Using Syntactic and Semantic Context to Explore Psychodemographic Differences in Self-reference

Masoud Rouhizadeh[†§], **Lyle Ungar**[§], **Anneke Buffone**[§], **H Andrew Schwartz**[†§]
[†]Stony Brook University, [§]University of Pennsylvania
mrouhizadeh@gmail.com, ungar@cis.upenn.edu, buffonea@sas.upenn.edu, has@cs.stonybrook.edu

Abstract

Psychological analysis of language has repeatedly shown that an individual's rate of mentioning 1st person singular pronouns predicts a wealth of important demographic and psychological factors. However, these analyses are performed out of context — syntactic and semantic — which may change the magnitude or even direction of such relationships. In this paper, we put "pronouns in their context", exploring the relationship between self-reference and age, gender, and depression depending on syntactic position and verbal governor. We find that pronouns are overall more predictive when taking dependency relations and verb semantic categories into account, and, the direction of the relationship can change depending on the semantic class of the verbal governor.

1 Introduction

Approximately 1 in 18 English words on Facebook are first-person singular pronouns.[1] Extensive work in psychological analyses of language has consistently found strong relation between first-person pronoun use and psychological attributes of individuals (Kendall, 1998; Pennebaker and Stone, 2003; Pennebaker, 2011; Twenge et al., 2012; Oishi et al., 2013; Carey et al., 2015). Although such findings have been replicated extensively, little is known about how the syntactic or semantic context of the pronouns may affect their relationship with human traits. Usage in subject or object position may vary, and the type of verb governing the reference may further change its relationship. For instance, while younger individuals are more likely to use 1st-person singular pronouns overall, older individuals may be more likely to use them as the subject of social verbs.

In this study we dive deep into this one type of word which makes up a large portion of our daily lives. We first look at the relationship between first person singular pronouns and age, gender, and depression. We then consider the syntactic position of the pronoun and its occurrence in the subject and direct object position. Next, we explore the self-referenced use of verbs compared to their general use across different semantic categories, followed by an examination of the rate of 1st-person singular pronoun as the subject and the object with different verb categories.

We ultimately show that pronoun relationships with human outcomes can change drastically depending on their syntactic position and the category of their verbal governor. To be more specific, our contributions include: (a) taking the role of context into account in the psychological analysis of personal pronouns, (b) distributional clustering of verbs using Canonical Correlation Analysis (CCA), and (c) exploring the integration of verbal semantic categories in the analysis of pronouns. Utilizing verb categories instead of actual verbs, enables generalization and less sparsity in the semantic comparison of the contexts in which personal pronouns are used.

[1]Within the study dataset, 5.45% of all words from self-identified English speakers were first-person pronouns.

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2054–2059,
Austin, Texas, November 1-5, 2016.

2 Background

A wealth of studies have explored pronoun use with regard to age, gender, and personality types. In fact, a whole book, "The Secret Life of Pronouns" has been dedicated summarizing such studies which have built up over several decades of work (Pennebaker, 2011). [2]

We could not come close to a full survey of such work, but rather list some of the most notable and recent results for outcomes related to those of this study. Pennebaker et. al. (2003) and Chung & Pennebaker (2007) found that the use of self-references (i.e. 'I', 'me') decreases over age. Pennebaker et. al. (2003), and Argamon et. al. (2007) showed that females use significantly more first-person singular personal pronouns compared to males. Bucci and Freedman (1981), Weintraub (1981), and Zimmermann et. al. (2013) found that first-person singular pronouns are positively correlated with depressive symptoms. These analyses do not take the role of syntactic and semantic context into consideration which may indicate interesting information about psychological factors.

3 Method

Data Set: Facebook Status Updates. Our dataset consists of the status updates of 74,867 Facebook users who volunteered to share their posts in the "MyPersonality" application (Kosinski et al., 2013), sent between January 2009 and October 2011. The users met the following criteria: (a) have English as a primary language, (b) indicated their gender and age, (c) be less than 65 years old (due to data sparsity beyond this age), and (d) have at least 1,000 words in their status updates (in order to accurately estimate language usage rates). This dataset contains 309 million words within 15.4 million status updates. All users completed a 100-item personality questionnaire (an International Personality Item Pool (IPIP) proxy to the NEO-PI-R (Goldberg, 1999). User-level degree of depression (DDep) was estimated as the average response to seven depression facet items (nested within the larger Neuroticism item pool of the questionnaire) (Schwartz et al., 2014).

Dependency Features. We used dependency annotations in order to determine the syntactic function of personal pronouns i.e. subject (S) and direct object (DO). We obtained dependency parses of our corpus using Stanford Parser (Socher et al., 2013) that provides universal dependencies in (*relation, head, dependent*) triples. In the next step, we extracted the words in in the nominal subject (*"nsubj"*) and direct object (*"dobj"*) positions including *nsubj* 1st-person singular pronoun *"I"*, and *dobj* 1st-person singular pronoun *"me"*. We also extracted the corresponding verbs for each of the nominal subjects, and direct object words.

Verb categorization. In order to integrate the verbal semantic categories in the syntactic analysis of pronouns, we utilize two verb categorization methods (a) linguistically-driven Levin's Verb Classes, and (b) empirically-driven verb clustering based on CCA.

Levin's verb classes (Levin, 1993) includes around 3100 English verbs classified into 47 top level, 193 second and third level classes. This classification is based on Levin's hypothesis that the syntactic behavior of a verb is influenced by its semantic properties, indicating that identifying sets of verbs with comparable behavior at the syntax level will lead to coherent clusters of semantically similar verbs. In this paper we used all of the 193 second and third level Levin's classes (*Lev*). As an alternative way, we also used the 50 top most frequent sub-classes in our social media data (*LevTop*).

To derive empirically driven clusters we use Canonical Correlation Analysis (CCA), a multi-view dimensionality reduction technique. CCA has previously been used in word clustering methods such as multi-view learning of word embeddings (Dhillon et al., 2011), or multilingual word embeddings (Ammar et al., 2016). The advantage of a multi-view technique is that we can leverage both the subject and object context. More precisely, we performed sparse CCA on matrix x that includes 5k by 10k verb-by-nominal-subject (*nsubj*) co-occurrences, and matrix z that includes 5k by 10k verb-by-direct-object (*dobj*) co-occurrences. The output of CCA is a subject by component matrix

[2] To quantify the pervasiveness of pronoun studies in social science, we consider the citation count, via Google Scholar (July, 2016), to works mentioning "pronoun" by one of the top researchers, James W. Pennebaker, which number over 10,000.

Feature Set	Gender (AUC)	Age (MSE)	Dep (MSE)
$P(1p)$	.512	78.9	90.1
$P(1p\|r)$	.589	76.4	90.3
$P(1p\|r,c)$, *Lev*	.660	70.0	89.8
$P(1p\|r,c)$, *Lev* & sent	.695	**68.3**	**89.1**
$P(1p\|r,c)$, *LevTop*	.660	71.5	89.8
$P(1p\|r,c)$, *LevTop* & sent	**.669**	69.0	89.3
$P(1p\|r,c)$, *CCA-D*	.634	73.4	90.3
$P(1p\|r,c)$, *CCA-D* & sent	.649	71.5	89.7
$P(1p\|r,c)$, *CCA-KM*	.632	72.6	90.3
$P(1p\|r,c)$, *CCA-KM* & sent	.645	70.9	89.9

Table 1: Area under the ROC curve (AUC) for gender (higher is better), and Mean Square Error (MSE) for age and depression prediction (lower is better), and the prediction using 1st-per pronoun use overall, in subject and object position, and given verb categories.

Verb Clusters	r
1st person singular pronoun use	-.17
1st person singular nominal subject	
thank, celebrate, welcome, greet, applaud	.09
shake, freeze, melt, collect, bend, twist, squeeze	.08
hate, fear, regret, dislike, despise, dread, tolerate	-.16
write, draw, type, print, scratch, plot, sketch	-.10
1st person singular direct object	
join, pool, merge	.05
deny, suspect	.04
hate, fear, regret, dislike, despise, dread, tolerate	-.09
bore, worry, scare, bother, annoy, disappoint	-.08

Table 2: Linear regression coefficient of age and 1st person pronoun use in different verb clusters.

(u: *subject-view*), and object by component matrix (v: *object-view*). We then build matrix S by multiplying x by u and matrix O by multiplying z by v to get the verbs by CCA-components from *subject-view*, and verbs by object components from *object-view* respectively. In order to cluster verbs from direct CCA components, we use the average score of *subject-view* and *object-view* components, assigning verbs to those components for which they have a non-zero absolute weight (*CCA-D*). Sparse CCA zeros-out verbs from multiple components so as to assign verbs to components, but we also explore normal CCA and cluster the verbs using k-means ($k = 30$) clustering from the z-scaled values of S and O matrices (*CCA-KM*).

Both Levin's and CCA-based verb classes are derived from syntactic behavior. As a result, they often do not distinguish antonyms. For instance, Levin's "admire" verb class contains both 'love' and 'hate". Building on research showing positive and negative emotions differ across age and gender (Schwartz et al., 2013), we integrate valence information in our verb clustering. We used positive and negative sentiment scores from EmoLex word-emotion association lexicon (Mohammad and Turney, 2013), dividing each of our clusters into *positive*, *negative*, and *neutral* sub-classes.

Analysis. We explore the use of 1st-person singular pronouns across age and gender in different syntactic and semantic contexts. Features are encoded as the mean from maximum likelihood estimation over the probability of mentioning a first person singular pronoun in a given context.

(a) The overall usage first person singular pronoun:

$$P(1p) = P(PN = 1p)$$

(b) The probability of using first person singular pronoun in the *nsubj*, and the *dobj* positions:

$$P(1p|r) = P(PN = 1p \mid rel = r)$$

where $rel \in \{nsubj, dobj\}$.

(c) The probability of using first person singular pronoun in the *nsubj*, and the *dobj* positions of a given verb category:

$$P(1p|r,c) = P(PN = 1p \mid rel = r, vcat = c)$$

where $rel \in \{nsubj, dobj\}$ and $vcat$ is the set of all verb categories being considered.

4 Evaluation

The goal of our work is to expand the knowledge of how the first-person singular pronoun, one of the most common word types in English, is related to who we are – our demographics and psychological states. We work toward this goal in an empirical fashion, by first replicating known general relationships of 1st-person singular pronouns with gender, age, and depression, exploring how their use in different syntactic positions, and, finally, by looking at relationships within specific semantic contexts according to the verb classes described earlier.

Verb Clusters	β
1st person singular pronoun use	.11
1st person singular nominal subject	
love, enjoy, respect, adore, cherish, admire	.29
miss, like, appreciate, trust, support, value	.28
destroy	-.08
kick, shoot, slap, smash, shove, slam	-.07
1st person singular direct object	
make, blow, roll, hack, cast	.22
hold, handle, grasp, clutch, wield, grip	.18
hit, kick, strike, slap, smash, smack, bang, butt	-.10
add, mix, connect, link, combine, blend	-.04

Table 3: Logistic regression coefficient between gender and 1st person singular pronoun use in different verb clusters (positive is female).

Replication. We use standardized linear and logistic regression to correlate gender, age, and depression with $P(1p)$ (first-person singular pronoun use). We control for age in the case of gender, gender in the case of age, and both gender and age in the case of depression by including them as covariates in the regression and reporting the unique coefficient for the variable in question. Logistic regression is used for gender, since it is binary, while linear regression is used for the continuous age and depression variables. Confirming past results, we found significant relationships between first-person pronoun usage and gender ($\beta = .11$, $p < .001$), age ($r = -0.17$, $p < .001$), and depression score ($r = -0.06$, $p < .01$).

Syntactic Context. Taking dependency relationships into account ($P(1p|r)$), we observed shifts in the magnitude of correlations. Specifically, we found significant negative correlations between age and using 1st-person singular pronoun in the subject ($r = -0.12$, $p < .001$), and the object positions ($r = -0.17$, $p < .001$). For gender we found a significant positive correlation between being female and the probability of using 1st-person singular pronoun ($r = 0.11$, $p < .001$), and 1st-person singular pronoun in subject position ($r = 0.16$, $p < .001$). For depression a significant positive correlation between with $P(1p)$ ($r = 0.06$, $p < .05$), and using 1st-person singular pronoun in the subject position ($r = 0.07$, $p < .05$).

Syntactic and Semantic Context. Table 1 reports the area under the ROC curve (AUC) for gender prediction and the Mean Square Error (MSE) for predicting age and depression based on $P(1p)$, $P(1p|r)$, and $P(1p|r, c)$, driven from various categorization approaches. We used AUC since it can capture more differences in performance by evaluating the class probabilities of test instances rather than just finding whether it was right or wrong. We applied 10-fold cross-validation with a linear-SVM in the case of gender, and ridge-regression in the case of depression. The obtained results reveal a consistent pattern: in gender, age, and depression prediction all the features that take context into account outperform $P(1p)$ which is the vastly reported measure of self-reference in the literature. This suggests that there is more information to be gained by utilization syntactic and semantic context. In other words, we can achieve a more meaningful, deeper insight into the relationship of subject and object position of the first person in different contexts, revealing a more complex, and more insightful set of relations.

We achieve the best performance by utilizing verb categories. We first observe that integrating sentiment helps in nearly all verb categorization approaches. Next, we see that while both CCA and Levin verb clusters yield improvement in prediction accuracy, our performance gains using the data-driven CCA-based verb clustering are not as large as that from Levin's linguistically-driven classes.

While we believe our features can improve prediction accuracy, that is not the primary application of social science research. Rather, it is correlating the behavior of referencing the self with psychological conditions, like depression, in order to gain human insights. In the case of correlating behavior with a psychological measure, Pearson coefficients above .1 are considered noteworthy and above .3 are considered approaching a "correlational upperbound" (Meyer et al., 2001).

Tables 2, 3, and 4 show the most predictive features, using the best performing clustering method (i.e. Levin & Sentiment). Note that in the case of age and gender, we see that not only does the magnitude of the relationship change, but it's possible that the direction can completely change.

For example, while males are less likely to

Verb Clusters	r
1st person singular pronoun use	.06
1st person singular nominal subject	
cry, worry, suffer, fear, bother, ache, mourn, anger	.11
scare, annoy, confuse, depress, upset, disappoint	.11
1st person singular direct object	
kill, murder, slay, slaughter, butcher	.09
scare, annoy, confuse, depress, upset, disappoint	.07

Table 4: Linear regression coefficient of depression score and 1st person singular pronoun use in different verb clusters.

use first-person singular pronouns overall, they are much more likely to use them as the subject of aggressive physical contact verbs like “kick", “shoot", “slap", and “smash", suggesting men are more likely to express themselves as agents of aggressive contact. On the other hand, women use first-person singulars in the social sphere, particularly in an affiliative context. They assert themselves as agents of empowering and encouraging others (e.g. “love", “enjoy", “cherish", “admire") and faith in others (e.g. “trust", “value", “support", “respect").

5 Conclusion

We have shown that the well-studied link between the first-person singular pronoun and human psycho-demographics is largely dependent on its syntactic and semantic context. Many theories and conclusions are built on such relationships, but here we show these relationships depend on verbal context; correlations can shrink, grow, and even change directions depending on the verbs governing the pronoun. For example, while the usage of 1st person singular pronoun decreases over age, it increases if it is used as the subject of verbs such as “thank", and “celebrate", or as the object of verbs such as “join". Similarly, while females tend to use 1st person singular pronouns more than males, they use them less often as the subject of “destroy” verbs or as the object of “hit” and “kick” verbs.

By integrating syntactic dependency relationships along with semantic classes of verbs, we can capture more nuanced linguistic relationships with human factors. Beyond pronouns, we ultimately aim to expand the regimen of *open-vocabulary* techniques available for the analysis of psychologically-relevant outcomes.

Acknowledgments

The authors acknowledge the support from Templeton Religion Trust, grant TRT-0048.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9).

Wilma Bucci and Norbert Freedman. 1981. The language of depression. *Bulletin of the Menninger Clinic*, 45(4):334.

Angela L Carey, Melanie S Brucks, Albrecht CP Küfner, Nicholas S Holtzman, Mitja D Back, M Brent Donnellan, James W Pennebaker, Matthias R Mehl, et al. 2015. Narcissism and the use of personal pronouns revisited. *Journal of personality and social psychology*, 109(3):e1.

Cindy Chung and James W Pennebaker. 2007. The psychological functions of function words. *Social communication*, pages 343–359.

Paramveer S. Dhillon, Dean Foster, and Lyle Ungar. 2011. Multi-view learning of word embeddings via cca. In *Advances in Neural Information Processing Systems (NIPS)*, volume 24.

Lewis R Goldberg. 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, 7(1):7–28.

Lori Kendall. 1998. Meaning and identity in âĂIJcyberspaceâĂİ: The performance of gender, class, and race online. *Symbolic interaction*, 21(2):129–153.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Gregory J Meyer, Stephen E Finn, Lorraine D Eyde, Gary G Kay, Kevin L Moreland, Robert R Dies, Elena J Eisman, Tom W Kubiszyn, and Geoffrey M Reed. 2001. Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist*, 56(2):128.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Shigehiro Oishi, Jesse Graham, Selin Kesebir, and Iolanda Costa Galinha. 2013. Concepts of happiness across time and cultures. *Personality and Social Psychology Bulletin*, 39(5):559–577.

James W Pennebaker and Lori D Stone. 2003. Words of wisdom: language use over the life span. *Journal of personality and social psychology*, 85(2):291.

James W Pennebaker. 2011. The secret life of pronouns. *New Scientist*, 211(2828):42–45.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

H Andrew Schwartz, Johannes Eichstaedt, Margaret L Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125. Citeseer.

Richard Socher, John Bauer, Christopher D Manning, and Andrew Y Ng. 2013. Parsing with compositional vector grammars. In *ACL (1)*, pages 455–465.

Jean M Twenge, W Keith Campbell, and Brittany Gentile. 2012. Male and female pronoun use in us books reflects womenâĂŹs status, 1900–2008. *Sex roles*, 67(9-10):488–493.

Walter Weintraub. 1981. *Verbal behavior: Adaptation and psychopathology*. Springer Publishing Company.

Johannes Zimmermann, Markus Wolf, Astrid Bock, Doris Peham, and Cord Benecke. 2013. The way we refer to ourselves reflects how we relate to others: Associations between first-person pronoun use and interpersonal problems. *Journal of research in personality*, 47(3):218–225.