DATA STATEMENT FOR SOLO CORPUS

A. Curation Rationale

The SOLO Corpus comprises over 4 million English tweets, each of which contains at least one of the following tokens: *solitude*, *lonely*, and *loneliness.* The corpus has been collected to analyze the language and emotions associated with the state of being alone in English tweets.

The state of being alone can have a substantial impact on our lives, though experiences with time alone diverge significantly among individuals. On the one hand, loneliness—a negative and unwanted state of being alone—has been shown to be correlated with increased cognitive decline, dementia, depression, suicide ideation, self-harm, and even death. On the other hand, solitude—a positive and self-driven state of being alone—has been shown to improve autonomy, creativity, and well-being. The SOLO corpus allows carrying out large-scale computational analyses of the language of being alone to validate psychological theories and get new insights into the relationship between the state of being alone and our well-being.

After consulting with psychologists who specialize in the study of solitude and utilizing different thesauri, a list of words and short phrases related to the state of being alone was created. It included the following terms: *alone*, *alone time*, *aloneness*, *confinement*, *desert*, *detachment*, *get away from it all*, *get away from people*, *hermit*, *isolation*, *loneliness*, *lonely*, *lonesomeness*, *me time*, *peace and quiet*, *privacy*, *quarantine*, *reclusiveness*, *retirement*, *seclusion*, *separateness*, *serenity*, *silence*, *solitariness*, *solitude*, *tranquility*, *undisturbed*, *wilderness*, *withdrawal*. Tweets were collected using these query terms for a few weeks, and then manually checked for their relevance to the topic of interest. Some query terms (e.g., *solitariness*, *reclusiveness*, *lonesomeness*, *aloneness*, *get away from it all*) were rarely used on Twitter and, therefore, were discarded. Some terms (e.g., *silence*, *privacy*, *retirement*, *desert*) were often used in other senses, not related to the state of being alone. After this manual inspection, three terms were kept: *solitude* and *loneliness* (nouns), and *lonely* (adjective).

These three words, *solitude*, *lonely*, and *loneliness,* were used as query terms to collect tweets by polling the Twitter API from August 28, 2018 to July 10, 2019. Duplicate tweets, short tweets (containing less than three words), and tweets with external URLs were discarded. Further, only up to three tweets per user account were kept. This minimized the impact of prolific tweeters and bots on the corpus.

The SOLO corpus consists of three sub-corpora, one for each query term (*solitude*, *lonely*, and *loneliness*). Each sub-corpus is released as a list of tweet IDs. In this way, if a user deletes their tweet at any point of time, the tweet will no longer be accessible.

B. Language Variety

The data was collected via Twitter API with the language option set for English; therefore, any variety of English recognized by the Twitter language identification tool as English can be present.

C. Speaker Demographic

No direct speakers' demographic information is available. Over 3 million user accounts are included.

According to Statista, Twitter users worldwide tend to be male, between the ages of 18 and 49. The United States of America has the most users. According to Pew Research Center, in the US, Twitter users are younger, more highly educated and have higher income than the general public. However, the SOLO corpus was collected using specific query terms and restricted to English-language tweets. Thus, its user demographics might differ from the general Twitter demographics.

D. Annotator Demographic

There are no annotations.

E. Speech Situation

The SOLO corpus was collected between August 28, 2018 and July 10, 2019. The tweets mostly represent informal, spontaneous, asynchronous written language. The intended audience is friends and followers of the user or the general Twitter audience. Each tweet is limited to 280 characters.

F. Text Characteristics

The tweets in the SOLO corpus mostly describe the writer's or other people's experiences of being alone, provide general statements about positive and negative aspects of being alone, offer support, or cite relevant quotes from literary sources. A small percentage of tweets (~6%) do not refer to the state of being alone. In these tweets, the query word (*lonely*, *loneliness*, *solitude*) is used as part of a title (of a book, song, etc.) or a name (of a place, a stadium, etc.).

G. Recording Quality N/A

H. **Other** N/A

I. **Provenance Appendix** N/A