Left-Handed or Right-Handed? A Data-Driven Approach to Analysing Characteristics of Handedness based on Language Use

Ho-Gene Choe
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109, USA
hgchoe@umich.edu

Rada Mihalcea
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109, USA
mihalcea@umich.edu

Abstract

Numerous studies have identified differences between left-handed and right-handed people, especially in the fields of psychology and neuroscience. Using a social media setting, this paper presents a data-driven approach to explore whether a person's handedness can be identified given his or her writing, and shows handedness characteristics that can be inferred from language.

Introduction

There exists a long history of research in the fields of psychology, neuroscience, and economics that worked on identifying characteristics or correlations based on handedness. Perhaps due to the preconceived notion on left-handedness rooted from the origin of the word "left" or "sinister," numerous findings showed unpleasant correlations with left-handed people, including prevalence among psychotic disorders (Webb et al. 2013), higher negative emotional valence (Propper et al. 2010), and lower wages compared to right-handed people (Goodman 2012). Other research claimed a rather positive side of being left-handed, especially among males, such as being more creative (Coren 1995) and earning more after college education (Ruebeck, Harrington Jr, and Moffitt 2007).

A stream of work in brain lateralization, which has to do with functional specialization of the brain, suggests possible differences in the use of language based on handedness. While language processing is dominated by the left-hemisphere of the brain for most people, higher percentage of left-handed people exhibit right-hemisphere language dominance compared to right-handed people (Knecht et al. 2000). Moreover, some studies related the interplay between language and motions to emotions based on handedness, where the dominant-hand gestures were associated with positive-valence speech (Casasanto 2011). Another study went beyond simple correlation to show that movement of dominant or non-dominant arm affects evaluation of neutral words to be positive or negative, respectively (Milhau, Brouillet, and Brouillet 2013).

Independent from this stream of research, recent studies in computational linguistics have been active in capturing a person's latent attributes given his or her writing. These studies are mostly based on results in psycholinguistics, which found significant differences in language use depending on diverse factors, including the writer's gender, age, and personality (Pennebaker, Mehl, and Niederhoffer 2003). To improve upon the small scale experiments in psychology research, many of these results were tested and confirmed on larger scale, through computational studies in social media settings (Nguyen et al. 2013), which demonstrate the feasibility of using social media to carry out similar kinds of observations. Despite these efforts and the aforementioned results in language lateralization, to the best of our knowledge, there have been no studies that directly relied on language use as a proxy to identify characteristics of left and right-handed people.

This paper presents a data-driven approach to explore differences between left-handed and right-handed people by analysing their language. Using a collection of tweets written by Twitter users identified as left-handed or right-handed, we explore whether a person's handedness can be identified given his or her writing, and examine whether there are any significant characteristics associated with handedness, that can be inferred from language.

Data Collection

At the early stage of the study, several data sources were considered, such as autobiographies of presidents and novels by famous writers, whose handedness is well known. We eventually decided to use Twitter to collect texts for left and right-handed people: despite the bias in user demographics, there are significant advantages that come with the use of social media, including a large number of people and a large number of texts.

Twitter data was collected in five steps: searching for tweets with selected keywords, handedness annotation, crawling annotated users' timeline, data cleaning, and additional gender annotation. Since there is no label indicating a Twitter user's handedness, we first searched the Twitter stream with several keywords that may contain relevant information to infer the writer's handedness. Search phrases included *"I am left handed," "I'm left handed," "I am leftie," "I'm leftie"* for left-handed users and *"I am right handed," "I'm right handed," "I am rightie," "I'm rightie"* for right-handed users. Between February 10–20, 2015, seed

@username I'm a left handed pitcher ;) ⇒ **Left-handed (techincal: sports or musical instrument)**

@username2 I'm right handed and I always keep my phone in my left pocket and use my left hand for it. ⇒ **Right-handed (minor/simple tasks)**

I'm not left handed but I always draw myself left handed ⇒ **Ambidextrous (technical: drawing, writing)**

I'm always that person that gets assigned to the left handed desk during 2 hour exams ://// ⇒ **Right-handed (inference)**

Figure 1: Sample Decisions on Seed Tweets

tweets containing the aforementioned search phrases were collected through the Twitter Search API.[1]

After obtaining the seed tweets, one of the authors manually verified the tweets and tagged each tweet's user as one of "left-handed," "right-handed," "ambidextrous," or "cannot identify." As handedness is defined as a continuous rather than a discrete variable (e.g., strong right, weak left), and is also dependent on the task at hand (e.g., write with left hand, but throw with right), a set of decision rules were used to tag each user consistently. Whenever possible, we tried to identify each user's dominant hand with the given information, ignoring uses of non-dominant hand for very simple tasks, as shown in the top two examples in Figure 1. When it was possible to identify that the user has a dominant hand but uses non-dominant hand for technical activities such as sports or drawing, the user was tagged as ambidextrous. Furthermore, when the user's dominant hand could be inferred from the tweet, it was used to tag the user as well.

Among these annotated users, tweets for users identified as left-handed or right-handed (excluding ambidextrous) were crawled through their timeline, starting from their most recent tweet and excluding retweets. Since some users wrote their tweets in more than one language, to restrict the analysis on English, users were selected for analysis only when they wrote their tweets in English. To select these users, 300 tweets were sampled from each user and each tweet's language was identified using the `langid.py` tool (Lui and Baldwin 2012). A user was selected for analysis when more than 75% of sampled tweets were identified as English.

Following previous work (Nguyen et al. 2013), we also restricted the analysis to "typical" users, and removed users with less than 300 tweets as well as users with more than 5000 followers. After this process, 358 left-handed users and 205 right-handed users were identified.

Additional gender annotations were also carried out for the users identified above, not only to normalize linguistic differences based on gender (Pennebaker, Mehl, and Niederhoffer 2003), but also to normalize possible differences in lateralization based on gender (Tomasi and Volkow 2012). As many Twitter users do not use their actual names for their username, we manually visited each user's timeline to annotate the gender if there is either one definitive evidence (e.g., explicitly identifying gender on their profile) or more than one indirect evidence (e.g., repeated uploads of one person's selfie). Table 1 presents the user distribution in our dataset. As shown in the table, despite all our efforts, the gender for about 30% of the users could not be identified.

	Female	Male	Unknown	Total
Left	153 (43%)	104 (29%)	101 (28%)	358
Right	87 (42%)	60 (29%)	58 (28%)	205

Table 1: User Distribution

Throughout the following analyses, subsets of users were randomly selected to balance the dataset. More specifically, when performing analysis between left and right-handed users as a whole (i.e., disregarding the gender label), 200 users from each class were randomly selected, and 300 tweets from each user were randomly chosen, resulting in 60000 tweets per each class. When performing gender-based analysis, 60 male and 80 female users were randomly selected from each class, each with 300 randomly selected tweets, resulting in 18000 male tweets and 24000 female tweets per each class.[2] We will refer to each dataset as `All` dataset and `Gender` dataset, respectively.

Predicting Handedness

In order to explore whether the user's dominant hand could be identified from tweets, we first perform a classification based on simple lexical features, including unigrams; length of tweets; part-of-speech tags (Gimpel et al. 2011); word classes based on Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Francis, and Booth 2001). Different preprocessing and feature selection mechanisms were tested, including stemming and chi-square feature selection. For evaluation, the classification performance was calculated by averaging test set accuracies during five-fold cross validation.

The classification is performed on two different levels of granularity: tweet-level and user-level. During tweet-level classification, although each tweet is treated as a single document, all tweets from a single user are ensured to be only in either training or test set, so that user specific characteristics are not used to boost the accuracy. Among several classifiers, random forest and logistic regression performed the best depending on the level of granularity, where the parameters for each classifier are learned through cross validation on the training set (i.e., four-fold cross validation on the training set for each run during the five-fold cross validation).[3]

Results are shown in Table 2, where separate classifications are performed on the `All` and `Gender` datasets. Although the accuracies are small, they are all significant with

[1] https://dev.twitter.com/rest/public/search

[2] Note that the tweets including the search phrases used to identify the handedness of the users are explicitly removed.

[3] All the experiments are carried out using Python and scikit-learn (Pedregosa et al. 2011).

Dataset	Tweet-level	User-level
`All`	0.523***	0.575*
`Gender`: Male	0.555***	0.642*
`Gender`: Female	0.528***	0.588*
Baseline	0.500	0.500

$^{***}p < 0.001$, $^{*}p < 0.05$

Table 2: Classification Accuracy Results

respect to the baseline, which suggests that there are indeed differences between the language used by left and right-handed people. While the user-level accuracies are higher compared to the tweet-level, they have lower degree of significance, due to a drastically reduced number of instances in the task (e.g., 60000 vs. 200 instances per class in case of `All`). Also, note that in both cases a higher accuracy is achieved for males.

In order to further ensure the significance of the results, we compare the original accuracy with the accuracy that could be obtained from classifying a random mix of users, in which users are randomly labeled as one of two different classes disregarding their handedness. This checks whether similar accuracies could be achieved by capturing any arbitrary hidden characteristics of users in each class. Best average accuracies obtained from such settings were significantly lower, with absolute differences of at least 0.02 for tweet-level and 0.05 for user-level as compared to the numbers reported in Table 2. This suggests that while it is difficult to predict a person's handedness from the language, there are some differentiating characteristics between left and right-handed users.

Linguistic Analysis

We also perform a linguistic analysis to determine if there are any characteristics of handedness that can be inferred from language use. We use lexical-based methods, using the LIWC (Pennebaker, Francis, and Booth 2001) and WordNet Affect lexicons (Strapparava and Valitutti 2004). LIWC includes about 2300 word-stems grouped into 70 different categories related to psychological processes, and WordNet Affect includes about 1100 words grouped into 6 basic emotion categories. For LIWC, as the lexicon contains both words and word-stems (e.g., "fail*" to cover "fails," "failed," etc.), word-stems were expanded to include all variations of such stems during the analysis.

To examine the differences between the left-handed and right-handed groups, a two-sample t-test was performed for each category in the lexicons. Since left-handed users appear to write more words per tweet on average compared to right-handed users (12.29 vs. 11.75 words per tweet), to normalize for the number of words, the percentage of categories used per each user is calculated and used in the analysis. For reasons of space, and since they are more insightful, we only report linguistic analyses performed on the `Gender` dataset.

Analyses comparing left-handed males to right-handed males are shown in the top part of Tables 3 and 4. Table 3 shows LIWC categories that had significant p-value in two-sample t-test, in the increasing order of p-value. It can be observed that right-handed males have a higher usage of language related to negative emotions, compared to left-handed males. Such observation is supported by the WordNet Affect analysis in Table 4. In Table 3, note also that left-handed males comparatively use more optimistic language.

Results for females are shown in the bottom part of Table 3. Compared to the results for males, fewer categories are identified as significant in the results for females. Table 3 shows that right-handed females have a higher usage of language that refers to other people, compared to left-handed females. No category in WordNet Affect is identified as significantly different for females.

Discussion

The classification results as well as the lexical-based analyses suggest possible differences between left-handed and right-handed people based on their language use. There seems to be more observable differences between males than females, which is in line with previous work in psychology that found stronger effects associated with handedness among males (Coren 1995; Ruebeck, Harrington Jr, and Moffitt 2007), and could possibly be explained by the higher degree of brain lateralization among males (Tomasi and Volkow 2012).

Across analyses, there is a strong association of negative emotional categories with right-handed males and some association of positive emotional categories with left-handed males. While it is difficult to concretely understand or explain this result, one speculation could be a possible emotional effect due to left-hand-oriented keyboard typing behaviors required for writing in online settings. In the dataset, the count of characters that require left-hand typing are about 10% higher than the count of characters that require right-hand, and a previous work suggests a bias in evaluation of neutral words to be positive or negative, depending on the movement of dominant or non-dominant hand, respectively (Milhau, Brouillet, and Brouillet 2013). While biased evaluation of neutral words does not directly account for using emotional words, it may be possible that a positive feedback mechanism is present to affect their language use. Experiments in constrained settings would be required to evaluate this speculation and to explain the causality of this result.

Conclusions

This paper presented a data-driven approach to explore whether there are any differences between left-handed and right-handed people in an online social media setting, as can be observed from their language use. While there are several limitations inherent to this study, most importantly having to do with the bias present in the user sample (e.g., given that users included in the study identified their handedness on the web, it is possible that the user sample is biased in some personality dimensions, such as how easily one reveals oneself), both the classification results and the lexical-based analyses showed possible indications of linguistic differences based on handedness, mostly in male users. Further experiments and analyses are required to concretely evaluate and understand the findings in this study.

Category	Sample Words	LH (Mean)	RH (Mean)
`Gender`: Male			
NEGEMO	bad, hate, wrong, sorry, miss, crazy, lost, weird, sad	2.377	**2.996**
ANGER	hate, damn, bitch, mad, stupid, kill, fight, dumb	1.224	**1.758**
SIMILES	like	0.466	**0.591**
SWEAR	shit, fuck, ass, damn, bitch, hell, dick, suck, piss	0.882	**1.351**
OPTIM	best, hope, win, free, top, ready, super, definitely, easy, won	**0.711**	0.607
NEGATE	not, don't, no, can't, never, didn't, nothing, doesn't, isn't, won't	2.069	**2.315**
SEXUAL	love, fuck, dick, gay, sex, pussy, nude, hug, kiss	0.694	**0.919**
`Gender`: Female			
OTHREF	you, your, he, they, we, she, her, them, someone, his	4.784	**5.321**
ARTICLE	the, a, an	**4.493**	4.145
YOU	you, your, ya, y'all, yourself, yours, you're, your, thee, thy	2.376	**2.761**

Table 3: LIWC Categories Used Differently by Left-handed (LH) and Right-handed (RH) Users (p-value < 0.05)

Category	Sample Words	LH (Mean)	RH (Mean)
`Gender`: Male			
ANGER	hate, mad, annoying, score, fit, pissed, angry, bother, jealous, evil	0.222	**0.299**
SADNESS	down, bad, sorry, sad, bored, blue, dark, poor, low, weight	0.339	**0.403**
DISGUST	sick, foul, horror, offensive, disgusting, disgusted, wicked, disgust, sickening	0.034	**0.056**

Table 4: WordNet Affect Categories Used Differently by Left-handed (LH) and Right-handed (RH) Users (p-value < 0.05)

Acknowledgments

This material is based in part upon work supported by the National Science Foundation (#1344257), the John Templeton Foundation (#48503), and the Jeongsong Cultural Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the John Templeton Foundation, or the Jeongsong Cultural Foundation.

References

Casasanto, D. 2011. Different bodies, different minds the body specificity of language and thought. *Current Directions in Psychological Science* 20(6):378–383.

Coren, S. 1995. Differences in divergent thinking as a function of handedness and sex. *The American journal of psychology* 311–325.

Gimpel, K.; Schneider, N.; O'Connor, B.; Das, D.; Mills, D.; Eisenstein, J.; Heilman, M.; Yogatama, D.; Flanigan, J.; and Smith, N. A. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, short papers*, 42–47.

Goodman, J. 2012. The wages of sinistrality: Handedness, brain structure and human capital accumulation.

Knecht, S.; Dräger, B.; Deppe, M.; Bobe, L.; Lohmann, H.; Flöel, A.; Ringelstein, E.-B.; and Henningsen, H. 2000. Handedness and hemispheric language dominance in healthy humans. *Brain* 123(12):2512–2518.

Lui, M., and Baldwin, T. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the Association for Computational Linguistics 2012 system demonstrations*, 25–30.

Milhau, A.; Brouillet, T.; and Brouillet, D. 2013. Biases in evaluation of neutral words due to motor compatibility effect. *Acta psychologica* 144(2):243–249.

Nguyen, D.; Gravel, R.; Trieschnigg, D.; and Meder, T. 2013. "how old do you think i am?"; a study of language and age in twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. AAAI Press.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71:2001.

Pennebaker, J. W.; Mehl, M. R.; and Niederhoffer, K. G. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54(1):547–577.

Propper, R. E.; Brunyé, T. T.; Christman, S. D.; and Bologna, J. 2010. Negative emotional valence is associated with non-right-handedness and increased imbalance of hemispheric activation as measured by tympanic membrane temperature. *The Journal of nervous and mental disease* 198(9):691–694.

Ruebeck, C. S.; Harrington Jr, J. E.; and Moffitt, R. 2007. Handedness and earnings. *Laterality* 12(2):101–120.

Strapparava, C., and Valitutti, A. a. o. 2004. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, 1083–1086.

Tomasi, D., and Volkow, N. D. 2012. Laterality patterns of brain functional connectivity: gender effects. *Cerebral Cortex* 22(6):1455–1462.

Webb, J. R.; Schroeder, M. I.; Chee, C.; Dial, D.; Hana, R.; Jefee, H.; Mays, J.; and Molitor, P. 2013. Left-handedness among a community sample of psychiatric outpatients suffering from mood and psychotic disorders. *SAGE Open* 3(4):2158244013503166.