Interspeech 2018
2-6 September 2018, Hyderabad

Modeling Self-Reported and Observed Affect from Speech

Jian Cheng[1], *Jared Bernstein*[1], *Elizabeth Rosenfeld*[1]
Peter W. Foltz[2], *Alex S. Cohen*[3], *Terje B. Holmlund*[4], *Brita Elvevåg*[4,5]

[1]Analytic Measures Inc., Palo Alto, CA, U.S.A.
[2]Institute of Cognitive Science, University of Colorado & Pearson, CO, U.S.A.
[3]Department of Psychology, Louisiana State University, LA, U.S.A.
[4]Department of Clinical Medicine, University of Tromsø – the Arctic University of Norway, Norway
[5]Norwegian Centre for eHealth Research, University Hospital of North Norway,Tromsø, Norway

jian.cheng@analyticmeasures.com, brita@elvevaag.net

Abstract

Listeners hear joy/sadness and engagement/indifference in speech, even when linguistic content is neutral. We measured audible emotion in spontaneous speech and related it to self-reports of affect in response to questions, such as "Are you hopeful?" Spontaneous speech and self-reports were both collected in sessions with an interactive mobile app and used to compare three affect measurements: self-report; listener judgement; and machine score. The app adapted a widely-used measure of affective state to collect self-reported positive/negative affect, and it engaged users in spoken interactions. Each session elicited 11 affect self-reports and captured about 9 minutes of speech; with 118 sessions by psychiatric patients and 227 sessions by non-clinical users. Speech recordings were evaluated for arousal and valence by clinical experts and by computer analysis of acoustic (non-linguistic) variables. The affect self-reports were reasonably reliable (α 0.73 to 0.84). Combined affect ratings from clinical-expert listeners produced reliable ratings per session (α 0.75 to 0.99), and acoustic feature analysis matched the expert ratings fairly well ($0.36 < r < 0.72$, mean 0.57), but neither human nor computed scores had high correlation with standard affect self-reported values. These results are discussed in relation to common methods of developing and evaluating affect analysis.

Index Terms: positive affect, negative affect, arousal, valence, mental health

1. Introduction

Our emotions are often evident in both what we say (the content) and how we voice it (acoustic features). Indeed, it is often maintained that a speaker's state can be accurately measured from the acoustic properties of speech (e.g. [1, 2]). However, most such studies have been conducted in controlled laboratory settings and little is known about such speech properties when produced in more natural settings and across different speaking tasks. Put simply, how well can speech-based methods model speaker states when diverse speech data are collected in real-world settings? Also, how closely are acoustic properties of vocal signals related to a person's subjective self-reports of their own inner mental state (i.e., their own estimates of arousal and valence levels)? Put differently, would we expect a close relationship between an individual's self-reported emotions and those that an external expert rater would assign by listening to a sample of speech from that same individual?

2. Delta mental state examination

We developed an iOS app for research purposes called the *delta* Mental State Examination (*d*MSE [3]). This telemental health monitoring tool is designed to be remotely administered to track changes (hence 'delta') in participants' emotional, cognitive, and psychomotor states over time. It is a tool for researching the dynamics of mental and cognitive states using both standard subjective assessments derived in clinical sciences and emerging objective technologies. This tracking is done by providing a series of short engaging tasks that require participants to listen, watch, speak, and touch to interact with the smartdevice, and thereby collectively assesses cognition, motor skill, and language. Additionally, several questions prompt participants to self-report on their emotional well-being by moving a slider to indicate their current level of positive and negative affect. We report data from the alpha version of this app that took participants on average 28 minutes to complete a daily session.

3. Self-reported and observed affect

One frequently used tool for assessing general affective states is the Positive and Negative Affect Schedule (PANAS) [4, 5] which measures Positive Affect (PA) and Negative Affect (NA). PA and NA are two primary dimensions of mood and are relatively independent. "NA - but not PA - is related to self-reported stress and (poor) coping, health complaints, and frequency of unpleasant events. In contrast, PA - but not NA - is related to social activity and satisfaction and to the frequency of pleasant events" [4].

In the *d*MSE, we included 15 digital sliders coded on a scale from 0 to 100 to let participants self-report their affect states at the present moment, with increasing scores reflecting increasing intensity of the states. Seven PA slider questions prompted self-evaluation as to how the participant felt in terms of hopefulness, calmness, appreciation, strength, ability to concentrate, happiness, and levels of energy. Eight NA slider questions prompted self-reflection regarding levels of anxiety, frustration, fear, sadness, stress, anger, pain, and feelings of helplessness. These slider questions were derived from [4, Appendix]. At the beginning of each *d*MSE session, participants were asked to respond to 11 PA and NA sliders (drawn from the 15), of which at least four concerned positive affect and at least four concerned negative affect. On average, five PA and six NA sliders were presented per session. The final self-reported PA (or NA) value per session is the average of the PA (or NA) slider response values. These values represent participants' self-reported emotional state.

10.21437/Interspeech.2018-2222

Note that our self-reported affect measures are significantly different from some previous reports of self-rated or self-reported emotion ratings from speech [6, 7]. In [6], participants were asked to rate their own speech, and in [7], participants were asked to recall what they were feeling when they produced their speech. One early study [8] had psychologists repeat neutral material as if in a particular affective state.

Arousal (degree of excitement) and valence (positive vs. negative) are two key components of emotion (e.g. [9]), and many human emotions can be modeled within these two-dimensions. Voice characteristics and others (e.g. facial expressions, hand gestures, posture, etc.) may reveal emotions so that they are observable. After the self-reported sliders in the *d*MSE, a series of structured speech interactions were used to record participants' non-acted, spontaneous speech, which may unobtrusively reveal participants' emotions.

Research to automatically recognize emotion from speech signals is ongoing [1, 2, 7, 10, 11]. Previous research on automatic measurement of continuous-valued affective arousal and valence in speech (e.g. [1, 2]). [1, 2] has shown that arousal in speech can be measured reliably, but not valence.

4. Elicited speech data

The *d*MSE app elicited on average 8.9 minutes of speech data per session for the 345 sessions analyzed in this paper. Table 1 lists the speech interaction types.

Table 1: *The structured speech interaction types per session. n is the number of different items per type presented in a session.*

Type	n	Description
greeting	7	Greeting questions, *e.g. What's new? What's on your mind?*
picture	4	Describe what is happening in a picture.
video	3	Describe what is happening in a silent video.
talk	2	Free talk on a topic, *e.g. How would you boil an egg? How has television changed family life?*
pataka	4	Rapid repetitions of a single syllable, */pa/*, */ta/* or */ka/* (alternating motion rate) or a syllable sequence, */pataka/* (sequential motion rate) [12, 13].
retell	2	Hear a story, then retell it immediately in detail. (Stories about 72 words (std = 4.6)).
re-retell	1	Retell the story you previously heard. (This appears a few item types (about 17 minutes) later in the same session).
repeat	9	Repeat the sentence you just heard.
Stroop test	1	Words (including color names) appear in different colors. Say the print color as fast as possible.
suggestion	4	Request suggestions, *e.g. What would make this app better?*

5. Speech processing and machine learning

We explored the relationship between affect apparent in speech and self-reported affect states. Apparent affect is observed by expert human raters listening to speech recordings, and affect states are those self-reported using sliders at the beginning of the sessions.

5.1. openSMILE speech signal features

The openSMILE audio feature extractor [14] is the state-of-the-art open source package that can generate a lot of low-level features based on speech signal processing (e.g. energy, loudness, MFCC, PLP, F0, probability of voicing, voice quality: jitter and shimmer, formants: F1, F2, F3, F4, etc., harmonics-to-noise ratio, etc.) and data processing (means, extremes, moments, segments, peaks, linear and quadratic regression, percentiles, durations, onsets, etc.) for voice research, affective computing, and many other applications. The measured parameters consider subglottal pressure, transglottal airflow, and vocal fold vibration, include parameters in the time domain (e.g. speech rate), the frequency domain (e.g. fundamental frequency or formant frequencies), the amplitude domain (e.g. intensity or energy), and the spectral distribution domain (e.g. relative energy in different frequency bands). We used the 2013 COMPARE feature set [11] that contains 6,373 features.

5.2. Machine learning method

Support Vector Regression (SVR) is a machine learning method that is commonly used and well-suited to emotion estimation [1, 2]. SVR is good at predicting targets when there are a lot of continuous predictor (independent) variables. It can easily be used to check if prototype machine learning ideas will work. We used the libSVM [15] in Weka [16]: ϵ-SVR with a radial basis function (RBF) kernel, degree = 3, cost = 10, eps = 0.2, loss = 0.1, normalize = true. Since our main interest was to see how relationships change under different scenarios, we did not tune these parameters further.

Results are reported from a 10-fold cross-validation scheme, in which all model parameters are learned using 9 of 10 data subsets for training, and accumulating results on the 10th (left-out) subset for estimating performance. We made sure that the same participant did not appear across different sets. To remove random effects, the reported results are the average of 10 random trials when applicable.

6. Experimental results

6.1. Data

During the data collection stage, the participants self-administered and completed a *d*MSE session on iOS devices every day for 3 to 6 days continuously. In 2016, we collected 227 valid sessions from 79 undergraduates enrolled in psychology courses at Louisiana State University, yielding 2.87 sessions per student. Students participated in return for extra course credit. We collected 118 valid sessions from 25 stable clinical participants with a range of serious mental illnesses, yielding 4.72 sessions per participant. More details about these clinical participants can be found in [17]. All the speech recordings were natural reactions: non-acted, spontaneous speech. They were recorded at 16 kHz and 16 bit resolution. Compared with previous studies based on speech data generated by acting, the variances of the different states for analysis here is relatively small. This paper focuses on these two different groups of populations. Here data are analyzed per session, rather than per participant.

6.2. Self-reported affect reliabilities

We checked the internal consistency reliabilities (Cronbach's α [18]) of the self-reported PA and NA, and their intercorrelations. When reporting results, we divide participants into two groups: Normative (undergraduate students), and Clinical, as in Table 2. Although we used a very different format (an iOS device with audio prompts and sliders compared with the traditional PANAS, which uses paper and pencil with a five-point response format: from not-at-all to quite-a-lot/extremely) to present the PA and NA questions, and we included only half the number of traditionally presented items [4], our α reliabilities were nonetheless consistent with earlier findings [4, Table 2]. Note that intercorrelations between PA and NA are fairly high in our sample. Still, the internal consistency reliabilities are in the 'good' range for both PA and NA in the clinical data sets, and in the 'acceptable' range for both PA and NA in the normative data set. The higher reliabilities in the clinical data set may be caused by wider response ranges in that population (standard deviation: 24.6 vs 17.3 for PA and 28.4 vs. 17.6 for NA). Test-retest or intra-participant reliabilities in a one-day retest interval in Table 2 are comparable to similar data reported in [4, Table 3]. These reliabilities exhibit a significant level of stability of the self-reported PA and NA, and at the same time should be sensitive to fluctuations in mood, which supports the use of PA and NA self-reports as targets for machine prediction.

Table 2: *The internal consistency reliabilities (Cronbach's α) of the self-reported PA and NA, and their intercorrelations.*

Group	n	α reliabilities		Intercorrelations
		PA	NA	
Normative	227	0.73	0.79	-0.56
Clinical	118	0.80	0.84	-0.49

Table 3: *The test-retest reliabilities of the self-reported PA and NA (one day retest interval).*

Group	The number of pairs	Test-retest reliabilities	
		PA	NA
Normative	148	0.66	0.47
Clinical	81	0.58	0.66

6.3. Observed affect reliabilities

The selected speech responses were rated by different human raters for their *Arousal* (or *Valence*) on 7 categories, with '0' representing silence, insufficient sample, or noncompliant due to a non-clinical distraction, '1' representing clinically/extremely low (or negative), and '6' representing clinically/extremely high (or positive). Ratings of '0' were excluded from data analysis. The detailed rating rubrics were designed by clinical professionals. The raters were four clinical professionals and five graduate students majoring in clinical psychology. Due to restrictions associated with schedules and cost, only a limited number of responses were rated.

From a sample of sessions from the normative group, 28 sessions were selected using stratified sampling such that sessions would be spread across the range of response styles, such as self-reported slider values and total number of words spoken. All speech responses of personal greeting questions, describing still pictures, describing silent videos, free talk, retell a story, retell the story a second time, and suggestions for improving the app were rated on average by four different raters for each rubric (arousal and valence). So each session has about 23 X 4 ratings per rubric (repeat responses were not rated). The average of these values per session per rubric provides reliable estimates of observed affect. The correlation between the averaged arousal and valence at the session level is 0.94. This may imply that for this normative population the difference between arousal and valence is not so obvious. After averaging, the ranges for arousal and valence were narrowed down to [2.7, 4.2] and [3.0, 4.0].

Sampling of clinical participants occurred at two different time periods. There are 55 sessions (17 participants) collected in the first period with responses from describing silent videos and free talk which were rated for arousal and valence. There are on average 14 ratings per session per rubric. These could be used to create stable estimates of arousal (Cronbach's α = 0.91) and valence (Cronbach's α = 0.85) for these sessions. The correlation between averaged arousal and valence in the session level is 0.66. After averaging, the ranges for arousal and valence were narrowed down to [2.0, 4.3] and [2.5, 4.3]. This spread is wider than the normative population. We noted it as ClinicalGroup1.

Later, two clinical professionals rated all clinical participants' free-talk responses for their arousal and valence. Their correlations at the response level: r(*Arousal*) = 0.64 (n = 201), r(*Valence*) = 0.43 (n = 198). The correlation between averaged arousal and valence is 0.50 (n = 213) at the response level, and 0.62 (n = 116) at the session level. After averaging, the ranges for arousal and valence were [1.0, 5.0] and [2.0, 4.5]. They are significantly wider than the previous two sets. We noted it as ClinicalGroup2. We used the average rating per response from these two raters as the prediction targets.

6.4. Self-reported affect vs. human ratings

We reported the correlations between self-reported affect and the observed arousal or valence ratings from speech by human raters in Table 4. They are generally low, ranging from -0.24 to 0.21. These correlation values indicate relative independence: self-reported affect seems independent of apparent affect observed in varied samples of speech. This finding is consistent with the result in [6, 7] although their self-reported emotion ratings were directly tied to the speech materials produced. In the three data sets, there is a mismatch between apparent (heard) emotions and self-reported emotions. Note that the session-level PA and NA scores in Table 4 are very reliable: arousal reliabilities in the three participant groups are 0.99, 0.91, 0.88. For valence, the score reliabilities are 0.98, 0.85, and 0.75.

Table 4: *The correlations between the self-reported affect and the observed affect from speech by human raters.*

Group	n	Arousal		Valence	
		PA	NA	PA	NA
Normative	28	0.08	0.21	0.15	0.15
ClinicalGroup1	55	0.06	0.08	0.14	-0.24
ClinicalGroup2	116	-0.01	0.07	0.02	0.02

6.5. Machine prediction results

6.5.1. Self-reported affect vs. speech features

We used the average of the PA (or NA) slider response values in a session as the prediction target for the responses of the session. All speech responses listed in Table 1 were used to predict the targets. We reported the correlations at the session level between the targets and the averaged predictions in Table 5. We checked the correlations between the targets and each individual feature and reported the best ones. These best values provide an indication of how strong these speech features are when they are used to predict the targets. The weak correlations in Table 5 further confirm the conclusions we arrived at in Subsection 6.4. When using all the different types of speech responses in combination (Table 1), the machine methods cannot predict the self-reported affect from the speech signal alone. We observe that in Table 5, overall, the clinical data set has better correlations than the normative one.

Instead of using all speech responses together to predict tar-

Table 5: *Using all speech responses together in a session to predict targets, correlations between self-reported affect and one best individual openSMILE speech feature; correlations between self-reported affect and SVR predictions using openSMILE features.*

Group	n	Prediction method	PA	NA
Normative	227	A best feature	0.14	0.14
		SVR	0.05	0.11
Clinical	118	A best feature	0.22	0.20
		SVR	0.26	0.29

gets, we checked each type to see if the responses from a certain type may predict the targets better. The results are reported in Table 6. For the normative set, all these correlations are in the negligible range. It is consistent with what we observed when using all speech responses together in a session to predict. Interestingly, for the clinical set the correlations become moderate when analysing by individual types. Specifically, the self-reported NA is easier to predict than the self-reported PA. Speech data from two types (*pataka, the Stroop test*) that did not produce meaningful linguistic content except for a few fixed syllables, such as the task *pataka* was used to measure alternating motion rate and sequential motion rate [12, 13], can produce moderate correlations when they were used to predict the self-reported NA.

In Table 5, the correlations from the best single individual feature for PA and NA when mixing different type responses together are significantly lower than the correlations obtained when treating different type responses separately. It could imply that treating different type responses separately may be a superior method.

Table 6: *Predicting with speech responses from single task types. Correlations between self-reported affect and single best individual openSMILE feature; correlations between self-reported affect and SVR predictions using openSMILE features*

Type	Target	Normative, n=227		Clinical, n=118	
		A feature	SVR	A feature	SVR
greeting	PA	0.19	0.22	0.25	0.40
	NA	0.16	0.29	0.24	0.46
picture	PA	0.24	0.29	0.35	0.41
	NA	0.18	0.19	0.32	0.45
video	PA	0.23	0.16	0.29	0.29
	NA	0.19	0.18	0.30	0.44
talk	PA	0.20	0.26	0.36	0.32
	NA	0.21	0.22	0.32	0.41
pataka	PA	0.24	0.09	0.34	0.30
	NA	0.19	0.10	0.34	0.41
retell	PA	0.23	0.27	0.41	0.42
	NA	0.18	0.27	0.35	0.41
re-retell	PA	0.27	0.22	0.41	0.31
	NA	0.30	0.15	0.36	0.44
repeat	PA	0.21	0.09	0.27	0.34
	NA	0.20	0.17	0.25	0.37
Stroop test	PA	0.25	0.12	0.35	0.43
	NA	0.23	0.14	0.35	0.47
suggestion	PA	0.17	0.27	0.27	0.23
	NA	0.19	0.16	0.27	0.40

6.5.2. *Observed affect vs. speech features*

We report our machine learning results for arousal and valence in Table 7. Valence was significantly more difficult to assess than arousal by machine methods. It could be because the variance of human valence ratings is significantly smaller. Human raters have difficulty judging valence from speech alone. The higher correlation in valence for the normative data set could be a result of human ratings not distinguishing very well arousal and valence (as evident by the intercorrelation of 0.94). A little lower values in ClinicalGroup1 could be because the rating distribution did not spread sufficiently. A better spread in ClinicalGroup2 shows that we can assess arousal in the clinical data set quite reliably. The assessment for valence can provide valuable information also.

Table 7: *The correlations between the machine predictions and human observed ratings for arousal and valence in the session level using 10-fold cross validation.*

Group	n	Arousal	Valence
Normative	28	0.70	0.67
ClinicalGroup1	55	0.54	0.36
ClinicalGroup2	116	0.72	0.43

We cross-validated the trained arousal and valence models, and so we trained a model on the normative data set and tested on the clinical data set, and vice versa. Results are reported in Table 8. These reasonable correlations further confirmed that machine modeling observed affect from speech is viable. The correlation 0.19 could be caused by a conflation of arousal and valence in the normative data or by a limited score range in the valence ratings for the normative data set.

Table 8: *Correlations between machine predictions and human observed ratings for arousal and valence at session level using cross-domain validation.*

Training set	Test set	Arousal	Valence
Normative	ClinicalGroup2	0.75	0.19
ClinicalGroup2	Normative	0.61	0.65

7. Result summary and conclusions

In real life, emotion is measurable in speech, behavior, subjective states and physiology, with measures showing modest convergence [19]. Self-reports of positive and negative affect states as measured in the *d*MSE mobile app provided two consistent and stable dimensions of affect, which show statistical properties very similar to those reported for the original instrument (PANAS [4]). Unlike many speech-affect research studies, in this data set the self-reports and speech data were collected in close time proximity. However, arousal and valence observed in speech still seems to reflect independent dimensions of emotions; which may pose a methodological issue when self-reports are the target of an acoustic-feature-based affect analysis. Even when we derived very reliable person-level measures of observed affect, the predictive relation to self-reported affect was slight, but machine scoring could match the listener data on apparent affect in the speech.

Analysis of both normative and clinical data sets suggests that arousal can be assessed quite reliably via speech, but valence is less evident in speech acoustics. Although the states from self-reported measures of affect may be difficult to observe via speech in the normative sample, we found evidence that affect is (not surprisingly) reflected in the speech of a clinical population, especially for negative affect states. Cross-domain validation results of the observed affect concludes that machine modeling of observed affect from speech is viable.

8. Acknowledgements

This project was funded by grant 231395 (2014 – 2017) from the Research Council of Norway awarded to Brita Elvevåg.

9. References

[1] M. Grimm, K. Kroschel, and S. S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *ICASSP*, 2007, pp. 1085–1088.

[2] M. Asgari, G. Kiss, J. van Santen, I. Shafran, and X. Song, "Automatic measurement of affective valence and arousal in speech," in *ICASSP*, 2014, pp. 965–969.

[3] Analytic Measures Inc., "dMSE," 2018, Computer Software. Apple App Store, Version 1.1.5. [Online]. Available: https://itunes.apple.com/us/app/dmse/id1134008913?mt=8

[4] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the PANAS scales," *Journal of Personality and Social Psychology*, vol. 54, pp. 1063–1070, 1988.

[5] J. R. Crawford and J. D. Henry, "The positive and negative affect schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample," *British Journal of Clinical Psychology*, vol. 43, pp. 245–265, 2004.

[6] C. Busso and S. S. Narayanan, "The expression and perception of emotions: Comparing assessments of self versus others," in *Interspeech*, Brisbane, Australia, 2008, pp. 257–260.

[7] K. Truong, D. van Leeuwen, and F. de Jong, "Speech-based recognition of self-reported and observed emotion in a dimensional space," *Speech Commun.*, vol. 54, no. 9, pp. 1049–1063, 2012.

[8] E. Rosenfeld, D. Massaro, and J. Bernstein, "Automatic analysis of vocal manifestations of apparent mood or affect," in *Proceedings of the 3rd international workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, Firenze, Italy, 2003, pp. 5–8.

[9] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[10] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Interspeech*, 2009, pp. 312–315.

[11] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech*, 2013, pp. 148–152.

[12] F. L. Darley, A. E. Aronson, and J. R. Brown, *Motor speech disorders*. Philadelphia: WB Saunders, 1975.

[13] J. R. Duffy, *Motor speech disorders: Substrates, differential diagnosis, and management*. St. Louis, Mo: Elsevier Mosby, 1995.

[14] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, 2013, pp. 835–838.

[15] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[17] A. S. Cohen, T. L. Fedechko, E. K. Schwartz, T. P. Le, P. W. Foltz, J. Bernstein, J. Cheng, T. B. Holmlund, and B. Elvevåg, "Ambulatory vocal acoustics, temporal dynamics and serious mental illness," 2018, manuscript submitted for publication.

[18] L. J. Cronbach, "Coefficient alpha and internal structure of tests," *Psychometrika*, vol. 16, pp. 297–334, 1951.

[19] M. M. Bradley and P. J. Lang, "Measuring emotion: Behavior, feeling, and physiology," in *Cognitive Neuroscience of Emotion*, R. D. Lane and L. Nadel, Eds. New York, NY: Oxford University Press, 2000, vol. 25, pp. 242–276.