Big-Five Backstage: A Dramatic Dataset for Characters Personality Traits & Gender Analysis

Marina Tiuleneva[1*], **Vadim Porvatov**[2], **Carlo Strapparava**[1]
[1]University of Trento, [2]University of Amsterdam
[1]Via Calepina 14, 38122 Trento, Italy,
[2]Science Park 904, 1098 XH Amsterdam, The Netherlands
[*]mari.tyuleneva@gmail.com

Abstract

This paper introduces a novel textual dataset comprising fictional characters' lines with annotations based on their gender and Big-Five personality traits. Using psycholinguistic findings, we compared texts attributed to fictional characters and real people with respect to their genders and personality traits. Our results indicate that imagined personae mirror most of the language categories observed in real people while demonstrating them in a more expressive manner.

Keywords: psycholinguistics, dataset, natural language processing, Big-Five personality traits, LIWC, MRC

1. Introduction

Can fictional characters be written so skillfully as to be indistinguishable from real people? Reading fiction opens up the inner worlds of the characters, their experiences and emotions, allowing the reader to take part in their life lessons, enhancing imagination and social competence (Boyd, 2009). It has been experimentally shown that reading different types of literary genres influences the social cognition of the reader (Kidd and Castano, 2013), (Heyes, 2018). One of the creative aspects of fiction that enables readers to immerse themselves in a character's perspective is the unique ability of the author to compose dialogue resonating with the authenticity of human speech. Previous research analysing a limited number of theatre plays written in verse (Ireland and Pennebaker, 2010) and movie scripts (Nalabandian and Ireland, 2022) has shown that certain authors can successfully imitate real people's speech, while others intentionally or not fail to do so (e.g., Shakespear's female characters speak like men; the same thing can be seen in Woody Allen movies).

The exploration of how fiction mirrors real-life speech finds its foundation in the distinct traits that individuals exhibit in their communication, both spoken and written. A gender-specific vocabulary-based approach has shown persistent differences in language use by males and females (Pennebaker and King, 1999). These findings were further confirmed on a big corpus of various types of texts (Newman et al., 2008): for example, women tend to use more emotional words and negations than men, and express thoughts, emotions and senses to other people. In contrast, men typically refer to external occurrences, objects, and processes, as well as utilize technical linguistic elements (numbers, articles, prepositions, and long words).

Further, the linguistic properties of texts written by people with different personalities have been extensively studied for the Big-Five taxonomy (Mairesse and Walker, 2007). This framework is centered around 5 major personality traits (Goldberg, 1990): *Extraversion* (EXT), *Neuroticism* (NEU), *Agreeableness* (AGR), *Conscientiousness* (CON), and *Openness* (OPN). For example, results show that informal speech is more common for extraverts than for introverts (*Hi* vs. *Hello*), neurotics more often use negative emotional vocabulary and first-person singular pronouns (*I*, *my*), conscientious individuals avoid negations, open people prefer longer words and vocabulary related to curiosity. These findings have also been confirmed for texts written on social media (Mewa, 2020).

Such discoveries from psycholinguistic research can be applied as a starting point for comparative analysis of authentic texts produced by real people and text written for fictional characters (Picca and Pitteloud, 2023). Studying imagined personae brings insights into the properties of separate works produced by the same author, whose intent is to mimic natural communication (Boyd and Pennebaker, 2015). Knowing patterns in the behaviour of fictional characters can give us a better understanding of sociocultural norms, and the extent to which it is possible for professional writers to imitate real-world speech.

The primary goal of this research is to evaluate the capability of authors to produce texts that convincingly mimic the speech of different genders and personalities. We focus on theater plays by internationally renowned authors as our primary source, as these plays rely on direct speech for character portrayal. Previous studies have mostly focused on movie scripts and have not investigated other narratives, such as those belonging to literary fiction.

The Workshop on Cognitive Aspects of the Lexicon, pages 114–119
20 May, 2024.

	EXT	AGR	OPN	NEU	CON
Acc.	0.872	0.889	0.894	0.791	0.9
Pr.	0.67	0.949	0.807	0.95	0.838
Rec.	0.88	0.862	0.852	0.663	0.94
F1	0.76	0.903	0.829	0.781	0.886

Table 1: GPT-3.5 performance measured by accuracy, precision, recall, and F1-score

		EXT	AGR	OPN	NEU	CON
M	0	0.4	0.26	0.44	0.37	0.35
	1	0.21	0.35	0.17	0.24	0.25
F	0	0.28	0.11	0.27	0.19	0.26
	1	0.11	0.28	0.12	0.2	0.13

Table 2: Big-Five personality traits distribution between male and female characters

In order to further boost research in this area and extend it to new domains, we prepared the Big-Five Backstage dataset comprised of fictional characters lines. To demonstrate its potential, we performed character analysis based on their genders and Big-Five personality traits. Character comparisons based on linguistic categories have shown that fictional males and females generally repeat language patterns observed in real people. The same trend can be seen for personality traits. Moreover, we found that specific language categories demonstrate a more drastic difference in imagined personae than in real people.

2. Data

2.1. Data Extraction & Preprocessing

The raw data consisted of 178 files containing theatre plays downloaded from the Project Gutenberg website. After having excluded non-English literary works along with those composed in verse, 400 theatre plays remained, written by 132 different authors. Next, we extracted the lines belonging to each character in the plays, and excluded the ones with fewer than 5 lines. The obtained text was normalized and tokenized with the help of the Stanza framework (Qi et al., 2020). The resulting dataset consists of 3 265 text samples corresponding to the concatenation of lines spoken by each character. Overall, it contains 3 419 136 words with a mean equal to 1047.2 words per character. The auxiliary part of the dataset includes author-level labels reflecting their gender, country of origin, and years of life.

2.2. Annotation Process

Each character was manually labelled as *male* (M) or *female* (F). For Big-Five personality traits annotation, GPT-3.5 (gpt-3.5-turbo) was applied to label each trait. These results were further compared with human annotations in order to validate predictions. The choice of GPT as an annotation tool was informed by prior research indicating the capacity of Large Language Models (LLM) to properly mimic the language of individuals across different personalities and genders, showing linguistic characteristics aligned with real human speech (Jiang et al., 2023). To achieve this, it is sufficient to create a prompt that clearly specifies certain personality traits and gender, according to which the LLM can generate text that reflects the linguistic patterns of specified traits according to the Five Factor Model (McCrae and John, 1992).

We sampled 10% of texts from the initial dataset and randomly distributed them among two human annotators. These annotators were blind to the GPT-3.5 responses and were subsequently instructed to read the texts and categorize them based on the Big-5 personality traits, marking a '1' if a trait was perceived in the text, and '0' if it was not. Definitions of the Big-5 personality traits were adopted from (Mairesse and Walker, 2007) (e.g., "Extraverts are sociable, assertive, playful, whereas Introversion is characterized by being aloof, reserved, shy."). Consequently, each text received annotations from one human and the GPT-3.5. We provide the following example that shows GPT-3.5 was not always able to correctly understand the intent of the characters. Here, in text id=94 GPT-3.5 misidentifies the Extraversion trait and labels the character as an introvert:

> **Text:** <...> the fact is, Dick, I'm lonely in this great house. You, who pass most of your time in a store, with people constantly coming and going, take pleasure in spending a week or two in a quiet place. <...>
>
> **GPT-3.5**: *Extraversion 0*. The person mentions feeling lonely in their house, indicating a lack of social interaction and possibly a preference for solitude.
>
> **Ground truth**: *Extraversion 1*. Experiencing loneliness does not imply a preference for solitude.

To evaluate the preliminary annotation's effectiveness, we assessed the performance of GPT-3.5 in terms of accuracy, precision, recall, and F1-score, as presented in Table 1. Although accuracy varied across traits, the lowest F1-scores were recorded for *Extraversion* and *Neuroticism*. Table 2 displays the final distribution of traits among the characters after validation by humans.

Additional information about the GPT-3.5 ablation study is available at the corresponding part of our GitHub repo.

3. Methods

We employ psychological findings that have identified connections between a range of linguistic markers (LMs) and genders (Pennebaker and King, 1999), as well as LMs and personality traits (Mehl et al., 2006). Linguistic markers represent clusters of words with a common characteristic, such as pronouns (*i*, *you*, *that*), prepositions (*to*, *of*, *in*), social (*you*, *we*, *he*, *she*) and cognitive (*but*, *know*) processes.

In order to study differences in texts of fictional men and women, we use the Linguistic Inquiry and Word Count (LIWC) dictionary (Boyd et al., 2022) and choose 44 LMs proposed in (Newman et al., 2008). This work analyses various groups of texts, finds LMs showing statistically significant differences between male and female writings, and presents Cohen's *d* coefficient (Cohen, 1992) obtained for each category of words. We compute frequencies of the LMs and compare Cohen's *d* of those showing statistical significance with Cohen's *d* calculated for real people. This analysis was performed for the whole dataset and for individual authors with at least 10 characters of each gender. On top of that, we extend the linguistic comparison to the characters' personality traits by applying MRC Psycholinguistic Database markers proposed in (Mairesse and Walker, 2007) in the addition to the mentioned subset of LIWC.

We took the results of statistical tests performed on real people texts from (Mairesse and Walker, 2007) and (Newman et al., 2008). In (Mairesse and Walker, 2007), the authors calculated Pearson's correlation coefficients between LIWC/MRC features and personality traits, while (Newman et al., 2008) provides the word frequencies and the effect size for the most common groups of words used by men and women.

During the analysis of the provided texts, we apply several methods from classic statistics: Mann-Whitney U test (Mann and Whitney, 1947) and Wilcoxon signed-rank test (Wilcoxon, 1945) for sample difference testing, Cohen's *d* coefficient to quantify the discovered differences, and point-biserial correlation (Lev, 1949) as a measure of dependency between LMs' frequencies and Big-Five personality traits considered as dichotomies. For all of the applied tests, we consider the level of significance α = 0.05.

4. Results

4.1. Genders

We performed the Mann-Whitney U test for male and female populations across the dataset and found 32 LMs that show statistical significance. Next, we calculated Cohen's *d* for these markers and compared them to real people. The difference in LMs between fictional characters repeats the patterns in the real world: men show a preference for long words (*BigWords*, >6 letters; *d*=0.33), prepositions (*d*=0.29), work-related vocabulary (*d*=0.23), numbers (*d*=0.2) and swear words (*d*=0.13), while women utilize language categories related to family (*d*=−0.36), home (*d*=−0.2) and social processes (*d*=−0.19), use pronouns (*d*=−0.25) and negations (*d*=−0.2). We also found LMs that show an opposite trend to the real world, which we call *reversed markers*. One such category, the second-person pronoun *you* (*d*=−0.1), occurs more often in fictional female speech, whereas in the real word it is used more often by men.

As shown in previous examples, the effect size for a number of LM's meets Cohen's *d* criteria for small ($0.1 \leq |d| < 0.3$) and medium effect ($0.3 \leq |d| < 0.5$), and for all of them it exceeds Cohen's *d* for real people, Figure 1(a). This indicates that there is an exaggerated difference for both men and women in fiction. Therefore, we continued our research focusing on individual authors, excluding those having fewer than 2 characters of each gender. We show the top-10 authors whose usage of LMs follows the same patterns as has been reported for real people, Figure 2.

For all the authors with at least one statistically significant LM, we calculated Cohen's *d* and did another comparison to real people. Thus, we confirmed the presence of an author-level exaggeration of gender-specific markers for males and females. In order to measure this effect, one can utilize the coefficients of a linear regression based on Cohen's *d* values for LMs, as shown for characters of August Strindberg, Figure 1(b). The slope of the linear regression line indicates the level of hyperbolization for both genders while the intercept sign demonstrates an imbalance in favor of females (negative) or males (positive). Conducted measurements on the sample of authors allow us to report that the mean value for the slopes is 4.5 with Q1=2.5 and Q3=5.5 while the mean value for the intercepts is −0.169 with Q1=−0.33 and Q3=−0.024. This indicates that the exaggeration is pronounced and slightly disproportional towards women.

4.2. Big-Five Personality Traits

We use 65 LMs (51 from LIWC, 14 from MRC) to analyze the linguistic differences in the personality

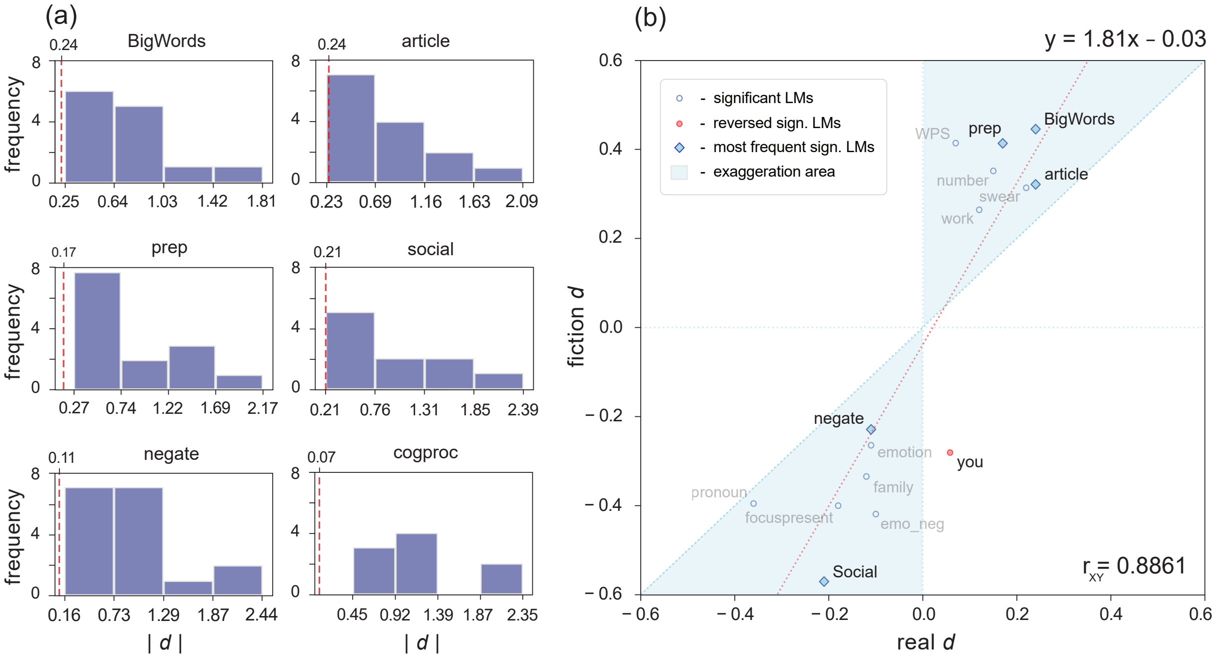

Figure 1: (a) Effect size distribution in 6 most frequent linguistic markers for the whole dataset: the red line shows Cohen's *d* for a corresponding LM in real people. (b) Example of author-level correlation between Cohen's *d* calculated for statistically significant LMs in real people and fictional characters. As Cohen's *d* is based on a mean difference of two samples, its positive values show that males used certain LM more than women, while the negative ones suggest the contrary.

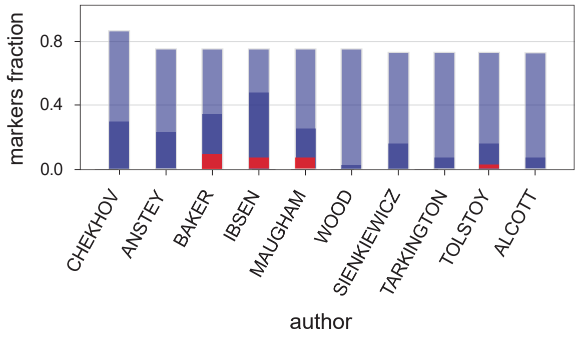

Figure 2: Fraction of linguistic markers indicating the difference between males and females (per author/ total LMs): real-world pattern of use (light blue), statistically significant (dark blue), reversed (red).

traits of characters. A significant point-biserial correlation was observed at least in one trait for all 14 MRC and 48 LIWC linguistic markers. We found 5 LMs showing statistical significance in all of the traits, Figure 3(a), and 24 LMs that are significant for 4 traits, Figure 3(b).

The strongest positive correlation is found for *Neuroticism* and LIWC markers corresponding to word count (*WC*) and linguistic categories related to affective vocabulary (*emo_neg*, *emo_anger*, *death*). In contrast, emotionally stable characters show a preference for punctuation marks (*AllPunc*). The pronounced dependency for *Conscientiousness* as a trait showing self-discipline was found in the case of the MRC summary variables: number of letters in one word (*NLET*) and number of phonemes (*NPHON*). Otherwise, unconscientious personae are typically depicted in plays by using excessive punctuation, such as exclamations, quotation marks, and non-fluent words (*nonflu*: *oh*, *um*). Similarly to real people, fictional extraverts communicate through vocabulary related to leisure, whereas introverts show a preference for long sentences. Agreeable characters tend to use positive emotional words (*emo_pos*), and their opposites rely on the negative ones (*emo_neg*, *emo_anger*). Finally, the presence of *Openness* correlates with Paivio's Meaningfulness (*MEANP*), spelling (*NLET*), and leisure. It also has the most discrepancies with the texts attributed to real people due to the largest number of reversed markers among the traits.

5. Limitations

This study employs GPT-3.5 as a tool for annotating Big-Five personality traits in textual data, complemented by the analysis of a single human annotator. A limitation of our methodology arises from the uncertainty surrounding the actual personality traits of the texts belonging to the fiction characters under examination. The ground truth cannot be established due to the nature of this data, and our

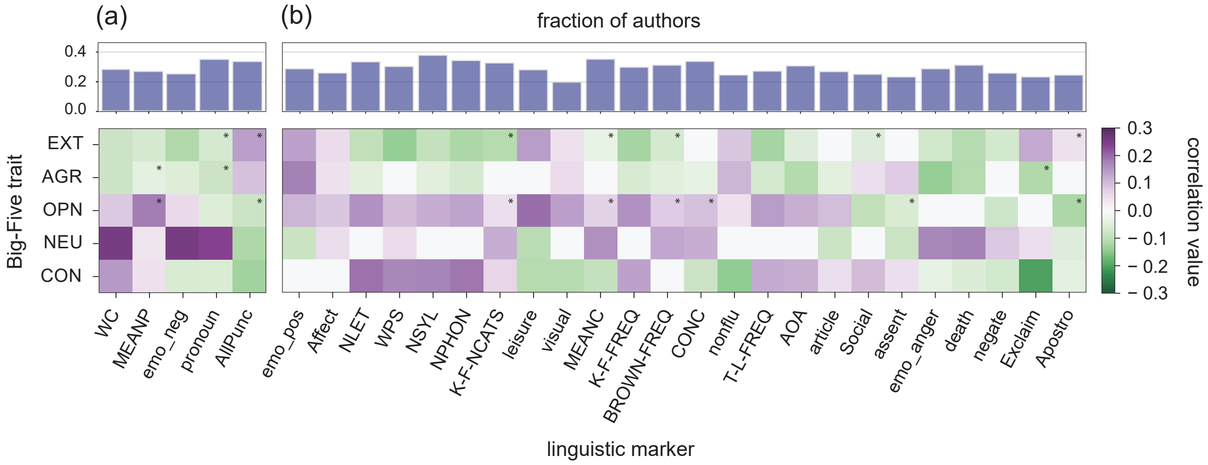

Figure 3: LMs demonstrating statistically significant correlation for 5 traits (a), and for 4 traits (b). The intersection of LM and trait represents a value of biserial point correlation between them; linguistic markers are sorted by the mean of correlations across all traits with asterisks denoting the reversed markers. The bar chart shows the fraction of authors that use a certain LM in at least one trait.

analysis operates under assumptions made by both the AI and human annotator based on the textual evidence available. Furthermore, the reliance on a single human annotator introduces potential biases and a lack of diverse interpretative perspectives that multiple annotators could provide. This limitation could affect the reliability and generalizability of the findings, as the interpretation of personality traits from text is subjective and may vary significantly among different readers.

6. Discussion

Our study presents a methodological framework that offers valuable insights not only into theater plays but also extends to real-life contexts. This framework has potential applications in various domains such as social media (for authorship attribution and detecting anomalous behavior) and cultural studies (exploring gender and social stereotypes, and analyzing authors through their characters). In the area of Human-Computer Interaction and robotics, our dataset and methodology could prove instrumental in assessing texts generated by large language models.

We have provided a statistical analysis of word usage shifts in theatrical texts across previously unexplored dimensions. While related research has focused on verse-based theater plays by a select group of authors (Ireland and Pennebaker, 2010), our study pioneers in examining the extent to which authors can replicate the speech of male and female characters and differentiate their characters from real individuals. Utilizing LIWC/MRC dictionaries, we observed that specific word categories correlate with certain personality traits, in line with prior studies (Mairesse and Walker, 2007), (Kosinski et al., 2013). Interestingly, our findings highlight a tendency among authors to overemphasize gender-specific vocabulary, particularly in depicting female speech. This suggests that while some authors successfully mirror real-world linguistic trends, others struggle to accurately represent these nuances in their characters.

We have also identified correlations between linguistic markers and personality traits, revealing dependencies for further investigation. For instance, emotive vocabulary is linked with Neuroticism, Extraversion, and Agreeableness, while summary variables can distinguish Conscientiousness, and specific punctuation usage is common among unconscientious and emotionally stable personalities.

Our findings underscore the challenge authors face in naturally replicating real speech patterns. Even when attempting to 'mimic' individuals of different genders, authors often exaggerate certain speech characteristics. Characters portraying various personality types exhibit more pronounced linguistic features than typically observed in real individuals. Our research invites further exploration into the nuances of generating speech that aims to mimic another's, whether by humans or machines.

In conclusion, our study demonstrates the potential of an automated approach for labeling Big-Five traits. Moving forward, we aim to delve deeper into the zero-shot capabilities of large language models in predicting personality traits, highlighting the need for more research in this area to refine and expand upon our promising results.

Acknowledgements

We thank Fredrik Hoeglund for help with the annotation process.

7. Bibliographical References

Brian Boyd. 2009. *On the origin of stories: Evolution, cognition, and fiction*. Harvard University Press.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, pages 1–47.

Ryan L Boyd and James W Pennebaker. 2015. Did shakespeare write double falsehood? identifying individuals by creating psychological signatures with text analysis. *Psychological science*, 26(5):570–582.

Jacob Cohen. 1992. Statistical power analysis. *Current directions in psychological science*, 1(3):98–101.

Lewis R Goldberg. 1990. An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Cecilia Heyes. 2018. *Cognitive gadgets: The cultural evolution of thinking*. Harvard University Press.

Molly E Ireland and James W Pennebaker. 2010. Language style matching in writing: synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology*, 99(3):549.

Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2023. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547*.

David Comer Kidd and Emanuele Castano. 2013. Reading literary fiction improves theory of mind. *Science*, 342(6156):377–380.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805.

Joseph Lev. 1949. The point biserial coefficient of correlation. *The Annals of Mathematical Statistics*, 20(1):125–126.

François Mairesse and Marilyn Walker. 2007. Personage: Personality generation for dialogue. pages 496–503.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

Matthias R Mehl, Samuel D Gosling, and James W Pennebaker. 2006. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, 90(5):862.

Tasneem Mewa. 2020. 'personality, gender, and age in the language of social media: The open-vocabulary approach'by h. andrew schwartz et al (2013). *Identifying Gender and Sexuality of Data Subjects*.

Taleen Nalabandian and Molly E Ireland. 2022. Linguistic gender congruity differentially correlates with film and novel ratings by critics and audiences. *PloS one*, 17(4):e0248402.

Matthew L Newman, Carla J Groom, Lori D Handelman, and James W Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse processes*, 45(3):211–236.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Davide Picca and Jocelin Pitteloud. 2023. Personality recognition in digital humanities: A review of computational approaches in the humanities. *Digital Scholarship in the Humanities*, page fqad047.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

8. Language Resource References

Qi, Peng and Zhang, Yuhao and Zhang, Yuhui and Bolton, Jason and Manning, Christopher D. 2020. *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. The Stanford NLP Group. PID https://github.com/stanfordnlp/stanza.