collection compared to traditional methods relying on questionnaires.

Replication of findings by other researchers is easier, as well. Furthermore, the tool kits coded into open-source packages in R, Python, or other programming languages, continue to expand in functionality and ease of use. These could compute word embeddings from the word cooccurrence matrix and identify highly related or interchangeable words through vector similarity. The advantage of the procedures we illustrated is that researchers do not need to program in Python or other languages to accomplish the same ends. One could use WORDij to produce the start word frequencies, then SPSS or other statistical packages to factor the matrix and identify the highly-weighted words.

The ease with which researchers can use packages enables simulations to assess the effects of alternative settings of hyper-parameters for statistical procedures. Researchers can consider our results as they empirically explore issues such as the window size for cooccurrence matrix formation, which principal component analysis algorithms and parameters to use, what thresholds for word vector similarity to implement, or what corpus sizes are optimal. Our approach should be replicated to see whether the design choices we made are generalizable.

Extending psychometric methods into the natural language domain enables scale construction that avoids the biases and costs of self-reported measures. There are additional benefits that emerge from the accelerating advancement of machine learning methods. Communication and other social research can increasingly take advantage of bridging the knowledge bases of social science theory and methods with those of computer science, as exemplified in the emerging area of computational communication research. This process produces a platform for hypothesis testing in which the increasing intelligence embedded in algorithms can be further leveraged to augment natural forms. In future research, the approach we illustrated can be advanced with deep learning, building from automatic scaling of constructs toward identifying patterns among them.

References

- Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
- Blei, D.M., Ng, A.Y., Jordan, M.I., Lafferty, J.: Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003). <https://doi.org/10.1162/jmlr.2003.3.4-5.993>
- Cantoni, L., Danowski, J.A.: Communication and technology—an itinerary. In: Cantoni, L., Danowski, J.A. (Eds.) *Communication and Technology*, vol. 5. Walter de Gruyter GmbH & Co KG, Berlin (2015)
- Danowski, J.A.: Network analysis of message content. In: Barnett, G., Richards, W. (eds.) *Progress in Communication Sciences XII*, pp. 197–222. Ablex, Norwood (1993a)
- Danowski, J.A.: Counterterrorism mining for individuals semantically-similar to watch list members. In: Wiil, U.K. (ed.) *Counterterrorism and Open-Source Intelligence: Models, Tools, Techniques, and Case Studies*. *Lecture Notes in Social Networks*, 1–6, 2nd edn, pp. 223–247. Springer, Berlin (2011). https://doi.org/10.1007/978-3-7091-0388-3_1

- Danowski, J.A.: Mining social networks at the organizational level. In: Ting, I.-H., Hong, T.-P., Wang, L.S.L. (eds.) *Social Network Mining, Analysis and Research Trends: Techniques and Applications*, pp. 205–230. IGI-Global, Hershey (2012)
- Danowski, J.A.: WORDij Version 3.0: Semantic Network Analysis Software [Computer Program]. The University of Illinois at Chicago, Chicago (2013). Retrieved from <http://wordij.net>. Accessed 31 Mar 2019
- Danowski, J.A.: WORDIJ: a word pair approach to information retrieval. In: *Proceedings of the DARPA/NIST TREC Conference*, pp. 131–136. National Institute of Standards and Technology, Washington (1993b)
- Danowski, J.A., Park, H.W.: Arab Spring effects on meanings for Islamist web terms and on web hyperlink networks among Muslim-majority nations: a naturalistic field experiment. *J. Contemp. East. Asia* **13**(2), 15–39 (2014)
- Davis, S.M., Lawrence, P.R.: Problems of matrix organizations. *Harv. Bus. Rev.* **56**(3), 131–142 (1978)
- Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **41**(6), 391–407 (1990)
- Diamantopoulos, A., Winklhofer, H.M.: Index construction with formative indicators: an alternative to scale development. *J. Mark. Res.* **38**(2), 269–277 (2001)
- Duncan, R.B.: Characteristics of organizational environments and perceived environmental uncertainty, pp. 313–327. *Adm. Sci. Q.*, SEPT (1972)
- Fowler Jr., F.J.: *Survey Research Methods*. Sage Publications, Thousand Oaks (2013)
- Golan, G.: Inter-media agenda setting and global news coverage: assessing the influence of the New York Times on three network television evening news programs. *Journal. Stud.* **7**(2), 323–333 (2006)
- Holland, P.W., Leinhardt, S.: Local structure in social networks. *Sociol. Methodol.* **7**, 1–45 (1976)
- Honnibal, M., Johnson, M.: An Improved non-monotonic transition system for dependency parsing. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal. Association for Computational Linguistics (2015)
- Kline, P.: *A Handbook of Test Construction (Psychology Revivals): Introduction to Psychometric Design*. Routledge, New York (2015)
- Kotsiantis, S.B., Zaharakis, I., Pintelas, P.: Supervised machine learning: a review of classification techniques. *Emerg. Artif. Intell. Appl. Comput. Eng.* **160**, 3–24 (2007)
- Lawrence, P.R., Lorsch, J.W.: *Organization and Environment: Managing Differentiation and Integration*. Harvard Business School Classics, Boston (1986)
- Loughran, T., McDonald, B.: Textual analysis in accounting and finance: a survey. *J. Account. Res.* **54**(4), 1187–1230 (2016)
- McCombs, M.E., Shaw, D.L.: The agenda-setting function of mass media. *Public Opin. Q.* **36**(2), 176–187 (1972)
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems*, pp. 3111–3119 (2013)
- Miller, G.A.: WordNet: a lexical database for English. *Commun. ACM* **38**(11), 39–41 (1995)
- Nunnally, J.C., Bernstein, I.H., Berge, J.M.T.: *Psychometric Theory*, vol. 226. McGraw-Hill, New York (1967)
- Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543 (2014)
- Perkins, J.: *Python Text Processing with NLTK 2.0 Cookbook*. Packt Publishing, Birmingham (2010)
- Rasmussen, C.E.: Gaussian processes in machine learning. In: Carbonell, J.G., Siekmann, J. (eds.) *Summer School on Machine Learning*, pp. 63–71. Springer, Berlin (2003)
- Rehurek, R., Sojka, P.: Gensim–Python framework for vector space modelling, vol. 3(2). NLP Centre, Faculty of Informatics, Masaryk University, Brno (2011)
- Shannon, C.E., Weaver, W.: Ch 6: Shannon Entropy. In: *The Mathematical Theory of Communication*, pp. 51–58 (1949)
- Smith, M., Ceni A., Milic-Frayling, N., Shneiderman, B., Mendes Rodrigues, E., Leskovec, J., Dunne, C.: NodeXL: a free and open network overview, discovery and exploration add-in for Excel 2007/2010/2013/2016, from the Social Media Research Foundation (2010). <https://www.smrfoundation.org>. Accessed 14 Jan 2019
- Wasserman, S., Faust, K.: *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge (1994)
- Woelfel, J. GALILEO (Version 5.6) [Computer software]. The Galileo Company, Amherst, NY (1993)
- Zipf, G.K.: *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge (1949)