Hansen Andrew Schwartz

University of Pennsylvania
3701 Market St., Room 219
Philadelphia, PA 19104

Phone: (407) 415-5280
Email: hansens@seas.upenn.edu
http://www.seas.upenn.edu/~hansens/

Research Interests

I study **large and scalable computational linguistics for health and social sciences**. This includes novel **natural language processing** and machine learning techniques for: **1)** discovering **new links with health and well-being** as manifest through language in **social media**, **2)** understanding **people and personality**, and **3)** developing **language-based metrics of psychological variables**. I also develop algorithms in **lexical semantics** for **word sense disambiguation**, **concept similarity**, and automatic **knowledge acquisition from the Web**.

Education

Postdoctoral Research Fellow, *University of Pennsylvania*, Current.
Mentors: *Lyle H. Ungar* and *Martin E. P. Seligman*

Ph.D. Computer Science, *University of Central Florida*, Spring 2011.
Advisor: *Fernando Gomez*
Dissertation: *The Acquisition of Lexical Knowledge from the Web for Aspects of Semantic Interpretation*

M.S. Computer Science, *University of Central Florida*, Spring 2006.

B.S. ***w/ honors*** **Computer Science**, *University of Central Florida*, Spring 2004.

Employment

Lead Research Scientist. *WWBP, University of Pennsylvania*. Philadelphia, PA. 2011 - current.

I am leading an interdisciplinary project between Computer & Information Science and the Positive Psychology Center at Penn which is pioneering techniques for measuring and understanding health and psychological well-being based on language in social media. I direct the daily research tasks, develop new analysis techniques, prioritize projects, hire staff, guide student research projects, organize team and collaborative meetings, and over-see a quarter-million dollar budget.
Other Affiliation: *Penn Social Media and Health Innovation Lab*

Research Assistant. *University of Central Florida*. Orlando, FL.

- **Lead Developer / Researcher**.
  Linguistic Preprocessing and Problem Report Trend Analysis. 2008 - 2011.
  NASA Engineering and Safety Center Grant/Cooperative Agreement NNX08AJ98A.

  We worked on research and development of a linguistic preprocessing system for trend analysis of space industry problems reports. I helped plan and develop a syntactic parser and a fully automatic domain-adaptable spell correction system. I was also responsible for communication with our colleagues at the Johnson Space Center in order to provide system support and updates.

- **Developer / Researcher**.
  NASA Snowy: Prototype Delivery, Extension, and Testing by End-users. 2004 - 2005.
  ASRC Aerospace Corporation Grant NAS1003006.

  We researched and programed *NASA Snowy*, a space shuttle problem report search tool utilizing natural language processing techniques. My notable contributions included programming and improving the efficiency of the search algorithm itself, adding a spell correction module to the system, and site visits to the Kennedy Space Center to update the system. We were told the system saved days of launch delay in Summer of 2009.

Instructor / Teaching Assistant. *University of Central Florida*. Orlando, FL.

- **Instructor of Record**. *Object Oriented Programming*, 2006.

  I taught a class of ~100 students, including lesson preparation, assignment and test creation, plus managing a team of teaching assistants.
- **Recitation Instructor (TA)**. *Introduction to Discrete Structures, Computer Science 2*, 2006.
- **Grader with Office Hours (TA)**. *Computer Science 1 and 2, Object Oriented Programming, Introduction to/Advanced Artificial Intelligence, Natural Language Understanding*, 2005 - 2006.

Lead Instructor. *Various High School Marching Bands*. Orlando, FL. 2004-2011

Web Software Developer. *Midera Solutions*. Orlando, FL. 2004

Assistant Manager. *Game Trader*. Casselberry, FL. 1998 - 2001

Refereed Publications

Note: In the field of natural language processing, the most reputable venues, in terms of acceptance rates and impact factors, are conferences. Both the journal and conference papers listed below, which were all peer reviewed, are full 3000 to 6000 word papers unless noted as short (+).

H. Andrew Schwartz, J. Eichstaedt, L. Dziurzynski, M. L Kern, S. M Ramones, M. Agrawal, A. Shah, D. Stillwell, M. Kosinski, M. E. P. Seligman and L. H. Ungar. 2013. **Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach**. *In PLOS ONE 8(9).*
(**> 100,000 views** and 3,000 pdf downloads in first month online – Oct. 2013. press: WIRED, Slate, Huffington Post)

We analyzed 700 million words, phrases, and topic instances collected from the Facebook messages of 75,000 volunteers, who also took standard personality tests. In our open-vocabulary technique, the data itself drives a exploration of language that distinguishes people, finding connections that are not captured with traditional closed-vocabulary word-category analyses. Our analyses, which resulted in state-of-the-art gender and personality prediction, also shed new light on psychosocial processes.

M.L. Kern, J.C. Eichstaedt, H. A. Schwartz, G. Park, L. H. Ungar, D. J. Stillwell, M. Kosinski, L. Dziurzynski, and M.E.P. Seligman. 2013. **From "sooo excited!!!" to "so proud": Using language to study development.** *To appear in Developmental Psychology.*

M.L. Kern, J.C. Eichstaedt, H. A. Schwartz, L. Dziurzynski, L. H. Ungar, D. J. Stillwell, M. Kosinski, S.M. Ramones, and M.E.P. Seligman. 2013. **The Online Social Self: An Open Vocabulary Approach to Personality.** *To appear in Assessment.*

H. Andrew Schwartz, Johannes Eichstaedt, Richard Lucas, Lukasz Dziurzynski, Margaret L Kern, Gregory Park, Megha Agrawal, Shrinidhi K Lakshmikanth, Shneha Jha, Martin Seligman and Lyle Ungar. 2013. **Characterizing Geographic Variation in Well-Being using Tweets.** *in Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013).* Boston, MA. **(acceptance rate: 20%)** *

The language used in tweets from 1,300 different US counties was found to be predictive of the subjective well-being of people living in those counties as measured by representative surveys. Topics derived from the tweets using *Latent Dirichlet Allocation* improved accuracy in predicting life satisfaction over and above standard demographic and socio-economic controls. Words relating to outdoor activities, spiritual meaning, exercise, and good jobs correlate with increased life satisfaction.

H. Andrew Schwartz, Johannes Eichstaedt, Eduardo Blanco, Lukasz Dziurzynski, Margaret L. Kern, Stephanie Ramones, Martin Seligman, and Lyle Ungar. 2013. **Choosing the Right Words: Characterizing and Reducing Error of the Word Count Approach.** *in *SEM-2013: Second Joint Conference on Lexical and Computational Semantics. Atlanta, GA.* **(acceptance rate: 31%)** *

Social scientists are increasingly counting words in social media as a measure of psychological state. We find lexical ambiguity the most prevalent type of error for such an approach, and show that one can reduce error with a simple automatic refinement to remove highly ambiguous words from lexica.

H. Andrew Schwartz, Johannes Eichstaedt, Lukasz Dziurzynski, Margaret Kern, Eduardo Blanco, Michal Kosinski, David Stillwell, Martin Seligman, Lyle H. Ungar. 2013. **Toward Personality Insights from Language Exploration in Social Media.** *In AAAI-2013 Spring Symposium: Analyzing Microtext. Stanford, California.* *

Although social media are widely studied, computational linguistics research has mostly focused on *prediction* tasks. We show how social media can also be used to gain *insights*, exploring language use as a function of age, gender, and personality from a large dataset of Facebook posts.

H. Andrew Schwartz, Fernando Gomez, Lyle H. Ungar. 2012. **Improving Supervised Sense Disambiguation with Web-scale Selectors.** *In COLing-2012: the 24th International Conference on Computational Linguistics.* Mumbai, India. **(acceptance rate: ∼30%)** *

This research improves word sense disambiguation performance by including a new class of semi-supervised features leveraging large unannotated corpora. Classifiers trained with the new features achieve an error reductions of 15.4% on standard coarse-grained disambiguation tasks.

Adam Kapelner, Krishna Kaliannan, H. Andrew Schwartz, Lyle Ungar and Dean Foster. 2012. **New Insights from Coarse Word Sense Disambiguation in the Crowd.** *In COLing-2012: the 24th International Conference on Computational Linguistics.* Mumbai, India. +

Sneha Jha, H. Andrew Schwartz, Lyle H. Ungar. 2012. **Using Word Similarities to better Estimate Sentence Similarity.** *In SemEval-2012: the 6th International Workshop on Semantic Evaluation.* Montreal, Canada. +

Hansen A. Schwartz, Fernando Gomez. 2011. **Evaluating Semantic Metrics on Tasks of Concept Similarity**. *In Cross-Disciplinary Advances in Applied NLP: Issues and Approaches.* IGI Global.

Hansen A. Schwartz, Fernando Gomez. 2011. **Evaluating Semantic Metrics on Tasks of Concept Similarity.** *In FLAIRS-24: Proceedings of the 24th Florida Artificial Intelligence Research Society.* Palm Beach, Florida. *

Past studies of similarity and relatedness metrics have focused entirely on relatedness or evaluated judgments over words rather than *concepts*. This research fills the gap by evaluating metrics through human judgements of *concept similarity* and through an disambiguation algorithm utilizing *similarity*.

Hansen A. Schwartz, Fernando Gomez. 2010. **UCF-WS: Domain Word Sense Disambiguation Using Web Selectors.** *In Proceedings of SemEval-2010: the 5th International Workshop on Semantic Evaluation.* Uppsala, Sweden. +

Explores the use of domain predominant sense information to improve the Web Selectors algorithm.

Hansen A. Schwartz, Fernando Gomez. 2009. **Using Web Selectors for the Disambiguation of All Words**. *In NAACL-2009 Workshop on Semantic Evaluations: Recent Achievements and Future Directions.* Boulder, Colorado. *

This research generalized the Web selectors algorithm to disambiguate verbs, adverbs, and adjectives in addition to nouns. It explores the effectiveness of each type of context selector by part of speech.

Hansen A. Schwartz, Fernando Gomez. 2009. **Acquiring Applicable Common Sense Knowledge from the Web**. *In NAACL-2009 Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics.* Boulder, Colorado. *

In this paper, a framework for acquiring common sense knowledge is presented. Relationships between nouns are retrieved from the Web, and through analysis over WordNet, we were able to successfully apply the knowledge to improve results of a state of the art word sense disambiguation system.

J. T. Malin, C. Millward, H. A. Schwartz, F. Gomez, D. R. Throop, C. Thronesbery. 2009. **Linguistic Text Mining for Problem Reports.** *In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics.* San Antonio, Texas.

Hansen A. Schwartz, Fernando Gomez. 2008. **Acquiring Knowledge from the Web to be used as Selectors for Noun Sense Disambiguation.** *In CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning.* Manchester, England. **(acceptance rate: 21%)** *

This paper presents a method of acquiring knowledge from the Web for noun sense disambiguation. We found *Web Selectors*, words which take the place of a target word instance in its local context, serve for a disambiguation system to essentially learn the concepts of WordNet that the sense of a target word should belong. Results were on par with state of the art knowledge-based noun sense disambiguation.

Hansen A. Schwartz, Fernando Gomez, Christopher Millward. 2008. **A Semantic Feature for Verbal Predicate and Semantic Role Labeling Using SVMs.** *In FLAIRS-21: Proceedings of the 21st Florida Artificial Intelligence Research Society.* Coconut Grove, Florida. *

This paper shows that semantic role labeling is a consequence of accurate verbal predicate labeling. In doing so, a novel type of semantic feature is presented for predicate labeling over a new corpus.

* included or will include full oral presentation.
+ short paper; all other papers listed were long format.

In Preparation

Predicting Well-Being through Language Use: Multi-level Modeling.

We present the task of predicting well-being, as measured by the *satisfaction with life* scale, through the language people use in social media. Well-being, encompassing much more than emotion or mood, is attributed to people rather than text. However, with individual *messages* themselves being the units that communicate well-being, we explore cascaded and joint message-to-user models which utilize message-level data to inform user-level predictions.

Language-Based Psychometrics.

This work evaluates regression models of personality based on language-use as if they were traditional psychometrics. The models are tested for reliability over time, against 3rd variables, and within the factors of the personality construct. Findings: language based psychometrics behave a lot like traditional psychometrics and should be considered a new approach in psychology for measuring one's psychological state and traits.

Twitter Correlates with Atherosclerotic Heart Disease Mortality across US Counties..

Language in tweets predicts mortality rates from atherosclerotic heart disease above and beyond standard community predictors: income, education, smoking rates, obesity rates. Furthermore, language in tweets provides a window into the factors affecting such rates– Communities that tweet more angry words ('anger', 'hate') are more likely to have higher rates of heart disease.

The Prediction of Community Disease and Mortality Rates Based on Twitter Language..

"Twitter Correlates with Atherosclerotic Heart Disease" demonstrated the power of information in twitter for health. Similar variables are available for a multitude of outcomes. This work looks at which diseases are easiest or most difficult to predict based on information in social media.

Keys are often Found in One's Pocket and other Common Sense Knowledge Acquired from the Web.
This journal-length paper, extends the 2009 workshop paper, "Acquiring Applicable Common Sense Knowledge from the Web". Notable additions include the acquisition of a more comprehensive set of knowledge, a revised utilization of a syntactic parser for result verification, an in-depth examination of the knowledge that we acquire, and utilizing a more advanced information-theoretic concept analysis.

Phrase Reconstruction for Knowledge Acquisition from the Web.
This work seeks to develop methods of reconstructing the syntactic structure of a phrase, such that the meaning is maintained. A particular focus is on using the reconstructed phrases to search and acquire knowledge from the Web.

Interdisciplinary Collaborations

Molly Ireland (Psychology, Texas Tech), Dolores Albraccin (Annenberg, Univ. of Pennsylvania).
Leveraging Social Media to Forecast HIV rates. Current.

Raina Merchant (Medicine, Univ. of Pennsylvania), Shawndra Hill (Wharton, Univ. of Pennsylvania)
Health Insights from Twitter. Current.

Jonah Berger (Wharton, Univ. of Pennsylvania)
Future-oriented language and Psychology. Current.

Barbara Mellers, Philip Tetlock (Psychology / Wharton, Univ. of Pennsylvania)
Language Analysis of Political Forecasting Comments. 2012 - current.

Eduardo Blanco (Lymba; Sothern Methodist University)
Lexical Semantics for Understanding Psychological States and Traits 2012 - 2013.

Adam Kapelner, Dead Foster (Statistics, Univ. of Pennsylvania)
Mechanical Turk for Word Sense Disambiguation. 2012.

David Stillwell and Michal Kosinski (Psychometrics Center, Cambridge Univerity)
Personality correlates in millions of Facebook messages. 2011 - 2013.

Martin Seligman, Johannes Eichstaedt, Margaret Kern, Greg Park (Positive Psychology Center, Univ. of Pennsylvania).
The World Well-Being Project. 2011 - Current.

Select Student Research Supervision

Maarten Sap (B.S. in Computer Science, *Johns Hopkins*).
Automatic Lexicon Creation and Topic Modeling for Continuous Psychological Variables. current.

Rigel Swavely (seeking B.S. in Computer Science and Cognitive Science, *UPenn*).
Creating a Language-Based Predictive Model of Optimism and Pessimism. current.

Libby Benson (B.S. in Psychology, *Univ. of Wisconsin*).
Finding Longitudinal Linguistic Correlates of Health and Psychology.. Summer 2013.

Tadas Antanavicius (seeking B.S. in Computer Science and Cognitive Science, *UPenn*).
Parsing Noisy Text of the Hardvard Red Books for longevity prediction. Summer 2013.

Sneha Jha (masters Computer Science, *UPenn*) .
Vector-Space Models of Word and Sentence Similarity. **(published in SemEval-2012).** Spring 2012.

Program Committees and Service

Program Committee. *ACL-2014 workshop: Computational Linguistics and Clinical Psychology – From Linguistic Signal to Clinical Reality.* 2014

Reviewer. *IEEE Transactions on Multimedia Manuscript Central database.* 2013

Program Committee. *EMNLP$_{2013}$: Empirical Methods in Natural Language Processing*. 2013

Program Committee. **SEM: Joint Conference on Lexical and Computational Semantics*. 2013

Program Committee. *FLAIRS-26: Florida Artificial Intelligence Research Society, Applied NLP Track*. 2013

Program Committee. *FLAIRS-25: Florida Artificial Intelligence Research Society, Applied NLP Track*. 2012

Program Committee. *FLAIRS-24: Florida Artificial Intelligence Research Society, Applied NLP Track*. 2011

Program Committee. *FLAIRS-23: Florida Artificial Intelligence Research Society, Applied NLP Track*. 2010

Program Committee. *FLAIRS-22: Florida Artificial Intelligence Research Society, Applied NLP Track*. 2009

Reviewer. *CIKM-2008: The 17th ACM Conference on Information and Knowledge Management*. 2008

Subreviewer. *ACL-2007: 45th Annual Meeting of the Association of Computational Linguistics*. 2007

Reviewer. *CIKM-2007: The 16th ACM Conference on Information and Knowledge Management*. 2007

Subreviewer. *ACL-2005: 43rd Annual Meeting of the Association of Computational Linguistics*. 2005

Notable Presentations

Presenter. (Invited – to appear) Data-driven Discovery of Psychosocial Health Factors via Social Media. *Penn Medicine Center for Health Care Innovation*. Philadelphia, PA, 2014

Presenter. (Invited) Text Mining Social Media for Health and Well-Being. *Robert Wood Johnson Foundation Fall Positive Health Meeting*. Philadelphia, PA, 2012

Presenter. (Invited) Acquisition of Common Sense Knowledge from the Web for Semantic Interpretation. *Sapienza - University di Roma, Linguistics Computing Lab*. Rome, Italy, 2011

Discussion Leader. (Invited) Applicable Common Sense Knowledge, *NAACL-2009 Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*. Boulder, CO. 2009

Discussion Leader. Comparing Symbolic to Connectionist Philosophies of AI, *UCF AI-Forum*. Orlando, FL. 2009

Topic Lecturer. Transformation-based, Part of Speech Tagging, *UCF Natural Language Understanding Course*. Orlando, FL. 2006 - 2008

Within the publications section, oral presentation of publications are indicated with *.

Select Media Coverage

WIRED. *Study: status update language used to predict Facebook users' age, gender, personality.* Oct, 2013.

Slate: Business Insider. *Scientists Used Facebook For the Largest Ever Study of Language and Personality*. Oct, 2013.

Huffington Post. *Facebook Study Reveals Links Between What You Post And Who You Are* Oct, 2013.

The Atlantic: Cities. *Twitter Can Tell Whether Your Community Is Happy or Not*. Jul, 2013.

Scholarship and Leadership

- CAE USA Graduate Fellowship
- UCF Graduate Research Fellowship
- UCF High Achievement Scholarship
- Florida Bright Future Scholarship
- UCF Honors College
- National Society of Collegiate Scholars
- Music Assistant, Marching Knights Band
- President, Kappa Kappa Psi National Honorary Music Fraternity
- Founding Member, UCF AI-Forum