Building An Annotated Corpus of Illocutionary Acts on Twitter

Laura Ryals, Georgetown University

1. Introduction

The expansive use of modern social media platforms suggests that these services succeed in satisfying some element of users' interpersonal and communicative needs. While the question of what that element is undoubtedly has a multi-faceted answer, one way of tapping into it is to ask a smaller question: what are users *doing* on social media? In particular, this study focuses on Twitter, investigating what Twitter users are "doing" linguistically through the lens of speech acts. To this end, we are building a corpus of Twitter data, segmented and annotated for a variety of factors related to speech acts.

2. Hypotheses

The primary hypotheses for this study were inspired either by the medium of Twitter and its peculiarities, or by theoretical proposals in the literature of illocutionary acts.

Twitter and its peculiarities

1) We will see a speech act landscape for Twitter that is distinct from that of other social media and spoken conversations.
2) There is a distinctive linguistic style to Twitter. That is, if we look at how all of the categories compare against how long the original poster has been on Twitter, we will see a convergence on a certain style.
3) The speech act profiles of unincorporated tags vs. the profiles of segments that incorporate tags will be distinct, suggesting that "incorporated vs. unincorporated" is pragmatically meaningful.
4) Tagging vs. non-tagging hashtags will have different speech act profiles, suggesting that this distinction is pragmatically meaningful.

Theoretical proposals

5) Certain overt elements available to Twitter users function as direct illocutionary force indicating devices (IFIDs, from Searle and Vanderveken (1985)). Candidate elements include lexical choice, capitalization presence and/or choice, punctuation presence and/or choice, emoji presence and/or choice, tag presence and/or choice (e.g. # vs. @, tagging vs. non-tagging, incorporated vs. unincorporated), unit position within the tweet, and status as an original post vs. reply post. This hypothesis will be tested through assessment of which, if any, of these elements are good predictors of a given force within the corpus.
6) Certain overt elements available to Twitter users signal the presence and/or choice of an indirect force. Candidate elements include all of those listed in the previous bullet, as well as direct force. This hypothesis will be tested through assessment of which, if any, of these elements are good predictors either of the mere presence of indirect force, or of the presence of a specific indirect force.

3. Building the Corpus

Building the corpus will ultimately involve five basic steps.

3.1 Data Collection

The raw data was gathered by hand, with the author collecting tweets in batches, and then re-collecting those tweets 24 hours later in order to also gather any favorites, retweets, and/or replies to the original tweet. The data was collected in 17 batches, over the course of 8 weeks in July-September 2016. The result of the data collection was 1018 Twitter conversations (i.e. threads). Most of these conversations consist of only the original post (no replies). Including all original posts and all replies, the data collection yielded a total of 1,316 tweets.

3.2 Motivating a Segmentation Scheme

Because a tweet may contain multiple speech act units, a set of guidelines for dividing the data into those units was necessary. The segmentation scheme was developed on both theoretical and practical grounds.

On the theoretical side, the author referred to classical speech act theorists (e.g. Austin, 1962; Searle, 1969; Bach and Harnish, 1979) as well as previous studies of speech acts in text to decide which syntactic segments should be considered basic units of speech act performance. For example, complement clauses were not considered separate segments because they complete the propositional content of the matrix unit. Meanwhile, relative clauses, subordinate clauses, and other parenthetical elements were considered segments separate from the surrounding text. Also, coordinated or conjoined elements were only separated if these elements were IPs or above.

On the practical side, the scheme included concessions for how to handle noisy or ambiguous syntax. For example, sentential fragments were mentally reconstructed into "full" clauses before proceeding with the remainder of the guidelines. Additionally, coordinated or conjoined elements that were ambiguous between IPs and VPs were separated by default, with the stipulation that annotators would be given the option to say that such units should actually be considered one unit.

The scheme also addressed what to do with technological affordances such as hashtags, @-tags, and emojis. For example, all tags and emojis could be either incorporated or unincorporated. Unincorporated elements were

considered separate segments, while incorporated elements were considered part of whatever greater unit they were contributing to semantically. For unincorporated emojis, multiple of the same emoji were considered one unit (e.g. 😀😀😀), while a sequence of different emojis were each considered a different unit (e.g. 😀🎃🦄 would be three separate units). In the case of a sequence of different emojis, annotators will have the option to say that they should actually be considered one unit together, in the case that they are perceived to form a semantic unit together (e.g. if the annotator thought that 😀🎃🦄 stood for something like "happy about your pumpkin unicorn").

3.3 Segmenting the Data

The author segmented the data by hand, according to the segmentation scheme.[1] Situations falling outside of the scheme, or about which the scheme was ambiguous, were noted and linked to the relevant tweet with the goal of transparency. Recurring notes were distilled into a list of post hoc clarifications on the original segmentation scheme, which the author used as the guide for a second pass of the segmented data, to ensure consistency. The segmentation yielded over 3,500 segments, about 30% of which consisted of solely emoji, an @-tag, or a hashtag.

3.4 Motivating an Annotation Scheme

The author is currently in the midst of step four, developing an annotation scheme for classifying each speech act unit along various relevant dimensions. The decision of which information will be annotated (and how) will be driven by three motivations:

Theoretical motivation

Which version of speech act theory one adheres to affects the annotation in several ways. First of all, how the task is framed, for the annotator, depends on the chosen definition of "illocutionary." Additionally, each theory motivates a different taxonomy, depending on how it expresses the relationship between content and force.

Comparability motivation

Seeing as one of the goals of this study is to see how the speech act profile of Twitter compares with other forms of CMC and spoken conversation, it is important to keep in mind what the studies of other media have done so that my results can be compared.

Practical motivation

Making sure that the task is accessible and intuitive to the average native speaker is very important, given the hope of using Amazon Mechanical Turk to complete the annotation.

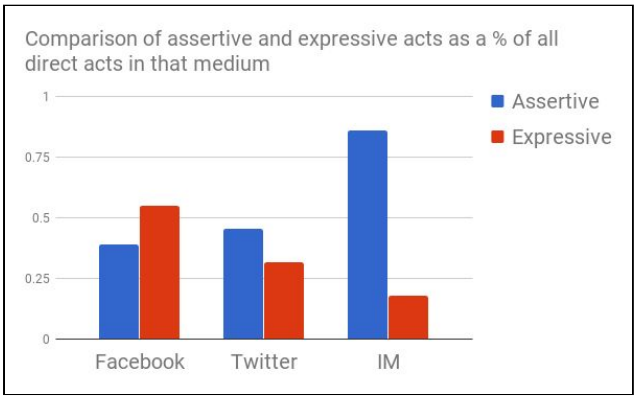
[1] An attempt to use Amazon Mechanical Turk to perform the segmentation was unsuccessful. Even with a short practice segmentation that eliminated Turkers with an insufficient knowledge of syntactic terms, the results were not satisfactory.

3.5 Annotating the Data

The author intends to use Amazon Mechanical Turk to perform the annotation. However, if a test run indicates that the annotation scheme is not satisfactorily accessible to an untrained worker, the author will instead hire and train 2-3 assistants to perform the annotation.

4. Preliminary Results

While the final annotation has yet to be completed, we have access to some preliminary results through a pilot study as well as the data collection and segmentation. For example, regarding hypothesis 1, the following chart compares the most frequent categories from Twitter (as per the pilot), Facebook status messages (Carr et al, 2012), and instant messenger away messages (Nastri et al., 2006) (to the extent possible, given methodological variation).



5. Goals and Contributions

This project has two primary goals, corresponding to its main contributions.

5.1 Theoretical Contribution

The first goal is to provide a descriptive snapshot of the speech act landscape of Twitter, a snapshot which can then be used to evaluate relevant theoretical questions and hypotheses such as those in section 2.

5.2 Practical Contribution

The second goal is to make the finalized corpus available to the academic community at large, along with the segmentation and annotation schemes used, as tools or references for future research.

6. Selected References

Carr, C.T.; Schrok, D.B. and Dauterman, P. (2012). Speech acts within Facebook status messages. *Journal of Language and Social Psychology*, pp.1-21.

Nastri, J.; Pena, J. and Hancock, J.T. (2006). The construction of away messages: A speech act analysis.. *Journal of Computer-Mediated Communication*, 4, pp.1025-1045.

Searle, J.R. and Vanderveken, D. (1985). *Foundations of Illocutionary Logic.* Cambridge, MA: Cambridge University Press.