Journal of Language and Social Psychology
2015, Vol. 34(2) 200–212

DOI: 10.1177/0261927X14546568
jls.sagepub.com
SAGE

Exploring Political Ideologies of Senators With Semantic Analysis Tools: Further Validation of CASS

Nicholas S. Holtzman[1], Simon Kwong[2], and Kirsten L. Baird[1]

Abstract

Phrase counting is an effective approach to capturing individual differences in language use. Specific phrases (e.g., "war on terrorism") powerfully predict a congressperson's political ideology. The question addressed in this study is whether there is additional information that can be extracted from the indirect relations among words in large Senatorial speech databases. Given that direct co-occurrence of target words is a very low likelihood event, we focus on higher-order co-occurrence (e.g., whether two target words appear in similar semantic contexts), using the free software Contrast Analysis of Semantic Similarity (CASS) to compute individual differences (www.casstools.org). We describe how we used CASS in detail and provide a tutorial. Using text transcripts from 86 Senators, totaling over 150 million words, we demonstrate that CASS can account for political ideology above and beyond phrase counts. By complementing phrase count methods, CASS may be a useful method for the digital humanities and social sciences more generally.

Keywords

CASS, co-occurrence, political ideology, semantic space models

Text data are accumulating quickly, leading to remarkable opportunities for text analysis. What should be the focus of text-analytic processing? One group of researchers in the social psychology of language have focused on word-counting (Matsumoto &

[1]Georgia Southern University, Statesboro, GA, USA
[2]New York, NY, USA

Corresponding Author:
Nicholas S. Holtzman, Department of Psychology, Georgia Southern University, PO Box 8041, Statesboro, GA 30469-8041, USA.
Email: nick.holtzman@gmail.com

Hwang, 2013; Tausczik & Pennebaker, 2010), as word-counts are rich in psychological information (for review, see Ireland & Mehl, 2014). Other researchers have used phrase-counting (Fetzer, 2008; Gentzkow & Shapiro, 2010) and similar approaches (Fast & Funder, 2008; Mehl, Gosling, & Pennebaker, 2006; Pennebaker, Francis, & Booth, 2001; Yarkoni, 2010). Indeed, there have been some very interesting observations in these literatures. For example, researchers have identified words and phrases that indicate social ideologies and social movements (Matsumoto & Hwang, 2013). However, language carries information beyond these simple types of word counts and phrase counts, and accordingly word counts and phrase counts have at times been called into question (e.g., by discourse analysts). One reason to question such counts is that the ways in which words are associated with one another is an especially rich aspect of language—one that is yet to be fully analyzed. Thus, one rationale for using an approach that explores how words are associated with one another is that such an approach may capture important outcomes that word counts and phrase counts cannot capture.

There is a growing literature emphasizing the utility of associations among words in large databases. One recent addition to this literature is the Contrast Analysis of Semantic Similarity (CASS). CASS is grounded on semantic models that focus on co-occurrences of words in sentences; it was inspired by Latent Semantic Analysis (Landauer & Dumais, 1997) and the Bound Encoding of the Aggregate Language Environment (Jones & Mewhort, 2007), and CASS shares features with the Hyperspace Analogue to Language (Lund & Burgess, 1996).

In a minimalist analysis, to reveal semantic similarity of two words, researchers would simply observe the direct co-occurrences of those words in sentences. However, direct co-occurrences of target words are unlikely to frequently co-occur. For this reason, and other reasons explained below, CASS focuses on higher-order associations among words (e.g., second-order associations) to reveal semantic similarity. As an example of a second-order association, CASS accounts for whether words like "conservative" and "good" (or other user-specified words) tend to have similar associations with other words (e.g., names of Republican politicians), which can reveal whether the speaker uses the words "conservative" and "good" in similar ways. Such a pattern would indicate the semantic similarity of words in a personal lexicon. This type of information can be captured even if "conservative" and "good" never directly co-occur. Thus, CASS arguably captures a residue of meaning in language that phrase counting methods may fail to capture. After capturing these higher-order associations from text, CASS can capture individual difference scores by contrasting the semantic similarities.

There is already some evidence that CASS captures meaning in large databases (Holtzman, Schott, Jones, Balota, & Yarkoni, 2011). For example, previous research linked CASS effects to media slant in expected ways (Fox News is more conservative than CNN, which in turn is more conservative than MSNBC). The way in which slant was measured (using CASS) was based on whether certain political words (e.g., "conservative") and valenced words (e.g., "good") had similar kinds of contexts. If the context for the word "conservative" is similar to the context for the word "good," then this is one sign that the media is slanted in the conservative direction. Additionally, this study revealed evidence that CASS could be linked to individual-level

ideological slant (in the newscasters Hannity and Colmes). CASS researchers are just beginning to gather evidence of its validity; it would be timely to demonstrate that CASS can capture individual differences on a large scale (i.e., analyzing many individual speakers).

To this end, the current study explores ideological slant in 86 U.S. Senators who provided a sizable amount of speech data between 1994 and 2010. However, before turning to the results, we will briefly describe how we used CASS for this study.

CASS: Technical Notes

Target by Context Word Co-Occurrence Matrix

The CASS is a semantic model based on the co-occurrences of words in sentences; each cell in the initial co-occurrence matrix represents the co-occurrence of a target word with a context word in sentences. See Appendix A for an example. Four *target words* are specified in CASS. In a study of political slant, these words might be "conservative," "liberal," "positive," and "negative." Then, CASS computes the co-occurrence of these words (i.e., the row labels in a matrix of co-occurrence data) with a set of other words that appear in the document (i.e., the column labels in a matrix of co-occurrence data); the latter are called the *context words*. A set of context words may be specified by the user or may be automatically extracted from the text. The unit of analysis is a sentence. Through this process, CASS computes a target by context word matrix of co-occurrence data. An example is provided in Appendix A.

Expected Co-Occurrence Matrix

Because individual cells may contain very large values simply due to high rates of both the target word and the context word in the document and their consequential chance-based co-occurrence, it is important to control for their mutual rates of occurrence (Church & Hanks, 1990; Recchia & Jones, 2009). If one fails to control for these base rates, then common context words would create outliers that would contaminate subsequent analytic steps. To control for these rates, CASS first calculates the expected rate of co-occurrence for each cell in the Target X context word matrix. The expected rate of co-occurrence equals the multiple of their individual likelihoods of occurrence in a given sentence, based on the full text. For example, if the likelihood of the target word is .10 and the likelihood of the context word is .20, then the expected rate of co-occurrence would be .02.

Actual Over Expected Co-Occurrence Matrix

The actual rate of co-occurrence is then divided by the expected rate of co-occurrence, in a cell by cell fashion, to arrive at the actual/expected co-occurrence matrix. The cell entries in this matrix indicate the following: given the occurrence rates in a certain text file, they represent the co-occurrence of terms relative to expected levels of

co-occurrence. A value less than 1.00 signifies that the words are co-occurring less than expected; a value of 1.00 indicates that the actual co-occurrence rate between the target word and the context word is exactly as expected (e.g., it was expected that there would be 2 co-occurrences out of 10 and it turned out that indeed there were 2 co-occurrences out of 10); a value greater than 1.00 indicates that the words are co-occurring at greater than expected levels. Because we did not want large cell values to affect the analyses that follow (as this would overweight the impact of certain context words), we capped the possible magnitudes in the actual/expected matrix at 2.00 (a value that maintains symmetry around 1.00, the expected rate).

In order to arrive at accurate contrasts of semantic similarity, the CASS software was modified to calculate the Fisher *r*-to-*z* transforms for the correlations of pairs of target word rows in the actual/expected matrix, after capping, and then it contrasts the Fisher *r*-to-*z* values:

$$[Z_{r(\text{conservative,positive})} - Z_{r(\text{conservative, negative})}] - [Z_{r(\text{liberal,positive})} - Z_{r(\text{liberal,negative})}]$$

A positive score would indicate conservative slant, while a negative score would indicate liberal slant. See the example in Appendix A for more details.

Method

Transcripts

The text files of the speech transcripts from a set of 197 Senators were downloaded from the Government Printing Office's digital collection of the Congressional Record (GPO, 2012). We downloaded every speech by every Senator between the years 1994 and 2010. To increase the stability of the target by context word matrix (i.e., reducing randomness in the matrix), the data were collapsed into one file per Senator. Of the 197 Senators who spoke in Congress during this period, 86 provided the minimum prespecified instances of the target words, 100 each; thus, we used the files for these 86 senators (42 were Democrats; 41 were Republicans; 2 were Independents; 1 switched from Republican to Democrat [Arlen Specter]). Using files that contain fewer instances of the target words tends to create an unstable target by context word co-occurrence matrix and thus some cutoff (such as 100) is required. Alternatively, if we had set the criterion to be higher, say 250, then the sample of senators would have been too small to explore individual differences. The 86 text files ranged from 1.90 to 50.36 megabytes (M = 10.66 megabytes). The files ranged from 313,691 words to 8,217,057 words (M = 1,801,110 words). In all, we analyzed over 150 million words.

CASS Specifications

There are some variables pertaining to the target words and context words that the user must set prior to running a CASS analysis.

Target Words. To fully sample the semantic space for the target concepts (conservative, liberal, positive words, negative words), we generated synonyms for these four targets. We have listed the synonyms in Appendix B. All the synonyms for the target words were replaced with the key target word prior to any CASS analysis. For example, the word "Democrat" (a synonym of "liberal") was converted to the target word "liberal" prior to CASS analysis. It is worth noting that it is generally wise to include synonyms of target words in the analysis. Failure to include enough synonyms will necessarily yield a more sparse co-occurrence matrix, which will diminish the reliability of the analysis. In contrast, including more target word synonyms tends to help fill-out the co-occurrence matrix and thus yield more reliable findings.

Context Words. In order to establish the context, we used custom software (available from the first author) to identify which words were used at least once by every one of the 86 Senators. There were 4,010 of these words. We eliminated 602 very common words (e.g., the), called stopwords: contractions (e.g., aren't), fillers (e.g., anyway), and prepositions (e.g., followed); this list of stopwords is available on request.

Outcome Variable: DW-NOMINATE First Dimension Scores

The variable we are trying to capture in this study is political ideology of U.S. Senators. Political ideology was operationalized with a measure of ideology that is based on Senatorial roll call votes. Specifically, we used the Dynamic Weighted Nominal three-step Estimation (i.e., DW-NOMINATE) first dimension scores, developed by McCarty, Poole, and Rosenthal (1997). This "first dimension" is the primary latent dimension underlying roll call voting by U.S. Senators, and it captures political ideology ranging from liberal (scoring on the low end) to conservative (scoring on the high end). The main distinction inherent in the scores is that liberals favor government intervention in the economy, whereas conservatives do not. We chose these scores as criteria because they have been widely used in political science and related fields (Carroll, Lewis, Lo, Poole, & Rosenthal, 2009), thus providing a consensus criterion to which we could relate our CASS scores. We downloaded DW-NOMINATE scores from the VoteView website (Carroll, Lewis, McCarty, Poole, & Rosenthal, 2013).

Predictor: Liberal and Conservative Phrase Counts

One goal of this study is to determine the extent to which CASS scores provide novel information about individual differences above and beyond word counts or phrase counts. Thus, we needed to obtain such counts of phrases indicating political ideology. Luckily, Gentzkow and Shapiro (2010) have already conducted a study that specifies which phrases congresspersons use that are liberal and which ones are conservative. Their analysis operationalized ideology with records of how congresspersons' constituency voted for President in 2004. Using this approach, they were able to show that congresspersons whose constituency tended to vote for John F. Kerry (the liberal

candidate in the 2004 election) tended to use liberal phrases including "African American" and "price gouging"; whereas congresspersons whose constituency tended to vote for George W. Bush (the conservative candidate in the 2004 election) tended to use conservative phrases including "ten commandments" and "war on terrorism." In all, Gentzkow and Shapiro identified 150 liberal phrases and 150 conservative phrases in congresspersons' speeches. We used a custom software program to count how many times U.S. Senators used those ideological phrases.

Results

Zero-Order Correlations

In order to establish the relationship between the CASS effects and DW-NOMINATE political ideology scores, we computed a zero-order correlation. The correlation was highly significant, $r(84) = .372$, $p < .001$. This result indicates that CASS captures Senators' political ideology.

To explore the convergent validity of the two predictors of political ideology (phrase counts and CASS scores), we correlated these measures. The correlation was significant, $r(84) = .271$, $p = .011$. This indicates that there is some degree of convergent validity between the two predictors of political ideology. Nevertheless, the two predictors were sufficiently distinct that it would be possible for CASS effects to explain unique variance in political ideology scores above and beyond phrase counts.

Hierarchical Multiple Regression

Next, we examined whether the CASS effects capture variance in political ideology scores above and beyond phrase-counting scores. In order to explore this possibility, we used counts of liberal and conservative phrases from Gentzkow and Shapiro (2010). The zero-order correlation between the phrase counts and the DW-NOMINATE ideology scores was very large, $r(84) = .613$, $p < .001$, supporting the notion that phrase counts are an excellent way to capture ideology. Given the very large magnitude of this effect, any incremental gain over this phrase count result would be particularly noteworthy. A hierarchical regression model revealed that, in Step 2, the phrase counts remained significant, $\beta = .553$, $p < .001$, and critically, the CASS scores were also significant, $\beta = .222$, $p = .012$. The ΔR^2 was .046. Thus, above and beyond the phrase counts indicative of political ideology, the CASS effects helped explain approximately 5% of the variance in political ideology.

Discussion

The literature on the social psychology of language is replete with examples of the utility of word counts and phrase counts (Gentzkow & Shapiro, 2010; Tausczik & Pennebaker, 2010). Indeed, such counts are linked to a plethora of individual

differences (Ireland & Mehl, 2014; Yarkoni, 2010). The list includes constructs that have a potent social impact, such as political ideology (Gentzkow & Shapiro, 2010).

The goal we set and attained in this study was to show that, above and beyond the aforementioned kinds of counts (counts that are easily obtained via LIWC), a higher-order association method—CASS—can capture unique information in an important outcome—political ideology. As researchers have quickly identified the easy-to-see effects with straightforward language analytic techniques (Pennebaker et al., 2001), now is the time to begin using additional more nuanced approaches to language analysis for modeling individual differences. CASS is one empirically supported option for this purpose.

It is important to point out that, although both of the extant articles about CASS have been about political topics, CASS could easily be applied to other topics as well. Indeed, we envision CASS as a useful tool for automated detection of individual differences more generally. Nevertheless, this particular analysis does suggest that people interested in political ideology, such as followers of Gentzkow and Shapiro (2010), could benefit from using CASS, as we demonstrated that the way in which politicians associate target concepts is important, above and beyond the words and phrases they say.

Limitations

One specific limitation of the current analysis is that we based our phrase counting on phrases that were popular in 2004; it may be the case that some of these phrases are specific to that political period. If there were a list of phrases available for 1994 to 2010, then certainly using that list would have been preferable. However, such a list is not available. Future research will have to continue to explore the relative efficacy of phrase counts versus CASS effects in the prediction of important outcome variables (e.g., political ideology). For now, it seems reasonable to tentatively conclude that CASS holds some promise in capturing variance that phrase counts cannot capture.

The primary limitation of the approach in general is that it requires large corpora in order to attain a stable co-occurrence matrix. Either low frequencies of the target words or low frequencies of the context words have the potential to introduce error into the analytic process. When such frequencies are low, the co-occurrence matrix is sparse, with many zeros, and that tends to destabilize the analysis. It is likely that some variant of factor analysis will have to be applied to the co-occurrence matrix, as was done in Latent Semantic Analysis, in order for CASS to be applicable to sparse matrices. At the moment, however, CASS can clearly be successfully applied to large text files.

Appendix A

This appendix provides an example of how a CASS analysis proceeded for this study. All of this is done automatically by our CASS code, and it is expanded here simply for pedagogical purposes.

A1. Text Example

This is the text analyzed by CASS.

conservative contextworda
conservative contextworda
liberal contextworda contextwordb contextwordc
liberal contextwordc
negative contextwordc
negative contextwordd
negative contextworda
positive contextwordd contextwordc
positive contextworda
positive contextworda

A2. Context by Target Word Matrix: Number of Sentences Containing Co-Occurrences

This matrix represents the co-occurrences of target and context words in the text example. For instance, "conservative" co-occurs twice with contextworda (i.e., cwa).

	Count→	6	1	4	2
Count ↓		cwa	cwb	cwc	cwd
2	conservative	2	0	0	0
2	liberal	1	1	2	0
3	positive	2	0	1	1
3	negative	1	0	1	1

A3. Actual Context by Target Word Matrix: Proportion of Sentences Containing Co-Occurrences

These values are obtained by dividing the raw co-occurrence values by the total number of sentences (which is 10 in the text). Thus, for example, conservative and contextworda co-occur twice out of 10 possible co-occurrences, and thus the corresponding cell entry is 2/10 = .20.

	Count →	6	1	4	2
Count ↓		cwa	cwb	cwc	cwd
2	conservative	0.200	0.000	0.000	0.000
2	liberal	0.100	0.100	0.200	0.000
3	positive	0.200	0.000	0.100	0.100
3	negative	0.100	0.000	0.100	0.100

A4. Expected Context by Target Word Matrix: Proportion of Sentences That Would by Chance Contain Co-occurrences

These values are obtained by multiplying the likelihood of encountering the context word by the likelihood of encountering the target word. For example, if the likelihood of encountering the context word is 6/10 (i.e., .600) and the likelihood of encountering the target word is 2/10 (i.e., .200), then the likelihood of encountering a co-occurrence is 12/100 (i.e., .120). This is an important step because it controls for mean differences in the rows.

	Count →	6	1	4	2
Count ↓		cwa	cwb	cwc	cwd
2	conservative	0.120	0.020	0.080	0.040
2	liberal	0.120	0.020	0.080	0.040
3	positive	0.180	0.030	0.120	0.060
3	negative	0.180	0.030	0.120	0.060

A5. Actual/Expected Matrix: Co-Occurrence of Terms Relative to Expected Rate

These values are obtained by dividing the Actual Matrix by the Expected Matrix in a cell-by-cell fashion. Thus, taking the upper left entry as the example, .20/.12 = 1.667.

	cwa	cwb	cwc	cwd
conservative	1.667	0.000	0.000	0.000
liberal	0.833	5.000	2.500	0.000
positive	1.111	0.000	0.833	1.667
negative	0.556	0.000	0.833	1.667

A6. Actual/Expected Matrix, Capped at 2.00 Maximum

These values are identical to the values in the previous matrix with the exception that values >2.000 are capped at 2.000 here.

	cwa	cwb	cwc	cwd
conservative	1.667	0.000	0.000	0.000
liberal	0.833	2.000	2.000	0.000
positive	1.111	0.000	0.833	1.667
negative	0.556	0.000	0.833	1.667

A7. Example Results

The final step is to correlate pairs of the rows across columns of the Actual/Expected Matrix, Capped; then, the correlations are converted to *z*-scores using Fisher's *r*-to-*z* transform.

$$Z_{r(\text{conservative, positive})} = 0.203$$
$$Z_{r(\text{conservative, negative})} = -0.203$$
$$Z_{r(\text{liberal, positive})} = -1.335$$
$$Z_{r(\text{liberal, negative})} = -1.019$$

Apply the following equation:

$$[Z_{r(\text{conservative, positive})} - Z_{r(\text{conservative, negative})}] - [Z_{r(\text{liberal, positive})} - Z_{r(\text{liberal, negative})}]$$

Thus,

$$\text{CASS Effect} = [.203 - -.203] - [-1.335 - -1.019] = .72,$$

indicating a conservative slant.

Appendix B

This is a list of the target words that were used in this study.

Liberal Words

democrat	democrats	left	progressive	progressives

Conservative Words

conservative	conservatives	republican	republicans

Positive Words

admirable	correct	good	perfect	terrific
advantageous	correctly	great	perfectly	useful
awesome	effective	important	pleasant	valid
best	effectively	importantly	promising	virtue
better	efficient	impressive	remarkable	wise
bright	excellent	love	satisfactory	wonderful
brilliant	fantastic	magnificent	strong	
commendable	favorable	marvelous	suitable	
competent	favorite	necessary	super	
constitutional	flawless	optimal	superior	

Negative Words

abuse	destructive	flop	invalid	stupid
abusive	devastate	flopper	irrational	terrible
alarming	devastating	flopping	liar	unconstitutional
appalling	devil	flunk	loser	unnecessary
arrogant	difficult	flunked	ludicrous	unremarkable
aversive	difficulty	flunking	messy	unsuitable
awful	disappoint	harm	nasty	useless
bad	disappointing	harmful	negative	vicious
careless	disgust	horrible	neglect	weak
carelessly	disgusted	horribly	neglectful	weakest
cheat	disgusting	horrify	outrageous	weakling
complain	disgustingly	horrifying	pain	whine
contradict	dislike	horror	painful	whining
contradicting	disliked	hurting	pains	wicked
contradiction	dismay	hurts	pathetic	wickedly
crap	doubt	ignorant	pitiful	worse
crying	dread	illegal	resent	worst
degrade	dumb	inadequate	resentful	worthless
degraded	evil	incompetent	ridiculous	wrong
deprivation	fail	incorrect	rude	wrongly
deprived	failed	ineffective	ruin	
deprives	failing	ineffectively	sinister	
depriving	failure	inferior	spoiled	

Acknowledgments

We thank Dave Balota, editor Howard Giles, and two anonymous reviewers for providing helpful feedback on previous drafts.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The first author was supported by funds from the Jack N. Averitt College of Graduate Studies at Georgia Southern University in Statesboro, GA.

References

Carroll, R., Lewis, J., McCarty, N., Poole, K., & Rosenthal, H. (2013). *DW-nominate scores with bootstrapped standard errors*. Retrieved from http://voteview.org/dwnominate.asp

Carroll, R., Lewis, J. B., Lo, J., Poole, K. T., & Rosenthal, H. (2009). Measuring bias and uncertainty in DW-NOMINATE ideal point estimates via the parametric bootstrap. *Political Analysis*, *17*, 261-275.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*, 22-29.

Fast, L. A., & Funder, D. C. (2008). Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *Journal of Personality and Social Psychology*, *94*, 334-346.

Fetzer, A. (2008). "And I think that is a very straightforward way of dealing with it": The communicative function of cognitive verbs in political discourse. *Journal of Language and Social Psychology*, *27*, 384-396.

Gentzkow, M., & Shapiro, J. M. (2010). What drives media slant? Evidence from U.S. daily newspapers. *Econometrica*, *78*, 35-71.

GPO. (2012). *Digital collection of the congressional record*. Retrieved from http://www.gpo.gov/fdsys/browse/collection.action?collectionCode=CREC

Holtzman, N. S., Schott, J. P., Jones, M. N., Balota, D. A., & Yarkoni, T. (2011). Exploring media bias with semantic analysis tools: Validation of the Contrast Analysis of Semantic Similarity (CASS). *Behavior Research Methods*, *43*, 193-200.

Ireland, M. E., & Mehl, M. R. (2014). Language use and personality. In T. Holtgraves (Ed.), *Oxford handbook of language and social psychology* (pp. 201-218). New York, NY: Oxford University Press.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1-37.

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*, 211-240.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods Instruments & Computers*, *28*, 203-208.

Matsumoto, D., & Hwang, H. C. (2013). The language of political aggression. *Journal of Language and Social Psychology*, *32*, 335-348.

McCarty, N. M., Poole, K. T., & Rosenthal, H. (1997). *Income redistribution and the realignment of American politics*. Washington, DC: SEI Press.

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*, 862-877.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: LIWC 2001*. Mahwah, NJ: Erlbaum.

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, *41*, 647-656.

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*, 24-54.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, *44*, 363-373.

Author Biographies

Nicholas S. Holtzman is an assistant professor of psychology at Georgia Southern University in Statesboro. He is one of the creators of CASS. A list of his publications can be seen at http://www.nickholtzman.com/publications.htm

Simon Kwong conducted his senior thesis under the guidance of Nick Holtzman, while both of them were at Washington University in St. Louis. He majored in political science. As of this writing, he lives and works in New York.

Kirsten L. Baird is, as of this writing, completing her undergraduate degree in psychology at Georgia Southern University. Her interests revolve around clinical and counseling psychology, and she plans to pursue graduate work in one of these areas.