École Polytechnique Fédérale de Lausanne

Master's Thesis

A Large-Scale Analysis of Personal Reference Expression Usage in Public Discourse

Marko Čuljak

Akhil Arora
Thesis Advisor

Andreas Spitz
Thesis Advisor

Robert West
Thesis Supervisor

December 8, 2022

Acknowledgments

First, I would like to thank Akhil Arora and Andreas Spitz for teaching me the ways of scientific research, as well as for their invaluable and insightful feedback throughout my 1-year stay at EPFL. Next, I thank Robert West for making all of this possible and welcoming me to his lab.

My sincerest gratitude goes to my parents, Zvonimir and Milena, my grandmother Elizabeta, my overseas aunts Sandra and Marijana, and to my Swiss aunt Mila, all of whom provided me with substantial moral and financial support during my studies.

Lastly, I thank all my friends for the fun times during my studies, but most importantly, Ivan for transporting me to Lausanne safely.

Lausanne, December 8, 2022 Marko Čuljak

Abstract

The choice of a reference type humans use to refer to each other is influenced by their relation, attitude toward each other, and status differences. In this thesis, our goal is to characterize the use of personal references in public discourse. For this purpose, we propose Quotegraph, a novel social network extracted from Quotebank, a large corpus of attributed quotations, based on the speaker-mention relation. Additionally, from the same corpus, we extract CoQuotegraph and CoMentiongraph based on co-quotations in an article and co-mentions of persons in quotations, respectively, and analyze their structure. Our results show that in professional communication captured by the news quotations, women are less likely to be referred to by their last name and to use last names when referring to others than men. The obtained results are consistent across various domains and nationalities when a man and a woman are mentioned in the same quotation. Using full names instead of last names to refer to a person has previously been shown to impact one's perception of their status negatively. Therefore, due to the significance of the mass media in everyday human life, such bias in the news contributes to the gender gap in the perceived women's prominence.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Arguably one of the most influential books in human history, The Bible, in line with the creationist theories of human origin, brings up the concept of an almighty creator who uses words to create us and the world we live in and set it in motion. The first verse of the King James Version of The Bible highlights the significance of *the Word* for human existence. Although creationism dominated human history, the currently widely accepted theory is evolutionism which argues that humans have originated from simple structures and organisms. As humans evolved, they developed complex communication systems, which allowed humans to form and maintain complex social relationships, and store previously acquired knowledge. Therefore, one could argue that the emergence of words, i.e. the development of language, indeed marked the beginning of humans as complex social beings and, consequently, the modern world. Having developed language, in addition to assigning labels to objects, phenomena, and other living beings, humans also label each other by assigning personal names to themselves. Although names are usually not unique and can change throughout one's life, they allow for the identification of specific humans. Therefore, they are a key component of human social structure, and an important part of an individual's social identity [1]. However, in social interactions, the way humans refer to each other varies based on their personal relations, status differences, and attitude toward the person they refer to [2]. For example, siblings or romantic partners are likely to refer to each other by their first name, partly because they usually share the same family name. On the other hand, students typically refer to the professors using their title and last name. Lastly, Donald Trump, the former president of the USA, often used a derogatory term *Crooked Hillary* to refer to his rival Hillary Clinton during the 2016 Presidential Campaign. Therefore, we can conclude that the use of different personal reference expressions is an important signal to consider when analyzing human communication.

In this work, our goal is to characterize the use of personal references (first name, last name, full name, alias) in public discourse while considering the attributes of both the speaker and a mention. Starting from Quotebank [3], a large corpus of attributed quotations, we first create Quotegraph, a novel social network based on the speaker-mention relation. Next, by leveraging the link between Quotebank and Wikidata [4, 5], we extract gender, age, nationality, occupation, and party affiliation for all the nodes. Furthermore, the network structure of Quotegraph allows us to take into account the status, i.e. the prominence of a person

as approximated by PageRank [6]. Our studies primarily focus on gender-based differences in different occupations and cultures. We also briefly cover the evolution of gender-based differences through time as well as the differences based on party affiliation in US politics.

Atir and Ferguson [7] highlight the gender bias in using last names when referring to female and male professionals, finding that male professionals are more than twice as likely to be referred to by their last name as women. Furthermore, they find that the researchers referred to by their last name are perceived as more likely to receive career awards, thus arguing that this nominal gender bias can be detrimental to the perceived eminence of women. However, while Atir and Ferguson cover the issue extensively, they focus only on the use of the last names in a small-scale setting where a non-professional mentions a professional. Extending their work, we aim to investigate whether their findings hold on a large scale in the communication captured by news quotations while also studying the usage of first name references. Furthermore, Since each quotation in Quotebank is attributed to a speaker, we also cover the differences with respect to the speaker's attributes and emphasize the interaction rather than solely a mention. By leveraging the attributes of speakers and mentions, we study the interactions of professionals in various domains, thereby investigating whether bias exists in professional communication.

Contributions. In addition to Quotegraph, we propose two novel social networks extracted from the news, CoQuotegraph (edges connect speakers quoted in the same article), and CoMentiongraph (edges connect two persons mentioned in the same quote). Furthermore, we provide a large-scale analysis of personal reference usage in public discourse across different domains and nationalities. Lastly, we highlight the shortcomings of Quotebank and propose directions for improvements.

Chapter 2

Related Work

Gender differences in the use of personal references. Recently, Marjanovic, Stanczak, and Augenstein [8] conducted a large-scale study on Reddit focusing on gender biases when mentioning politicians. They found that the odds of a male politician being named by his surname are 8 times greater than for a female politician. On the other hand, the odds of a female politician being named by her first name are 15 times greater than for a man. However, they do not consider biographic features other than the mention's gender. Next, Atir and Ferguson [7] conducted eight studies focusing on gender biases in the use of the last name reference. The first four studies focused on the differences in the use of the last name between female and male mentions in various domains, including politics and science. In the subsequent four studies, the authors investigated the consequences of nominal gender bias, finding that the professionals referred to by surname are perceived as more famous and eminent. Their second study is closely related to our work. In this study, the authors studied the last name reference used in the transcripts of American radio programs, finding that, when discussing politics, pundits and other commentators are more likely to refer to male politicians by their last name than to female politicians. Lastly, studies by Halbert and Latimer [9] and Messner, Duncan, and Jensen [10] indicate that sports commentators use the last name reference more frequently when referring to men than women.

Social network extraction from textual data. A large body of research has been focused on the extraction of social networks from various types of unstructured and semi-structured textual data, e.g. emails [11, 12], scientific citations [13, 14] and collaborations [15, 16], and Wikipedia [17]. Additionally, there is a line of research focused on the extraction of the social network of characters from literary works [18, 19, 20] and historical records [21, 22]. However, to the best of our knowledge, news corpora have not previously been used for social network extraction.

Chapter 3

Network Extraction from Quotebank

In this chapter, we cover technical aspects of network extraction from Quotebank. Before describing the networks (Section 3.2) and their extraction process (Section 3.3), as a prerequisite, we describe our approach in Quotebank speaker disambiguation in Section 3.1 and highlight its necessity. Lastly, in Section 3.4, we describe the peculiarities of Wikidata attributes and their extraction.

3.1 Speaker Disambiguation in Quotebank

The need for speaker disambiguation. Quotebank is a corpus of quotation-speaker pairs. Each speaker in Quotebank is identified by their name, which leads to ambiguity because (1) speaker names are ambiguous, and (2) speakers can have multiple aliases.

- **Speaker names are ambiguous.** It is not always trivial to identify which person the name of its speaker refers to if we isolate the quotation from the context of the articles it appears in and consider only its content. For example, Michael Jordan can refer to a basketball player but also a computer scientist.

- **Speakers can have multiple aliases.** In Quotebank, the surface form of the first mention of a speaker in an article is considered as their name. However, speakers can have multiple aliases, and the first surface form of a speaker can differ from article to article. For example, the former president of the United States, Donald Trump, can be mentioned as President Trump in one article and Donald Trump in another.

  Although the existence of multiple aliases for the same person does not impact local attribution (i.e., attribution within a single quotation context), it affects the attribution once the speaker probabilities are aggregated over multiple contexts [3]. Thus, if President Trump and Donald Trump are considered distinct entities, the attribution probability to the former US president is always less than or equal

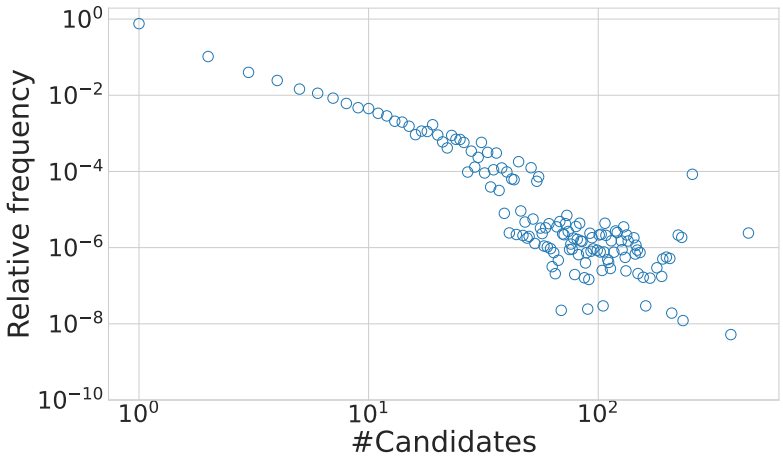


Figure 3.1: **Probability density function of the number of candidate Wikidata entities per person name in Quotebank.**

to the attribution probability obtained by considering them as a single entity. The proof of this statement is trivial. Let quotation q appear in k contexts. Let the name of an entity ε be A in m contexts and B in n contexts, $m+n=k$. Finally, let $P_l(A|q)=\{a_1,\ldots,a_m\}$ and $P_l(B|q)=\{b_1,\ldots,b_n\}$ be the sets of local attribution probabilities to q for A and B, respectively. The global attribution probability of q to A and B is calculated as $P_g(A|q)=\frac{1}{k}\sum_{i=1}^{n}a_i$, and $P_g(B|q)=\frac{1}{k}\sum_{i=1}^{n}b_i$, respectively [3]. If A and B are treated as distinct entities, $p_d := P_g(\varepsilon|q)=\max\{P_l(A|q),P_l(B|q)\}$, otherwise, $p_s := P_g(\varepsilon|q)=P_l(A|q)+P_l(B|q)$. Since the probabilities are non-negative, $p_s \geq \max\{P_l(A|q),P_l(B|q)\}$. Therefore, $p_d \leq p_s$. We can prove the statement analogously for any number of names per entity.

The presented drawbacks can significantly complicate Quotebank analyses and decrease the credibility of the results. Therefore, disambiguating person names by linking them to real-world entities is a crucial prerequisite for any type of automated Quotebank analysis, including but not limited to graph building (described in Section 3.3). Čuljak et al. [4] discuss the challenges of named entity linking in Quotebank and address them by proposing scalable and well-performing heuristics that link Quotebank names to their respective Wikidata entities. Thus, we use the proposed heuristics to disambiguate person names in Quotebank.

Quotebank ambiguity in numbers. Overall, approximately 76% of named entities in Quotebank are unambiguous. In general, as shown by Figure 3.1, names with a lower number of candidate Wikidata entities appear more often in Quotebank. However, we do not know how many unambiguous named entities correspond to correct Wikidata entities.

Article-level disambiguation. We use the proposed heuristics to disambiguate person names in Quotebank by linking them to their respective Wikidata entities. Specifically, we use the best-performing UIScore heuristic and optimize the proposed implementation by filtering out single-character words from

the entity representations. These tweaks lead to a 1.2 precision point gain on the same evaluation set. Additionally, by computing the entity representations in advance, we reduce the per-mention inference time by approximately 13 times on the same hardware as Čuljak et al. [4]. Finally, we parallelize the score computation using PySpark [23]. The overall score computation time amounts to approximately 130 minutes for all the unambiguous named entities in Quotebank on 20 Intel Xeon E5-2680 2.5 GHz CPU cores.

Quotation-level disambiguation. To attribute each quotation to the globally most probable Wikidata entity, we first set the local attribution probabilities of the entities to the previously computed local attribution probabilities of their respective names. Next, as proposed by Vaucher et al. [3], we sum the local probabilities over all the quotation contexts and attribute each quotation to the entity with the highest global attribution probability. It is important to note that for approximately 3.7M (non-unique) quotations ($\sim$ 0.9 %), the names listed in the quotation local probabilities field do not appear in the names field in article-level data. Therefore, in those cases, we cannot link the names to any disambiguated names in an article. Conveniently, the candidate Wikidata entities are still listed in these cases, so, for the sake of simplicity, we link such names to a Wikidata entity whose Wikidata item has the highest number of sitelinks (NS) [4]. Unfortunately, we could neither identify the cause of the issue nor fix it. However, considering the rarity of the issue, the good overall performance of the NS heuristic, and the high percentage of unambiguous names, we argue that the issue's impact on the overall attribution results is insignificant.

3.2 Network Definitions

Quotegraph, CoQuotegraph and CoMentiongraph have different properties due to the differences in their construction procedure. Therefore, before defining the networks and describing them on a high level, we first introduce the notation for key terms related to Quotebank. It is noteworthy that some design decisions we made while constructing the networks (Section 3.3) are not reflected in the network definitions since we aim to keep them as general and flexible as possible.

Quotebank preliminaries. Let A be the set of articles in the article-centric Quotebank, let Q be the set of quotations in the quotation-centric Quotebank, and let W_a be the set of all Wikidata entities mentioned in an article $a \in A$. Thus, $W = \bigcup_{a \in A} W_a$ is the set of all the entities mentioned anywhere in article-centric Quotebank. Assume that each quotation $q \in Q$ is either attributed to a speaker $w \in W$ or a NIL entity if the speaker cannot be determined. To formalize speaker attribution we define a function $s : Q \to W \cup \{\text{NIL}\}$[1]. Finally, let $W_q \subseteq W$ be a set of Wikidata entities mentioned in $q \in Q$ and let $Q_a \subseteq Q$ be a set of quotations appearing in an article a.

Quotegraph is a directed multigraph. For each $q \in Q$ attributed to a speaker $s(q) \neq \text{NIL}$, and for each $w \in W_q$, we construct a directed edge (s, w). Formally, Quotegraph is a pair $Q = (N_Q, E_Q)$, where $N_Q \subseteq W$ is

[1]Note that s is neither injective nor surjective since in Quotebank, some entities uttered multiple quotations, while to some other entities no quotations are attributed.

a set of nodes, while $E_q = \{(s(q), w) \mid q \in Q, s(q) \neq \text{NIL}, w \in W_q\}$ is a multiset of directed edges. Edges in Quotegraph correspond to quotations. Thus, since $\exists q \in Q$ such that $|W_q| > 1$, Quotegraph is a hypergraph. Essentially, Quotegraph is obtained by path contracting a directed bipartite graph $D = (N_Q \cup Q, E_D)$, where $E_D = \{(s(q), q) \mid q \in Q, s(q) \neq \text{NIL}\} \cup \{(q, w) \mid q \in Q, w \in W_q, s(q) \neq \text{NIL}\}$. Therefore, each edge E_B either points from a speaker to a quotation or from a quotation to an entity mentioned in it.

CoQuotegraph is an undirected multigraph. For each $a \in A$, we define a set of speakers $S_a = \{s(q) \mid q \in Q_a, s(q) \neq NIL\}$. For each $p, q \in S_a \wedge p \neq q$ we construct an undirected edge $\{p, q\}$. Therefore, CoQuotegraph is a pair $C = (N_C, E_C)$, where $N_C \subseteq W$ is a set of nodes, whereas $E_q = \{\{p, q\} \mid p, q \in S_a, a \in A\}$. Edges in CoQuotegraph correspond to articles in which the speakers that comprise the edges are co-quoted. Thus, since there can be more than two quoted speakers in an article, CoQuotegraph is a hypergraph. Analogously to Quotegraph, CoQuotegraph can be derived from an undirected bipartite graph $U = (N_Q \cup Q, E_U)$, where each edge $E_U = \{\{s(q), q\} \mid q \in Q, s(q) \neq \text{NIL}\}$ connects a quotation and its respective speaker.

CoMentiongraph is an undirected multigraph. For each $q \in Q$ such that $S(q) \neq NIL$ and $|W_q| > 1$, and for each $m, n \in W_q$ we construct an undirected edge $\{m, n\}$. Thus, QuoteCoMentiongraph is a pair $M = (N_M, E_M)$, where $N_M \subseteq W$ is a set of nodes, and $E_M = \{\{m, n\} \mid m, n \in W_q,\ q \in Q, S(q) \neq NIL,\ |W_q| > 1\}$. CoMentiongraph is conceptually similar to CoQuotegraph. Edges in CoMentiongraph consist of entities mentioned in the same quotation, while in CoQuotegraph, edges consist of entities quoted in the same article.

3.3 Network Construction

Preprocessing. Speaker disambiguation and quotation are performed using error-prone methods, which leads to noise in the obtained dataset. To ensure the good quality of the obtained networks, we aim to mitigate the noise as much as possible. Therefore, before network construction, we preprocess the quotations in Quotebank by (1) filtering short quotations and (2) grouping similar quotations.

1. **Short quotation removal.** We observed that short quotations are either generic quotes that can be correctly attributed to multiple people (e.g., *I love you.*) or only contain a person's name, a movie title or a book title. Additionally, in some cases, the same word is repeated multiple times within a single quotation (e.g., *Trump, Trump, Trump, Trump, Trump!*). Since the design of a robust solution that identifies such quotations is out of the scope of this work, to reduce the effect of the presented issues, we discard all the quotations with less than l_q case folded unique words, not considering punctuation and use $l_q = 5$ based on manual inspection.

2. **Grouping similar quotations.** As stated by Pavllo, Piccardi, and West [24], the same quotation can appear in slightly different forms or as a part of a longer quotation in different news outlets. Although one could argue that a high frequency of a quote in different articles can indicate its relevance, due to the noisy and heterogeneous nature of Quotebank, it is not clear whether this

assumption truly holds. Furthermore, different versions of the same quotation can be attributed to different speakers, thus leading to the emergence of non-existent interactions. Therefore, as proposed by Pavllo, Piccardi, and West [24], we group all the quotations that, when case folded, share the same substring of length at least $l_s = 8$ words excluding punctuation, and substitute them with the longest quotation in the group.

The preprocessing steps resulted in a corpus of 123M unique quotations attributed to 976k speakers. On the other hand, the unfiltered Quotebank contains 178M unique quotations attributed to 997k speakers. Thus, after applying the preprocessing steps, we remove 31% of quotes and 2% of speakers. Note that the number of unique speakers in the original dataset differs from the number presented (918k speakers) [3] as the original number is calculated on Quotebank before speaker disambiguation.

Quotegraph construction. We can break down the Quotegraph building procedure into seven steps: (1) determining quotation spans, (2) filtering spurious names, (3) filtering spurious mentions, (4) identifying the mentions inside quotations, (5) edge construction, (6) edge aggregation, and (7) removing loops.

1. **Determining quotation spans.** In the article-centric Quotebank, token spans that correspond to named entities mentioned in an article and the positions of starting tokens for each quotation are provided. Since neither quotation length in tokens nor the position of their ending tokens is provided, we first find ending tokens for each quotation in each article. Finding the positions of the ending tokens is trivial since Stanford CoreNLP [25] distinguishes between opening and closing quotation marks. Thus, to determine the position of the last token in a quotation, we iterate through all the tokens in an article, starting with the starting token of the quotation until we reach the closing quotation mark. As stated by Vaucher et al. [3], the innermost quote is extracted in the case of nested quotes. We rely on this design decision while finding the quotation ends and assume that the first closing quotation mark found does not mark the ending of any inner quotation.

2. **Filtering spurious names.** Although all the named entities mentioned in Quotebank are non-fictional humans, some real persons share names with fictional characters. For example, in Wikidata, there is an Australian rugby union player named Harry Potter (Q76164749) and a Canadian football player named James Bond (Q18377890). The named entity disambiguation heuristics designed for Quotebank would likely be capable of identifying the correct entity in each case, and we could safely remove fictional character nodes. However, Quotebank is designed so that all the speaker candidates correspond to real humans. Consequently, fictional characters Harry Potter (Q3244512) and James Bond (Q2009573) are not listed as candidate entities for their respective names. Thus, instead of designing a different candidate generation procedure, we discard all such names for simplicity. For the complete list of the names we ignore, please see Appendix A.

3. **Filtering spurious mentions.** Some entity names contain punctuation and stopwords. Due to the limitations of the system for speaker candidate extraction [24], if any entity alias contains such tokens, each of their occurrences in an article will be identified as a mention of the entity. To resolve

this issue, we do not consider (1) one-character tokens, (2) tokens with no alphabetical characters, and (3) stopwords as possible mentions for edge extraction.

4. **Identifying the mentions inside quotations.** Having defined the ending token for each quotation, the mentions inside quotations could be identified by comparing their token spans with the starting and ending offsets of the quotations. However, we found that the quotation offsets and name entity mention offsets are not always consistent. In Quotebank, the article content is represented as the concatenation of its tokens separated by the space character. The quotation spans correspond to the tokens obtained by splitting the article content by the space character. In contrast, the mention offsets are consistent with the tokens obtained by splitting the article content by any whitespace character.[2]

   Since identifying mentions inside quotations by comparing their respective token spans does not always yield correct results, we resort to a simple approach based on string matching. We first extract all the tokens of all the quotations in an article and concatenate them by the space character. Although quotation content is available in the article-centric Quotebank, it is provided in the detokenized form. For efficiency, and since detokenization is not injective,[3] we use already available article tokens instead of tokenizing the quotation content and extracting the tokens within each quotation based on the computed quotation spans. First, we concatenate the quotation tokens with the space character. Next, we extract all the tokens corresponding to the mentions of the named entities based on the mention spans listed in Quotebank. If a mention contains more than one token, we concatenate the tokens with the space character. We then look for exact matches of the mention string in the quotation string. If an article contains a pair of entities such that a mention of one entity is a substring of the other's mention, and both mentions can be found in the quotation, we simply discard the entity with a shorter mention.

5. **Edge construction.** Having identified the entity mentions in each quotation, as described in Section 3.2, we construct an edge from the Wikidata entity to which the quote is attributed after disambiguation to an entity appearing in the quotation.

6. **Edge aggregation.** Similar to speaker disambiguation, we aggregate the edges obtained in the previous step over all the quotation contexts by selecting the most common mention as the target node of the edge. Since a quotation may contain multiple named entity mentions, in such cases, analogously, we select the most common set of Wikidata entities mentioned in the respective quotations as the target nodes and create a distinct edge for each mentioned entity in the obtained set.

[2] See `https://en.wikipedia.org/wiki/Whitespace_character` for an overview of different whitespace characters

[3] Consider the following sentence: "The horse – which had been missing for days – suddenly returned to the pasture." Tokenizing this sentence using Stanford CoreNLP tokenizer would lead to the following tokens: "The", "horse", "–", "which", "had", "been", "missing", "for", "days", "–", "suddenly", "returned", "to", "the", "pasture", ".". In some Quotebank articles, no distinction is made between a dash and a hyphen, especially in the earlier phases of the data. In such articles, detokenizing the obtained tokens to the text: "The horse – which had been missing for days–suddenly returned to the pasture ." Finally, if we tokenize the obtained text again, we obtain the following tokens: "The", "horse", "–", "which", "had", "been", "missing", "for", "days–suddenly", "returned", "to", "the", "pasture", ".", which do not match the tokens obtained by tokenizing the original sentence, since "days", "–", and "suddenly" are concatenated into a single token.

7. **Removing self-loops.** Due to the limitations of quotation attribution in Quotebank, some quotations are erroneously attributed to the same persons mentioned in them. Thus, the edges derived from such quotations are self-loops. Although public personalities such as Zlatan Ibrahimović and Terry Crews often refer to themselves in the third person (illeists [26]), considering a large volume of such quotations attributed to distinct speakers (approximately 4.6 million quotes and 350 thousand different speakers), they are likely falsely attributed.

CoQuotegraph construction. The CoQuotegraph building procedure is significantly less challenging. The edge construction follows the description provided in Section 3.3. Similar to the Quotegraph construction procedure, before edge construction, we filter out all the spurious names according to the same spurious name list (Appendix A). Note that when analyzing CoQuotegraph, one must not ignore the properties of an article in which the speakers were co-quoted and the properties of a news outlet that featured the article. Thus, compared to Quotegraph, CoQuotegraph is better suited for outlet-level studies analyzing media bias rather than personal interactions.

CoMentiongraph construction. Once Quotegraph is built, we build CoMentiongraph by constructing an edge between all the entities mentioned in quotations where at least two entities are mentioned. CoMentiongraph, therefore, follows the same preprocessing procedure as Quotegraph.

3.4 Obtaining Node Attributes

In addition to solving the problems stemming from named entity mention ambiguity presented in Section 3.1, linking named entity mentions to their respective Wikidata items enables access to rich entity information stored in the form of statement-value pairs. This information substantially increases the value of our analyses as we are not limited to just abstract network-based features. Thus, by utilizing the information stored in Wikidata, we can analyze the interactions in the proposed networks from a person-centric perspective and put the findings into a social context. While extracting Wikidata statement-value pairs is not challenging, it is crucial to consider the cases in which multiple values are listed for a single statement and decide how to handle them. This section describes how we handle such cases for Wikidata statements that can hold multiple values. **Date of birth (P569).** While a person may have multiple dates of birth listed as a part of their Wikidata item, this is the case for only 0.1% of entities in all the extracted networks. Thus, in those cases, we extract the first listed birth date.

Nationality (P27). A person can have multiple nationalities, so we extract all the nationalities listed and do not consider countries that no longer exist, such as the Socialist Federal Republic of Yugoslavia or the Soviet Union.

Gender (P21). For gender, we consider three categories: female, male, and other. We deem an entity to be female or male if only female (Q6581072) or male gender (Q6581097) is listed in their gender statement. If either a non-binary gender or multiple genders are listed, we label gender as other.

Table 3.1: **Top-level occupations used to define respective domains in the left column.**

Domain	Top-level occupation
Art	artist (Q483501), creator (Q2500638)
Politics	politician (Q82955), lawyer (Q185351)
Sport	sportsman (Q50995749)

Political party affiliation (P102). A person can switch their political party throughout their lives. In this case, we assign a party to an entity based on the date of a quotation and Wikidata information about the starting and ending time of the party affiliation. If this information is not provided, we select the last party listed as it is likely the most recent one. For example, suppose only a year or a month of the start or the end of the party affiliation is provided. In that case, we extract the party an entity is affiliated with at the earliest possible date that meets the provided information. For example, if a person switched their party in 2008, we assume that the switch happened on January 1st, 2008.

Occupation (P106). We split the occupations into domains based on the Wikidata occupation hierarchy modeled by the *subclass of* property (P279). We define one or more top-level occupations for each domain and consider all the occupations below them in the hierarchy to belong to the specified domain. We focus on art, sport, and law & politics domains and label other occupations as other. The domain to top-level occupation mapping is shown in 3.1. Since art can take up many forms and can be performed on various levels of commitment, we deem an entity an artist if their domain is only art and not politics or sport. For example, according to Wikidata, Donald Trump (Q22686) is considered an actor (Q33999) and a writer (Q36180), both of which belong to the artist (Q483501) occupation tree. We do not disentangle other overlapping domains.

Given name (P735) and family name (P734). In the case of multiple given or family names, we use all the values listed.

US Congress membership (P1157). Similar to Külz et al. [27], we consider a US politician as a member of Congres if their US Congress Bio ID (P1157) is listed in their Wikidata item, making no distinction between former and active members.

Chapter 4

Network Overview

This chapter provides a high-level overview of Quotegraph, CoQuotegraph, and CoMentiongraph. We first present the structural properties of the networks in Section 4.1 and compare them to the properties of other real-world social networks. Next, in Section 4.2 we analyze node features. It should be noted that the networks are extracted from Quotebank data collected throughout a span of 12 years. However, in this section, we view all the networks as static and leave the analysis of their temporal evolution for future work.

4.1 Structural properties

In Table 4.1, we present an overview of structural network metrics. Looking into the number of nodes and edges, CoQuotegraph dominates in scale. Therefore, we can conclude it is more likely that two people get mentioned in the same article than that one mentions another. This also explains the small size of CoMentiongraph compared to Quotegraph and CoQuotegraph. All the networks are organized into one large connected component and many small ones. While this property is common in real-world social networks [28], one might expect that the news landscape is more fragmented. However, it is important to consider the noisy network building process as a single error in quotation attribution or entity disambiguation can induce a non-existent edge which may connect two disjoint components. Thus, more research is required to prove or disprove that this phenomenon is truly a property of the persons' interactions in the news. All the networks have positive degree assortativity, their degree distributions are heavy-tailed (Figure 4.1, and 4.2), and have a high clustering coefficient, all of which are typical properties of social networks. Lastly, given the heavy-tailed nature of the edge-multiplicity distribution (Figure 4.3), we can conclude that most of the interactions are event-specific. In our case, many such interactions might also be a product of the noisy graph-building process. On the other hand, a small fraction of interactions occurs consistently in the news.

Table 4.1: **Structural graph metrics for all three extracted networks.** Mean degree of is computed on its total degree, i.e., by taking into account both indegree and outdegree. CC stands for connected components. In Quotegraph, CC refers to weakly connected components

Measure	Quotegraph	CoQuotegraph	CoMentiongraph
#Nodes	528k	444k	214k
#Edges	8.63M	25.56M	1.93M
#Unique interactions	4.50M	7.52M	1.04M
Mean degree	32.77	115.01	18.05
#CC	5651	2190	10162
%Nodes in largest CC	97.53	98.98	89.20
Degree assortativity	0.034	0.006	0.159
Global clustering coefficient	0.265	0.539	0.509

4.2 Node Features

In this section, we present the distributions of the node features in the networks. It is noteworthy that the presented distributions are calculated based on edge ends rather than individual nodes, i.e., the distributions are weighted by the total node degree. For example, 10 woman-woman interactions contribute to the count of females in the interactions by 20, while 10 man-woman interactions increase both the male and the female count by 10. Consequently, we normalize the feature counts by twice the number of edges, as according to the handshaking lemma, the sum of node degrees equals twice the number of edges. Given the heavy-tailed nature of the node degrees, we argue that computing the feature distributions in such a way is a more accurate representation of their coverage in the news than computing the feature distributions on individual nodes.

Nationality. In Table 4.2, we show the distribution of the nationalities in the networks. Overall, persons from the USA receive the widest coverage across all the networks, followed by other countries of the English-speaking world.

Domain. In Table 4.3 we show the domain distribution. Sport dominates in Quotegraph and CoMentiongraph, while politics dominates CoQuotegraph. Lower representation of politics in Quotegraph and CoMentiongraph, as opposed to its representation in CoQuotegraph, likely indicates that the political debate is more focused on the ongoing issues in the society rather than on specific persons. Thus, as a direction for future work, it would be interesting to investigate whether the focus of the political debate becomes directed at certain political actors rather than on social issues. On the other hand, sportsmen and coaches frequently mention their teammates after games, evaluate their performances, and sometimes even praise their rivals. As an example, we show a quotation attributed to the former manager of FC Barcelona, Ernesto Valverde (Q359118):

"Coutinho stepped up without Messi against Inter. He is a constant threat with his shooting, he can pop up at any moment, and he can play in between the lines."

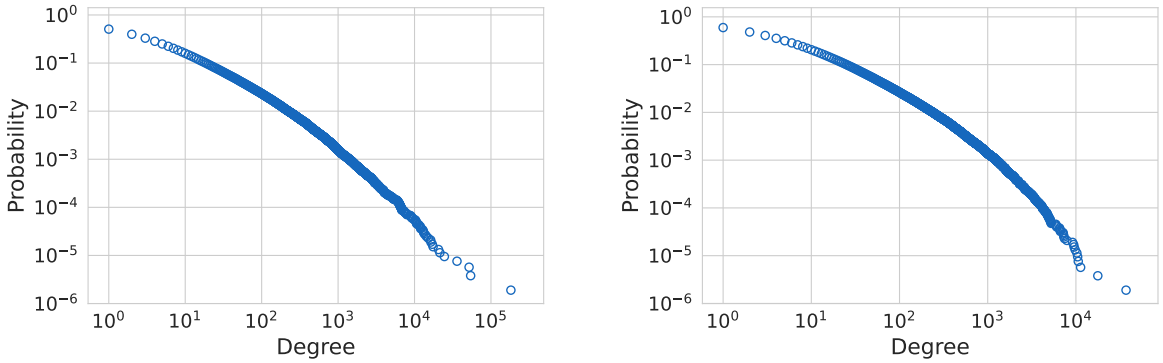
Age. The domain distributions are also reflected in the respective age distributions (Figure 4.4). In CoMentiongraph, we can see a clear peak around the age of 29, which can be explained by a high proportion of interactions involving sportspeople. The same peak is much less emphasized in Quotegraph and CoQuotegraph, in which sportsmen are less represented. On the other hand, in CoQuotegraph, we observe a peak after the age of 40, which can be explained by a high proportion of political interactions.

Gender. Table 4.4 reveals a clear gender bias in the news landscape. In Quotegraph and CoMentiongraph, men appear in approximately 87% of interactions and approximately 83% of CoQuotegraph. A slightly lower gender gap in CoQuotegraph may be explained by its lower coverage of sport which is generally known to be male-dominated.

Party. In Table 4.5 we show the party distribution in all three networks. In line with a high representation of political interactions in CoQuotegraph, a higher fraction of interactions revolve around entities affiliated with a political party than in the other two networks. We observe slight differences in party coverage in USA, India, and UK. A clear outlier is Conservative Party in the UK, which receives twice as much coverage as the opposing Labour Party.

Centrality. Although Quotegraph and CoMentiongraph are dominated by sportspeople, the top 10 nodes with the highest PageRank centrality are all politicians Table 4.6. This is not unexpected as the sport domain is more heterogeneous than politics since it is fragmented into various sports whose practitioners usually do not interact with each other. Comparing the most central nodes in the network, we argue that applying PageRank to Quotegraph or CoMentiongraph yields better results in the identification of key public figures. This is because the mention relation is better adapted to PageRank's definition of the centrality of a central node which can be summarized as *a node is central if it is pointed to by other highly central nodes that do not point to many other nodes*. If we view the definition of PageRank through the lens of the co-quotation relation, it would indicate that the nodes co-quoted with other highly central nodes are important. Thus, this definition may not highlight individual importance since the co-quotation does not correspond to the real interaction between two persons. While the same applies to the co-mention relation, the results of applying PageRank to the CoMentiongraph are similar to the results obtained on Quotegraph. We hypothesize that the similarity arises due to the construction of CoMentiongraph as it is obtained by path contraction of Quotegraph. In Table 4.7 we list the degrees of the nodes listed in Table 4.6.

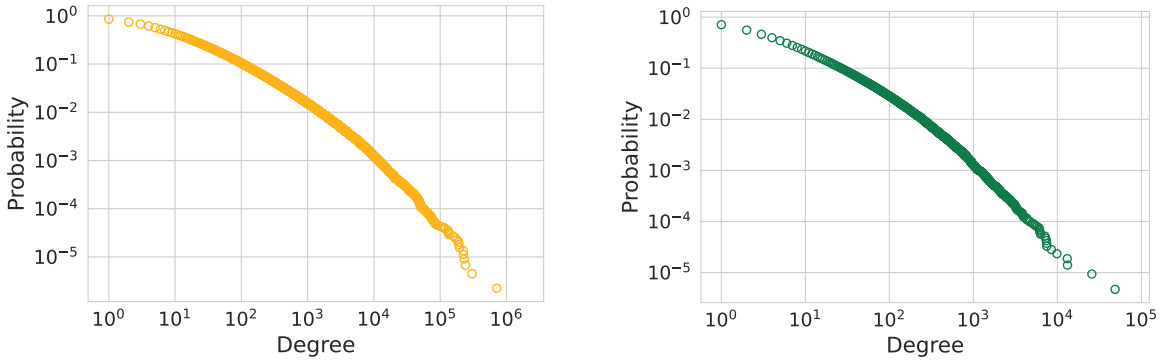
Assortative mixing. Lastly, using the definition of modularity proposed by Newman [29], we investigate the existence of assortative mixing by nationality, domain, gender, and party affiliation. As shown in Table 4.8, all the networks show high assortative mixing by nationality and domain and much lower assortative mixing by gender and party affiliation. Overall, the modularities in CoMentiongraph are the highest across all the categories, which, along with a high degree of assortativity, hints that persons mentioned in the same quotation are likely similar by their biographic features. We also observe that the UK is characterized by relatively high modularity by party in comparison to USA and India. However, we do not attempt to explain this phenomenon in this work.



(a) Indegree distribution.

(b) Outdegree distribution.

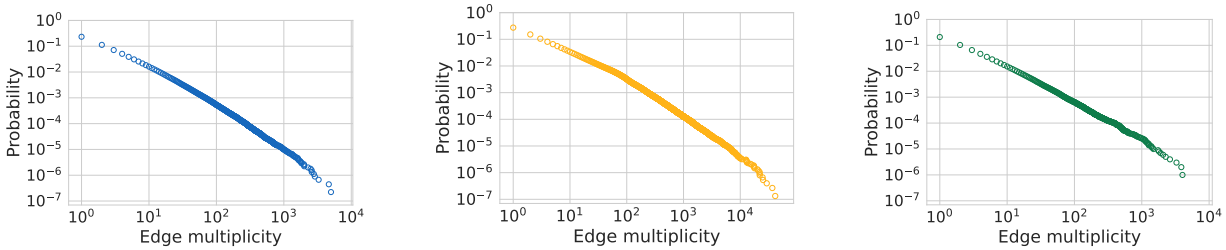
Figure 4.1: **Cumulative indegree and outdegree distribution in Quotegraph.** on the y-axis we show the probability $P(d \geq d_k)$ for a degree d_k on x-axis.



(a) CoQuotegraph degree distribution.

(b) CoMentiongraph degree distribution.

Figure 4.2: **Cumulative degree distribution in CoQuotegraph, and CoMentiongraph.** Probabilities are calculated as in Figure 4.1.



(a) Quotegraph edge multiplicity distribution.

(b) CoQuotegraph edge multiplicity distribution.

(c) Cumulative distribution function (CDF)

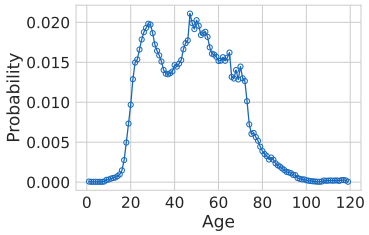
Figure 4.3: **Cumulative edge multiplicity distribution in CoQuotegraph, and CoMentiongraph.** Probabilities are calculated as in Figure 4.1, in this case modeling the probability $P(m \geq m_k)$ for an edge multiplicity m_k on x-axis.

Table 4.2: **Nationality distribution in Quotegraph, CoQuotegraph, and CoMentionGraph.**

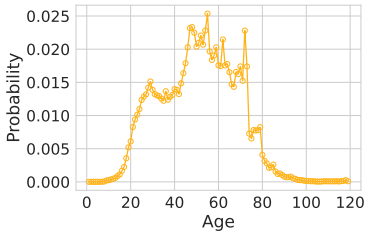
Nationality	Quotegraph	CoQuotegraph	CoMentiongraph
USA	48.47%	53.73%	49.31%
UK	11.99%	13.57%	12.09%
India	5.46%	2.61%	6.42%
Australia	4.33%	3.66%	4.15%
Canada	3.66%	3.57%	3.22%
New Zealand	1.25%	0.92%	1.31%
Republic of Ireland	1.09%	0.84%	1.14%
France	1.02%	1.21%	1.17%
Germany	0.96%	1.16%	0.89%
Italy	0.95%	0.75%	0.91%
Other	13.34%	11.62%	13.89%
None	5.69%	4.82%	3.31%

Table 4.3: **Domain distribution in Quotegraph, CoQuotegraph, and CoMentionGraph.**

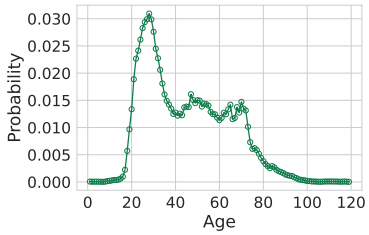
Domain	Quotegraph	CoQuotegraph	CoMentiongraph
Sport	35.12%	25.10%	40.59%
Art	31.95%	28.28%	32.26%
Politics	25.32%	39.22%	23.38%
Other	6.28%	6.31%	3.18%
None	1.33%	1.09%	0.58%



(a) Age distribution in Quotegraph.



(b) Age distribution in CoQuotegraph.



(c) Age distribution in CoMentiongraph.

Figure 4.4: **Age distribution in Quotegraph, CoQuotegraph, and CoMentiongraph.** y-axis represents $P(a = a_k)$ for an age a_k on x-axis.

Table 4.4: **Gender distribution in Quotegraph, CoQuotegraph, and CoMentiongraph.**

Gender	Quotegraph	CoQuotegraph	CoMentiongraph
Male	86.61%	83.29%	86.79%
Female	12.91%	16.50%	13.03%
Other	0.16%	0.19%	0.16%
None	0.33%	0.02%	0.02%

Table 4.5: **Political party distribution in Quotegraph, CoQuotegraph, and CoMentiongraph.**

	Quotegraph	CoQuotegraph	CoMentiongraph
Republican Party (USA)	8.90%	15.04%	8.96%
Democratic Party (USA)	7.75%	13.11%	7.88%
Bharatiya Janata (India)	1.03%	0.57%	0.89%
Indian National Congress (India)	0.77%	0.42%	0.73%
Conservative Party (UK)	0.70%	3.37%	0.60%
Labour Party (UK)	0.58%	1.68%	0.44%
Liberal Party of Australia (Australia)	0.37%	0.57%	0.29%
Australian Labor Party (Australia)	0.36%	0.30%	0.23%
Likud (Israel)	0.17%	0.33%	0.13%
United Russia (Russia)	0.16%	0.13%	0.19%
independent politician	0.37%	0.77%	0.46%
Other	5.99%	7.31%	3.88%
None	71.96%	55.23%	74.54%

Table 4.6: **Top 10 highly central nodes according to PageRank.**

Quotegraph	CoQuotegraph	CoMentiongraph
Barack Obama	Donald Trump	Donald Trump
Donald Trump	Barack Obama	Barack Obama
Hillary Clinton	Nancy Pelosi	Hillary Clinton
Mitt Romney	Mitch McConnell	Narendra Modi
Narendra Modi	Theresa May	Mitt Romney
Vladimir Putin	Hillary Clinton	Bill Clinton
Bill Clinton	Chuck Schumer	George W. Bush
John McCain	Boris Johnson	Vladimir Putin
George W. Bush	Joe Biden	Bernie Sanders
Theresa May	Sarah Sanders	John McCain

Table 4.7: **Degrees of highly central nodes.**

Node	Quotegraph	CoQuotegraph	CoMentiongraph
Donald Trump (Q22686)	299226	1771977	71083
Barack Obama (Q76)	231905	713812	48656
Hillary Clinton (Q6294)	65540	179768	25702
Narendra Modi (Q1058)	61152	84112	13196
Mitt Romney (Q4496)	53648	133086	13111
Vladimir Putin (Q7747)	27462	63805	7239
John McCain (Q10390)	26821	101939	7445
George W. Bush (Q207)	25814	46326	7402
Bill Clinton (Q1124)	25725	49773	9866
Bernie Sanders (Q359442)	24012	169274	8504
Theresa May (Q264766)	23526	240325	4347
Joe Biden (Q6279)	23035	195654	7470
Nancy Pelosi (Q170581)	17807	303020	6020
Mitch McConnell (Q355522)	14160	230059	4301
Boris Johnson (Q180589)	13279	224479	2924
Chuck Schumer (Q380900)	9628	222404	2629
Sarah Sanders (Q27986907)	3627	191177	384

Table 4.8: **Assortative mixing by nationality, domain and gender in Quotegraph, CoQuotegraph, and CoMentiongraph.**

Modularity	Quotegraph	CoQuotegraph	CoMentiongraph
Nationality	0.326	0.230	0.347
Domain	0.283	0.235	0.347
Gender	0.052	0.041	0.070
Party (USA)	0.024	0.070	0.124
Party (UK)	0.121	0.104	0.275
Party (India)	0.072	0.066	0.228

Chapter 5

Personal Reference Expressions in Quotegraph Interactions

In this chapter, we aim to characterize person referencing patterns in the interactions in Quotegraph, i.e. in the interactions covered in the news. Specifically, we study personal referencing patterns with respect to age, gender, and occupation of both the speaker and the mention. Additionally, by leveraging the network structure of Quotegraph, we study the impact of prominence as approximated by PageRank [6] on the reference type choices in the interactions. Although we are still constrained by the English language, to mitigate geographical biases, we do not focus only on the United States of America but conduct our analysis on the interactions in India and UK. We chose the UK and India for our studies because, in addition to satisfactory data availability (see Table 4.2), we hypothesize that, due to cultural differences, last name references would be less frequently used in those nations. In the UK, the members of the Royal Family frequently appear in the news, and they are usually referred to by their names and titles (e.g. Queen Elizabeth II or Prince of Wales). In contrast, in India, approximately 17% of the population shares one of the top 3 surnames.[1] We design our studies with the intention of answering the following research questions:

RQ1: Which entity features correlate with the usage of certain reference types?

RQ2: How does the reference type usage differ with respect to the speaker's and mention's gender?

RQ3: Can gender-based differences be consistently observed in different settings?

RQ4: How does the reference type usage differ with respect to the speaker's and mention's party alignment in US politics?

[1] Estimated based on `https://forebears.io/india/surnames` by summing the fractions in the *Frequency* column.

Table 5.1: **Reference type definitions and their abbreviations.** G and F are equal to one if a person's given or family name was used as their reference, respectively, and zero otherwise.

Reference type	Abbreviation	Definition
First name	FN	$G \cdot (1 - F)$
Last name	LN	$(1 - G) \cdot F$
Full name	FLN	$G \cdot F$
Alias	A	$(1 - F) \cdot (1 - G)$

We first define the reference types and entity features used in our analyses in Section 5.1 and Section 5.2. Next, in Section 5.3, we provide a high-level overview of the usage of the defined reference types in Quotegraph. Finally, we describe the conducted studies in Section 5.4 and present their results in Section 5.5.

5.1 Reference Type Definitions

We consider four disjoint reference types: (1) first name, (2) last name, (3) full name, and (4) alias. To clarify the definitions, we first define two indicator variables: G and F. $G = 1$ if any person's given name is used to refer to them in a quotation. Otherwise, $G = 0$. Similarly, $L = 1$ if any person's family name is used as their reference, and $L = 0$ otherwise. Based on the values of G and L, the definitions of the presented reference types and their abbreviations are given in Table 5.1. In our studies, we put less emphasis on the alias reference type due to its vague and heterogeneous nature and leave its detailed decomposition and analysis to future work.

5.2 Speaker and Mention Features

Biographic features. In all our studies, we focus on the person features presented in Table 5.2, along with their descriptions and abbreviations. In each interaction, we consider gender, PageRank, and age for both a speaker and a mention and name-related features (frequency and length) only for a mention. We limit our studies to female and male genders since non-binary entities are significantly less represented in Quotegraph (see Table 4.4). The interaction date corresponds to the year and month of the first utterance of the quotation from which a respective interaction is derived. When it comes to domains and nationalities, to simplify the analysis, we only focus on within-domain and within-nationality communication. Consequently, we view nationality and domain as interaction features[2] rather than as entity, i.e., person features. Suppose in an interaction, a speaker and a mention do not share the same domain, or the domain of either entity is labeled as *other*. In that case, we label the interaction domain as *other*. We proceed analogously with the nationalities.

[2]Not to be confused with interaction terms in regression analysis which model feature interactions.

Table 5.2: **Biographic and name-related features.**

Feature	Abbreviation	Description
Domain	D	factor with four levels: art, sport, politics, other
Nationality	N	factor with four levels: India, UK, USA, other
Interaction date	t	year and month of an interaction modeled as a discrete variable
Speaker gender Mention gender	G_s G_m	factor with two levels: female and male
Speaker age Mention age	a_s a_m	calculated at the time of the first utterance of a quotation
Speaker PageRank Mention PageRank	r_s r_m	calculated on Quotegraph
Last name frequency First name frequency	f_{LN} f_{FN}	calculated on unique Quotegraph nodes
Last name length First name length	l_{FN} l_{LN}	calculated as the number of characters

Name-related features. We consider the name length and frequency because we hypothesize that the usage of a name as a standalone reference is negatively correlated with its length and frequency in a population. To compute the respective name frequencies, we count all the given and family names listed for each unique node in Quotegraph. However, we use only the length and the frequency of the first given or family name listed in Wikidata as the respective mention's features.

Features specific to US politics. When focusing on US politics, we include features related to a person's political activity. Specifically, we consider a person's party affiliation and congress membership, as described in Table 5.3.

Table 5.3: **Additional features specific to US politics.**

Feature	Abbreviation	Description
Speaker party Mention party	P_s P_m	factor with two levels: Republican, Democrat
Speaker congress membership Mention congress membership	C_s C_m	factor with two levels: false, true

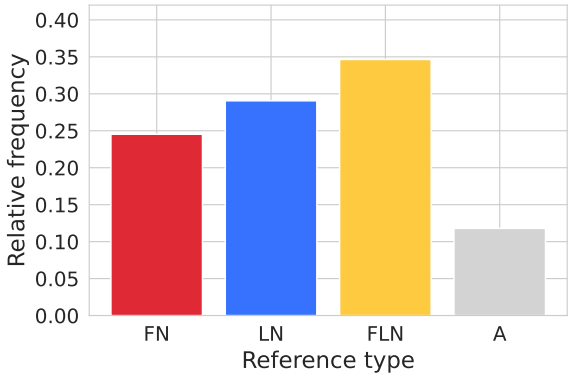


Figure 5.1: **Reference type distribution.** 95% confidence intervals cannot be seen due to the large sample size

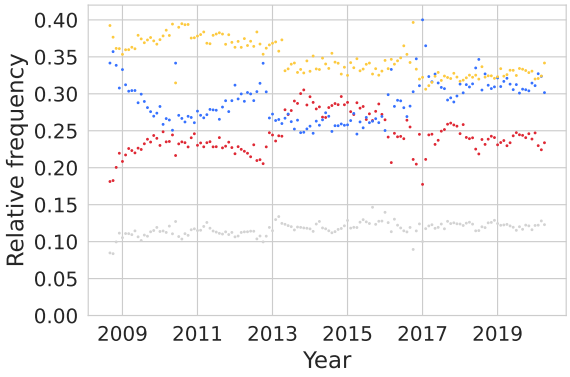


Figure 5.2: **Reference type distribution through time.** Colors represent the same reference types as in Figure 5.1

5.3 Reference Type Usage Overview

Additional preprocessing. Before diving deep into studying the usage of personal reference expressions, we provide an overview of the reference type usage in Quotegraph. It should be noted that before conducting our studies, we perform additional preprocessing steps. First, we discard all the edges in which the mention does not have a given or a family name as a part of their Wikidata item. Then, we discard all the mentions and speakers whose age is not the in the $[10, 110]$ interval at the time of the interaction. After preprocessing, approximately 39% of the Quotegraph edges are removed, resulting in a dataset of 5.28 million interactions.

Static and dynamic reference type distribution. In Figure 5.1 we show the general distribution of personal reference types. According to Mann-Whitney U-test, all the differences are statistically significant ($p < 0.001$). Overall, the Quotegraph interactions are dominated by the FLN reference, followed by LN and FN. Aliases are substantially less represented than the other reference types. In Figure 5.2, we present the evolution of the reference type distribution through time. We can notice a clear shift in the distribution starting with 2013. We found that the shift is due to an anomaly in the data and directly corresponds to the start of the Quotebank phase D.[3] In subsequent analyses, we analyze only the interactions that took place after the shift, i.e. we analyze only the data collected in the cleanest Quotebank data collection phases, resulting in a dataset of 3.30M interactions. Despite reducing the scope, the overall distribution does not change drastically, as shown in Figure 5.3. The only notable change can be seen in the relative frequency of the FLN reference, which drops from 0.346 to 0.330. However, it should be noted that, due to large sample sizes, all the changes are statistically significant ($p < 0.001$). Aside from the shift corresponding to

[3]Specifically, the shift occurs because in the phases A, B, and C, the percentage of self-loops is approximately 47% as opposed to 22% in the phases D and E. Furthermore, the mentions in the self-loops have fewer tokens on average. Since a mention of more than one token usually corresponds to a full name, the removal of self-loops, therefore, reduces the proportion of full name references in later phases. In other words, in early phases, a higher fraction of shorter mentions is removed, resulting in overall longer mentions on average as opposed to phases D and E. Additionally, due to the same phenomenon, the data in phases A, B, and C is less representative since we remove close to 50% of the data by excluding self-loops.

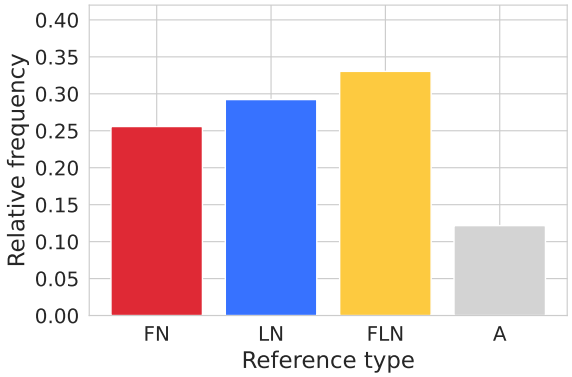


Figure 5.3: **Reference type distribution in the phases D and E of Quotebank data collection.** 95% confidence intervals cannot be seen due to the large sample size.

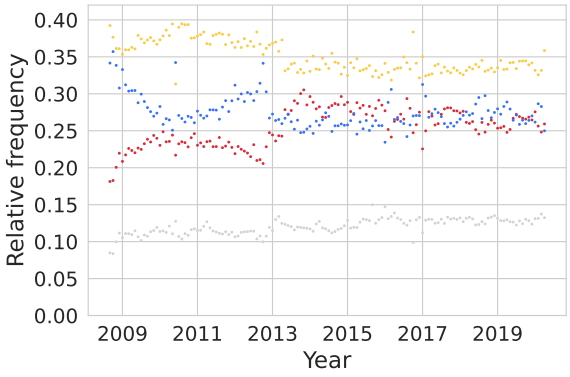


Figure 5.4: **Reference type distribution through time excluding the mentions of Donald Trump.** Colors represent the same reference types as in previous figures.

the transition from Quotebank phase C to phase D, another shift in the reference type distribution can be seen at the beginning of 2017. This shift can be explained by the rise to prominence of the United States president Donald Trump (Q22686), who is most commonly referred to as just *Trump*. Indeed, after excluding all of Trump's mentions, the shift is no longer visible.

5.4 Study Design

RQ1: Which entity features correlate with the usage of certain reference types? As a foundation for subsequent studies, we first aim to understand the correlations between different features and certain reference types. Therefore, we fit a logistic regression model using features described in Section 5.2 as predictors for each reference type. We fit each model on a binary outcome, which equals one if the respective reference has been used in a quotation. Before fitting the models, we preprocess the predictors. We model all the factor variables as binary indicators and set male as the default gender, the USA as the default nationality, and politics as the default domain. Furthermore, we rank-transform PageRank and name frequencies and map the resulting values to the $[-0.5, 0.5]$ interval. The coefficients corresponding to the features rank-transformed in the described way, therefore, indicate the difference in the outcome when the lowest ranked value of the feature increases to the highest ranked value. Next, we subtract 50 from the age features, 5 from the first name length, and 6 from the last name length. Additionally, since we don't expect that the usage of certain reference types would drastically change with the increase in speaker's or mention's age, we divide each age feature by 10. Finally, we model interaction date as a discrete variable. The value of zero in the date variable corresponds to the first month in the dataset (May 2013) and is increased by one for every subsequent month. Since we do not expect the person's reference usage to change drastically in one month, we divide the date by 12 to make the corresponding model coefficient indicate the yearly change. The intercept of a model fit on the predictors processed as previously described, therefore captures the average usage of a respective reference type in an interaction

that took place in May 2013, where both the speaker and the target are male 50-year-old US politicians with a moderate prominence as approximated by PageRank. In the same interaction, the mention's first and last name is moderately frequent while their first and last names are 5 and 6 characters long, respectively. The following equation illustrates the model fit for each reference type:

$$\begin{aligned}\log p_r - \log(1-p_r) = \beta_0 &+ \beta_1 D_{art} + \beta_2 D_{sport} + \beta_3 D_{other} \\ &+ \beta_4 N_{India} + \beta_5 N_{UK} + \beta_6 N_{other} \\ &+ \beta_7 G_s + \beta_8 G_m + \beta_9 a_s + \beta_{10} a_m + \beta_{11} r_s + \beta_{12} r_m \\ &+ \beta_{13} r_m + \beta_{14} f_{LN} + \beta_{15} f_{FN} + \beta_{16} l_{LN} + \beta_{17} l_{FN},\end{aligned} \tag{5.1}$$

RQ2: How does the reference type usage differ with respect to the speaker's and mention's gender? Next, we look closely at the differences in the reference type usage with respect to the speaker's and the mention's gender. For simplicity and clarity, we refer to *interactions that correspond to all the combinations of speaker and mention genders* as *gender interactions*. Starting from the model illustrated in Equation 5.1, we first add an interaction term for G_s and G_m and analyze gender interactions globally.

RQ3: Can gender-based differences be consistently observed in different settings? After analyzing gender-based differences globally, we first analyze gender interactions in sport, art, and politics domains by adding an interaction term for every domain and gender-based attribute (G_s, G_m, and G_sG_m). By adding analogous interaction terms, we study gender interactions with respect to different nationalities (USA, UK, India) and their evolution through time. We summarize the interaction terms added to the original model (Equation 5.1) for the studies focusing on gender interactions with respect to domain, nationality, and time in Table 5.4. Finally, quotations can be gathered in various settings, e.g. press conferences, phone interviews, or hallway conversations with the reporter [30]. Since communication styles may vary in different settings, we aim to examine whether gender-based differences remain when the context is controlled for. Therefore, we designed a matched study in which the reference types were used to refer to persons of different gender mentioned in the same quotation. To perform matching, we leverage CoMentiongraph. The edges we consider for the analysis must meet the following conditions:

- The edge must be derived from a quotation where exactly two persons are mentioned.
- In each edge, there has to be exactly one man and one woman.
- In each edge, persons must not share the same last name.

By imposing those conditions, we obtain a dataset of 101k interactions in which the mention's gender is balanced with respect to the speaker of the quotation and the context in which the quotation was uttered.

RQ4: How does the reference type usage differ with respect to the speaker's and mention's party alignment in US politics? In our final study, we briefly cover the personal reference usage in US politics, intending to investigate whether the choice of a personal reference differs with respect to the speaker and mention's party. We hypothesize that, since FN reference indicates a close personal relationship, it should be used more often when the speaker is a member of the same party as the mention than if the opposite is true. In other words, our goal is to find whether *first-name polarization* exists in US politics. Therefore, we

Table 5.4: **Interaction terms added while studying gender-based person differences in different domains and nationalities and their evolution through time.** We present the interaction terms in their factorized form

Study	Interaction terms
Domain	$(\gamma_1 D_{art} + \gamma_2 D_{sport} + \gamma_3 D_{other} + 1) \cdot (\gamma_3 G_s + 1) \cdot (\gamma_4 G_m + 1)$
Nationality	$(\delta_1 N_{India} + \delta_2 N_{UK} + \gamma_3 N_{other} + 1) \cdot (\delta_3 G_s + 1) \cdot (\delta_4 G_m + 1)$
Time	$(\varepsilon_1 t + 1) \cdot (\varepsilon_1 G_s + 1) \cdot (\varepsilon_2 G_m + 1)$

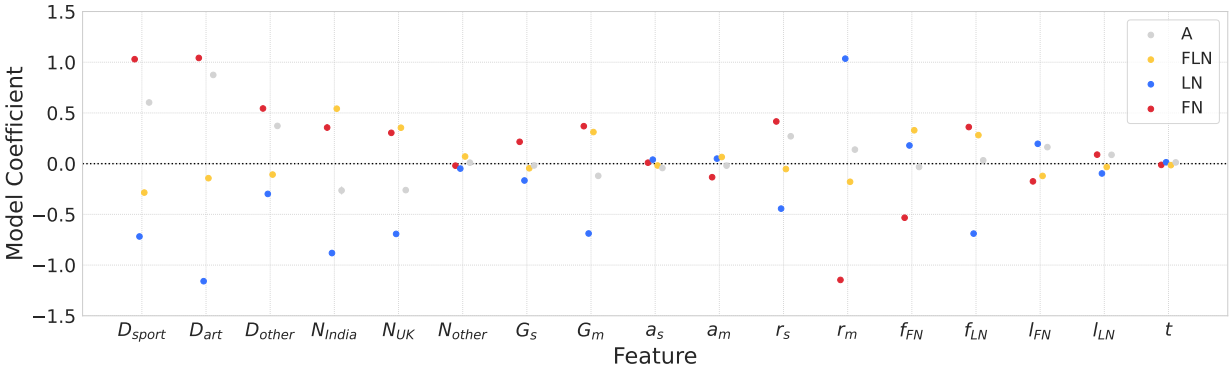


Figure 5.5: **Coefficients of the models fit on different reference types as outcomes.** Although plotted, the 95% confidence intervals are not visible due to their small size.

fit a logistic regression model based on the following equation:

$$\begin{aligned}\log p_r - \log(1-p_r) = \beta_0 &+ \beta_1 R_s + \beta_2 R_m + \beta_3 R_s R_m + \beta_4 C_s + \beta_5 C_m + \beta_6 C_s C_m \\ &+ \beta_7 G_s + \beta_8 G_m + \beta_9 a_s + \beta_{10} a_m + \beta_{11} r_s + \beta_{12} r_m \\ &+ \beta_{13} r_m + \beta_{14} f_{LN} + \beta_{15} f_{FN} + \beta_{16} l_{LN} + \beta_{17} l_{FN}.\end{aligned} \tag{5.2}$$

As opposed to the model in Equation 5.1, we do not include the nationality and domain features since we are focused only on US politics. Note that in this study, our goal is not to gain a deep understanding of the circumstances of using certain personal reference types in political communication. Instead, the purpose of this study is to serve as a starting point for future research.

5.5 Results

RQ1: Which entity features correlate with the usage of certain reference types? In Figure 5.5, we present the coefficients of the models fit for each reference type. The model coefficient, along with their standard errors and their significance levels, is shown in Appendix B (Table B.1). It is important to note that although some coefficients are small in magnitude, they are all statistically significant ($p < 0.05$). While we do not analyze the model fit for the alias reference type, we include its coefficients for completeness. We do the same in all the subsequent studies. Starting from the left-hand side of the figure,

we can observe that, on average, FN references are used more frequently in sport and art than in politics, while the opposite is true for FLN and LN. This is expected since sport and art are characterized by less formal communication than politics. As previously hypothesized, both in the UK and India, FN references are used significantly more often than LN and FLN references in comparison to the USA. The low standard error and magnitude of the coefficients corresponding to N_{other} indicate that the reference type usage in *other* nationalities is similar to their usage in the USA.

In line with previous studies, we find that women are, on average, the odds of women being referred to by their last name are approximately two times lower compared to men. On the contrary, the odds of the FN reference use for women are 1.45 times higher than for men. Similarly, women have 1.37 times greater odds of being referred to by their full name than men. Furthermore, we find that women are also more likely to use FN references and less likely to use LN and FLN references than men. Moving on to the age-based features, we find that a_m has a stronger effect on the reference type used than a_s. The odds of FN usage decrease by 12.5% with every ten years of mention's age. Conversely, with the same increase in mention's age, the odds of the LN usage increase by 5%, while the odds of FLN reference increase by 6.7%. Next, we can observe that the PageRank of the mention correlates with the usage of the LN reference, highlighting the importance of status. Interestingly, we also find that the opposite follows for the speaker's PageRank. Moving to the name-based features, we find that, as hypothesized, name length and frequency negatively correlate with its usage as a reference. Finally, we observe a slightly increasing trend for the odds of LN reference usage (approximately 1.4% per year) and decreasing trend of a similar magnitude for FN (approximately 1.2% per year) and FLN (approximately 1.5% per year).

RQ2: How does the reference type usage differ with respect to speaker's and mention's gender? In Figure 5.6, we present the average log odds of the reference type usage in different interactions calculated based on the coefficients corresponding to gender-based features based on the regression model (5.1) enhanced with the G_sG_m interaction term. The model coefficients are shown in We find that, when it comes to FN and LN usage, the effect of the speaker is stronger than in the case of FLN usage. The results confirm that in comparison to men, women are not only less likely to be referred to by the last name but also use the LN reference less frequently. The opposite follows for the FN reference type. This is indicated by a large gap in the within-gender interactions as opposed to cross-gender interactions.

RQ3.1: Can gender-based differences be consistently observed in different domains? Next, we proceed to study gender-based differences in different domains. In Figure 5.6, we show the average log odds of the reference type usage in different gender interactions across different domains. The average log odds are calculated based on the coefficients of the model with interaction terms shown in the first row in the interaction in Table 5.4. The gap between entirely male and entirely female interactions is consistent for all the reference types except for alias in politics. In sport and art domains, the difference in FN usage with respect to the speaker's gender is more emphasized than in politics. Interestingly, in the art domain, women are more likely to refer to men by their first names than to other women. We hypothesize that this is the case because the media often extensively covers actors' romantic affairs, and romantic partners usually refer to each other by their first names.

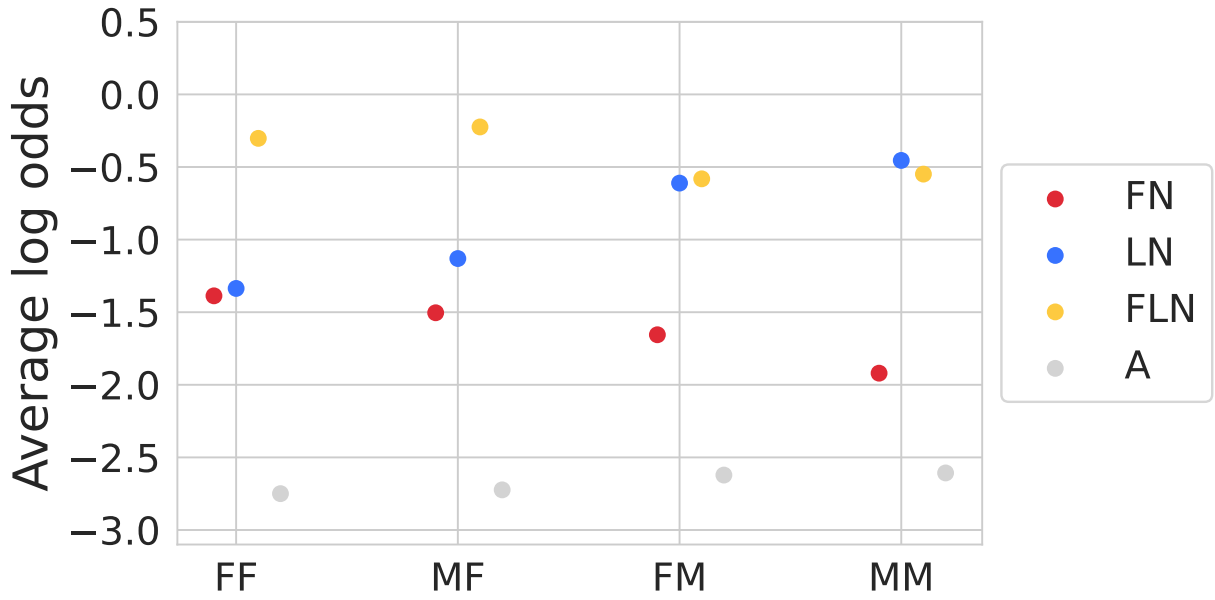


Figure 5.6: **Average log odds of the usage of different reference types in gender interactions.** Each interaction is encoded with two letters. The first letter indicates the gender of the speaker, while the second letter indicates the gender of the mention. F stands for female, while M stands for male. 95% confidence intervals are plotted but are not visible due to their small size.

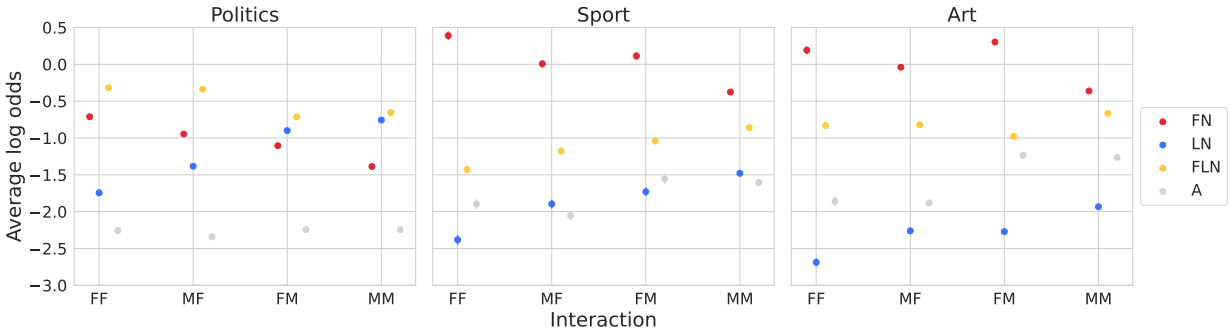


Figure 5.7: **Average log odds of the usage of different reference types in gender interactions in politics, sport, and art.** Interactions are encoded as in Figure 5.6. 95% confidence intervals are plotted but are not visible due to their small size.

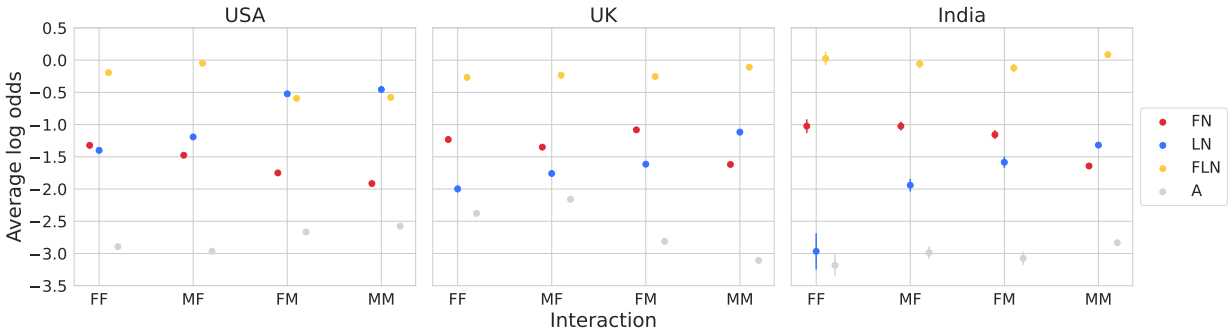


Figure 5.8: **Average log odds of the usage of different reference types in gender interactions in the USA, UK, and India.** Interactions are encoded as in Figure 5.6. Vertical lines indicate 95% confidence intervals.

RQ3.2: Can gender-based differences be consistently observed in different cultures? In Figure 5.8, analogously to Figure 5.6, we show the average log odds of the reference type usage in gender interactions with respect to a speaker's and a target's nationality. As before, differences in LN usage behave consistently across all nationalities. However, we also observe that, similar to the art domain, in the UK, women use FN when referring to men more frequently than when referring to women. This is likely because the UK quotations are dominated by prime ministers Theresa May and Boris Johnson. May is commonly referred to by her title and surname (*Prime Minister May*), whereas Johnson is frequently referred to as Boris.

RQ3.3: Do gender-based differences vary through time? In Figure 5.9, we show the effect of time on the reference type usage in gender interactions. We find that the usage of the LN reference slightly increases with each year and that the increase is stronger in cross-gender interactions than within-gender interactions. Note that the increase in last name usage in FF interactions is not statistically significant. On the other hand, we observe a significant decrease in the use of FLN reference in all the interactions. Since FN references are a characteristic of the interactions of persons who are close to each other, we hypothesize that its decrease, accompanied by the increase in the usage of LN references, might be correlated with the previously observed increase in negativity following Trump's entrance into US politics before 2016 US presidential elections [27]. However, we leave a detailed longitudinal analysis along with the study of linguistic correlates with the usage of certain reference types for future work.

RQ3.4: Can gender-based differences be observed within the same quotation? We conclude the analysis of gender-based differences in reference type used with the results of our matched study. In Figure 5.10, we present the coefficients of the regression model fit on the dataset matched by quotation. Overall, even after matching, the model coefficients do not drastically change, although due to a substantial reduction in sample size, certain coefficients become statistically insignificant ($p > 0.05$). Consistent with the previous findings, women are more likely to be referred to by their first name and to use their first name as a reference in communication than men. The opposite is true for the FN reference.

RQ4: How does the reference type usage differ with respect to the speaker's and mention's party alignment in US politics? As shown in Figure 5.11, within-party communication is characterized by

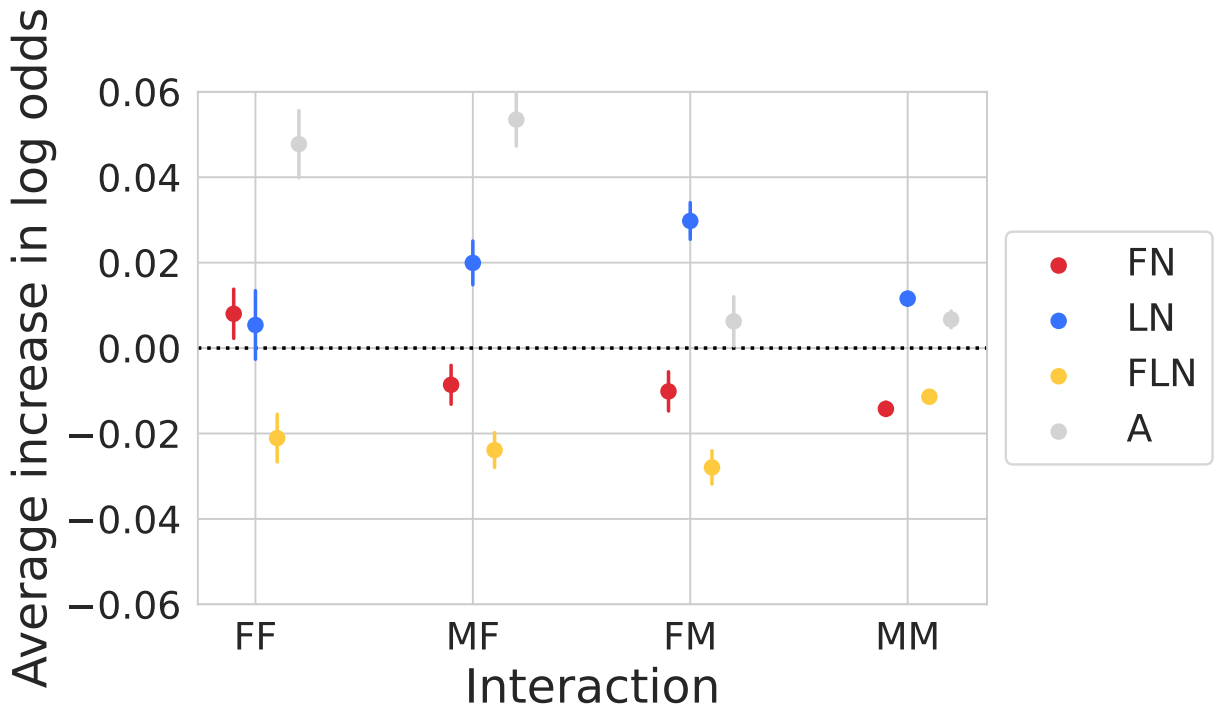


Figure 5.9: **Average yearly increase in log odds of the usage of different reference types in gender interaction.** Interactions are encoded as in Figure 5.6. Vertical lines indicate 95% confidence intervals.

higher usage of the FN reference type, hinting at the existence of close personal relationships between them. The opposite, although less evident, is true for the LN and FLN references. Furthermore, we find that in the communication within the Democratic party, the use of the FN reference is more frequent than in the Republican party. On the other hand, in cross-party communication, we find that Republicans tend to use first names more frequently when mentioning Democrats than vice-versa. We argue that the large difference in RD and DR communication can be explained by the Donald Trump-Hillary Clinton rivalry since Trump is a Republican who is commonly referred to as Trump, while Clinton is a Democrat who is commonly referred to as Hillary. In addition, unlike Trump, Clinton has been a member of the US Congress, which might explain a higher usage of FN references when referring to US Congress members. To back up the hypotheses, we exclude all Trump's or Clinton's mentions and fit the same model (Equation 5.2). Indeed, when Trump and Clinton are excluded, the gap in cross-party communication diminishes. Likewise, congressional affiliation seemingly only plays an important role in within-congress communication. However, even after excluding Trump and Clinton, within-party communication is still characterized by higher overall usage of the FN reference.

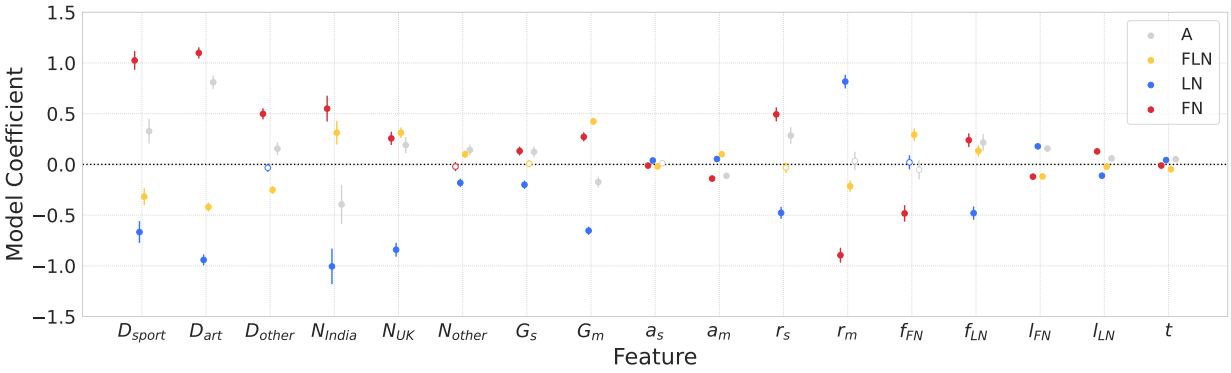


Figure 5.10: **Coefficients of the models fit on different reference types as outcomes on a dataset matched by the quotation.** Vertical lines denote 95% confidence intervals. Hollow points denote that the corresponding model coefficient is not statistically significant ($p > 0.05$).

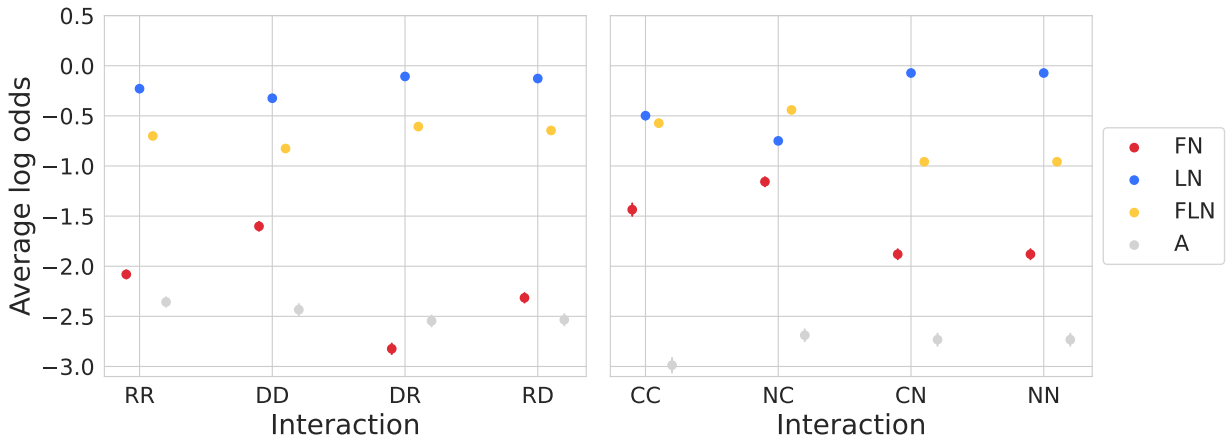


Figure 5.11: **Reference expression usage in US politics with respect to party affiliation and congress membership.** Each interaction is encoded with two letters. The first and second letter indicate the speaker's and mention's features, respectively. R stands for Republican, D for Democrat. On the other hand, C indicates US congress membership, while N indicates that a speaker or a mention is not a member of congress 95% confidence intervals are plotted but are not visible due to their small size.

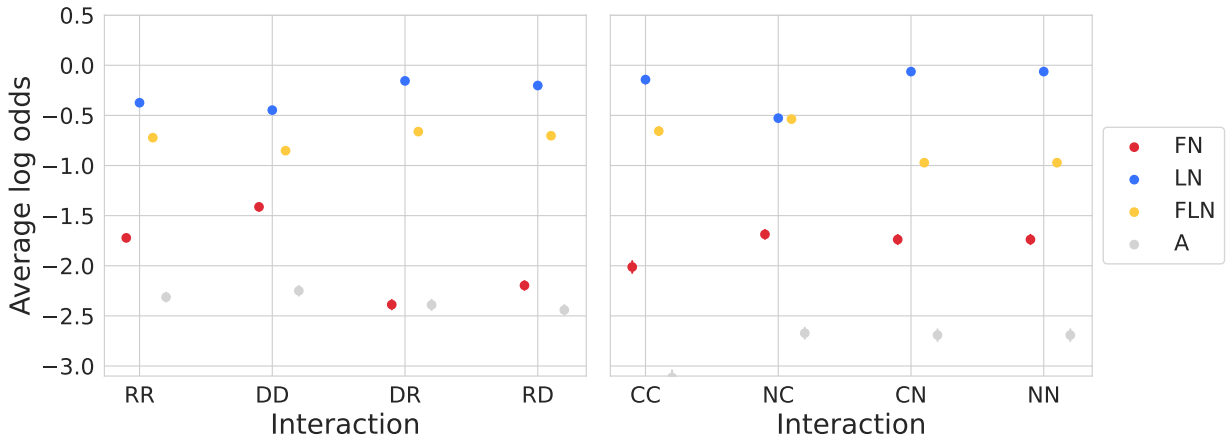


Figure 5.12: **Reference expression usage in US politics with respect to party affiliation and congress membership after excluding the mentions of Donald Trump and Hillary Clinton.** Each interaction is encoded with two letters. The first and second letters indicate the speaker's and mention's features, respectively. R stands for Republican, D for Democrat. On the other hand, C indicates US Congress membership, while N indicates that a speaker or a mention is not a member of Congress. Vertical lines denote 95% confidence intervals.

Chapter 6

Discussion

Summary of the results. The results of our studies of person referencing patterns in public discourse are consistent with previous research focused mainly on the general public's perception of professionals. We find that women are less likely to be referred to by their last names and more likely to use the LN reference when mentioning other persons than men. Regarding LN reference, our findings are consistent in the USA, UK, and India, as well as in politics sport, and art. On the other hand, we observed variations in the usage of the FN reference in the same occupations and nationalities. Furthermore, by conducting a matched study, we find that gender bias exists even when a man and a woman are mentioned in the same quotation. In general, the speaker's gender is a stronger predictor for using FN reference than for using LN reference. We also find that mention's PageRank correlates with higher use of FN and lower use of LN reference, while the opposite follows for speaker's PageRank. Lastly, we show that within-party communication in US politics is characterized by higher use of FN references, hinting at the existence of *first-name polarization* in US politics.

Implications. Atir and Ferguson [7] show that people tend to associate last name reference with fame and eminence as opposed to the use of the full name. Our findings suggest that in professional communication captured by the news, women are less likely to be referred to by their last name than men and that this nominal gender bias is consistent across different domains and cultures. Therefore, this phenomenon can influence the newsreader to perceive women mentioned in the news quotations as less eminent. This is particularly important in politics, where public perception of a politician determines the outcome of an election. Furthermore, according to our results, women are less likely to use last names in communication than men, thereby contributing to the perception of women as less eminent and increasing the gender gap. On the other hand, assuming that the findings by Atir and Ferguson are representative, this implies that by slightly altering their communication style, i.e. by choosing to refer to other women by their last name, women can contribute to reducing the gender gap.

Challenges. This thesis originally started as an effort to analyze person interactions by looking into linguistic cues in the quotations in which the persons mention each other. We found that the quotations

where both the speaker and the mention share the same party or gender are, on average, more positive than if the opposite is true. However, it is unclear whether sentiment is truly directed towards the mentioned person, which makes the finding difficult to justify. The problem becomes more apparent when we aggregate the scores over shorter periods or less-represented interactions. In this case, the overall score is more sensitive to noise. Thus, for sentiment-focused studies, it is crucial to develop well-performing targeted-sentiment analysis tools suitable for large-scale data. Aside from analyzing linguistic cues of the quotations using lexica such as LIWC [31], Empath [32] and VADER [33], inspired by Underwood, Bamman, and Lee [34], and Budhiraja et al. [35] we attempted to quantify party polarization in the United States of America by measuring classifier performance on a within-party vs. cross-party communication prediction task. While we found a positive correlation between the classifier performance and the year of the data snapshot used to train the classifier, although the finding is consistent with the literature, it is important to note that, based on manual inspection, quotation attribution accuracy increases significantly in later phases of Quotebank data collection. Therefore, we cannot justify that the increase in political polarization is truly the cause of the increase in classifier accuracy. Thus, we resorted to person referencing as a robust way of characterizing interactions in quotations.

Quotebank-related limitations. However, our studies of person referencing patterns are not without limitations. First, due to limitations of the heuristics used for entity linking [4], we cannot perfectly link mentions to their respective Wikidata entities, resulting in reduced accuracy of both the speaker and a mention in a respective quotation. Additionally, the heuristics always link a mention to one of the candidate entities. Thus, errors arise if an entity cannot be linked to any listed candidates. Since the candidate generator imposes an upper bound on the precision of the entity linker, it is important to discuss its limitations. As mentioned in Section 3.3, certain strings may be wrongly identified as person names. While we made a considerable effort to filter such spurious names, it is still possible that some spurious names went unnoticed. Another candidate generator limitation is that it requires that a person is mentioned by their full name at least once in an article to be identified as a named entity. This requirement negatively impacts recall of the generator as it can result in completely discarding mentions of prominent named entities who are well-known for their last name. Furthermore, name substrings shared by different entities in an article are ignored. For example, if Hillary Clinton and Chelsea Clinton are mentioned in the same article, their surname, *Clinton*, is not considered as an entity mention.

Aside from the entity linker, the limitations of the candidate generator also affect the performance of the quotation attribution system. Spurious strings identified as entity mentions can lead to erroneous attributions since a quotation can be attributed to an entity that did not appear in the quotation context. Analogously, if the candidate generator did not identify the actual speaker of a quotation as a speaker candidate, the quotation will either be falsely attributed or attributed to no speaker. Consequently, ignoring the actual speaker results in noise or loss of information. We argue that using simple preprocessing steps such as those described in Section 3.3 to filter out possible spurious mentions can significantly increase the attribution accuracy.

Quotegraph-related limitations. In Section 3.3, we described the removal of duplicated quotations and self-loops as an effort to mitigate noise in the data. However, by performing those steps, it is possible that

we excluded valuable information. For example, the quotations extracted from two completely different longer quotations may share the same substring of arbitrary length, in which case, we would falsely remove the shorter one. Similarly, although in self-loops, the speaker is most likely incorrectly identified and their removal is justified, we still lose the information about the interaction of the actual speaker and a person mentioned in the quotation. Thus, as an improvement to Quotebank, it would be beneficial to develop a robust method for quote deduplication.

Limitations related to personal reference usage analysis. The main limitation of our reference type analysis is that we conduct it on a quotation level. This causes the result to be biased towards more prominent speakers and mentions. However, this setting is more faithful to the news landscape and better reflects the usage of certain referencing types as perceived by the news reader. Furthermore, our analysis only covers given names and family names, ignoring the use of titles and nicknames. We also do not take into account the relationships (e.g. sibling or spouse) between speakers and mentions.

Chapter 7

Conclusions and Future work

Conclusions. In this thesis, we conducted a large-scale analysis of personal reference usage in the communication captured by the news. As a byproduct, we proposed three novel social networks extracted from Quotebank, namely Quotegraph, CoQuotegraph, and CoMentionGraph, and analyzed their properties. Our findings are consistent with the previous literature, showing that women are less likely to be referred to by their last name on average. Furthermore, we extended previous research in personal reference usage by (1) investigating personal reference usage in professional communication, (2) empirically demonstrating the importance of a person's status approximated by PageRank centrality calculated on Quotegraph, (3) highlighting the importance of speaker attributes by finding that women use LN reference more and FN reference less frequently than men, and (4) showing that the nominal gender bias exists in various occupations, cultures, and even when two persons are mentioned in the same quotation. Aside from providing new insights on personal referencing patterns, this thesis also showcases the applicability of Quotegraph, and consequently Quotebank, in computational social science.

Future work. The work conducted as a part of this thesis is far from finished. First, we merely scratched the surface in the analysis of the proposed social networks. Thus, a more comprehensive study is required to understand their properties deeply. For example, we propose to take a closer look into CoQuotegraph as a tool to uncover media biases through co-quotation patterns or to investigate when politicians focus more on humans than on social issues. When it comes to personal reference expressions, we propose to analyze linguistic correlates with certain reference types to investigate a possible link between gender-biased language and personal reference expressions. Next, the pool of personal reference expressions should be extended by titles and nicknames to enrich the analysis. More emphasis should also be put on the analysis of temporal trends. Lastly, it would be interesting to examine how certain significant events, such as death, impact personal reference use.

Bibliography

[1] Betsy Rymes. “Names”. In: *Journal of Linguistic Anthropology* 9.1/2 (1999).

[2] Gregory L Murphy. “Personal reference in English”. In: *Language in society* 17.3 (1988).

[3] Timoté Vaucher et al. “Quotebank: A Corpus of Quotations from a Decade of News”. In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 2021.

[4] Marko Čuljak et al. “Strong Heuristics for Named Entity Linking”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*. 2022.

[5] Denny Vrandečić and Markus Krötzsch. “Wikidata: A Free Collaborative Knowledgebase”. In: *Commun. ACM* 57.10 (2014).

[6] Lawrence Page et al. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. 1999.

[7] Stav Atir and Melissa J. Ferguson. “How gender determines the way we speak about professionals”. In: *Proceedings of the National Academy of Sciences* 115.28 (2018).

[8] Sara Marjanovic, Karolina Stanczak, and Isabelle Augenstein. “Quantifying Gender Biases Towards Politicians on Reddit”. In: *CoRR* abs/2112.12014 (2021).

[9] Christy Halbert and Melissa Latimer. “Battling Gendered Language: An Analysis of the Language Used by Sports Commentators in a Televised Coed Tennis Competition”. In: *Sociology of Sport Journal* 11.3 (1994).

[10] Michael A. Messner, Margaret Carlisle Duncan, and Kerry Jensen. “Separating the Men from the Girls: The Gendered Language of Televised Sports”. In: *Gender and Society* 7.1 (1993).

[11] Aron Culotta, Ron Bekkerman, and Andrew McCallum. “Extracting social networks and contact information from email and the Web”. In: *CEAS 2004 - First Conference on Email and Anti-Spam*. 2004.

[12] Jana Diesner, Terrill L Frantz, and Kathleen M Carley. “Communication networks from the Enron email corpus “It’s always about the people. Enron is no different””. In: *Computational & Mathematical Organization Theory* 11.3 (2005).

[13] Sune Lehmann, Benny Lautrup, and Andrew D Jackson. “Citation networks in high energy physics”. In: *Physical Review E* 68.2 (2003).

[14] Derek J De Solla Price. "Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front." In: *Science* 149.3683 (1965).

[15] "Evolution of the social network of scientific collaborations". In: *Physica A: Statistical Mechanics and its Applications* 311.3 (2002). ISSN: 0378-4371.

[16] James Moody. "The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999". In: *American sociological review* 69.2 (2004).

[17] Johanna Geiß, Andreas Spitz, and Michael Gertz. "Beyond Friendships and Followers: The Wikipedia Social Network". In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. 2015.

[18] Apoorv Agarwal et al. "Social network analysis of Alice in Wonderland". In: *Proceedings of the NAACL-HLT 2012 Workshop on computational linguistics for literature*. 2012.

[19] Prashant Jayannavar et al. "Validating literary theories using automatic social network extraction". In: *Proceedings of the fourth workshop on computational linguistics for literature*. 2015.

[20] David Elson, Nicholas Dames, and Kathleen McKeown. "Extracting Social Networks from Literary Fiction". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. 2010.

[21] John F Padgett and Christopher K Ansell. "Robust Action and the Rise of the Medici, 1400-1434". In: *American journal of sociology* 98.6 (1993).

[22] Michael Finegold et al. "Six degrees of Francis Bacon: a statistical method for reconstructing large historical social networks". In: *Digital Humanities Quarterly* 10.3 (2016).

[23] Matei Zaharia et al. "Apache Spark: A Unified Engine for Big Data Processing". In: *Commun. ACM* 59.11 (2016).

[24] Dario Pavllo, Tiziano Piccardi, and Robert West. "Quootstrap: Scalable Unsupervised Extraction of Quotation-Speaker Pairs from Large News Corpora via Bootstrapping". In: *Proceedings of the Twelfth International Conference on Web and Social Media*. 2018.

[25] *Stanford CoreNLP Official Documentation: Tokenization*. 1999. URL: https://stanfordnlp.github.io/corenlp-docs-dev/quote.html.

[26] Wikipedia. *Illeism — Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Illeism&oldid=1097984744. 2022.

[27] Jonathan Külz et al. *United States Politicians' Tone Became More Negative with 2016 Primary Campaigns*. 2022.

[28] M. E. J. Newman. *Networks: an introduction*. Oxford University Press, 2018.

[29] M. E. J. Newman. "Mixing patterns in networks". In: *Physical Review E* 67.2 (2003).

[30] The Associated Press. *The Associated Press Stylebook: 2020-2022*. Basic Books, 2020.

[31] Ryan Boyd et al. *The Development and Psychometric Properties of LIWC-22*. Tech. rep. 2022.

[32] Ethan Fast, Binbin Chen, and Michael S. Bernstein. "Empath: Understanding Topic Signals in Large-Scale Text". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016.

[33] Clayton J. Hutto and Eric Gilbert. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text". In: *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*. 2014.

[34] Ted Underwood, David Bamman, and Sabrina Lee. "The Transformation of Gender in English-Language Fiction". In: *Journal of Cultural Analytics* 3.2 (2018).

[35] Amar Budhiraja et al. "American Politicians Diverge Systematically, Indian Politicians do so Chaotically: Text Embeddings as a Window into Party Polarization". In: *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*. 2021.

Appendix A

Name Blacklist

Table A.1: **Name Blacklist.**

Name
war machine
is a goalkeeper
University of Bari
Harry Potter
model and actress
fund manager
Alexander the Great
Jon Snow
Vlad the Impaler
index finger
John the Divine
art collector
queen of hearts
Captain Marvel
Taj Mahal
Beverly Hills
James Bond
William Shakespeare

Appendix B

Model Coefficients

Table B.1: **Coefficients of the logistic regression model without interaction terms (Equation 5.1).**

	First name	**Last name**	**Full name**	**Alias**
Intercept	$-1.908(0.005)^{***}$	$-0.453(0.004)^{***}$	$-0.546(0.004)^{***}$	$-2.606(0.006)^{***}$
D_{sport}	$1.030(0.005)^{***}$	$-0.718(0.005)^{***}$	$-0.285(0.004)^{***}$	$0.603(0.006)^{***}$
D_{art}	$1.042(0.005)^{***}$	$-1.159(0.005)^{***}$	$-0.143(0.004)^{***}$	$0.875(0.006)^{***}$
D_{other}	$0.544(0.005)^{***}$	$-0.298(0.004)^{***}$	$-0.108(0.004)^{***}$	$0.372(0.006)^{***}$
N_{India}	$0.356(0.013)^{***}$	$-0.881(0.013)^{***}$	$0.542(0.011)^{***}$	$-0.264(0.017)^{***}$
N_{UK}	$0.304(0.005)^{***}$	$-0.693(0.006)^{***}$	$0.354(0.004)^{***}$	$-0.261(0.007)^{***}$
N_{other}	$-0.020(0.003)^{***}$	$-0.049(0.003)^{***}$	$0.071(0.003)^{***}$	$0.010(0.004)^{*}$
G_s	$0.215(0.004)^{***}$	$-0.166(0.004)^{***}$	$-0.045(0.004)^{***}$	$-0.018(0.005)^{***}$
G_m	$0.369(0.004)^{***}$	$-0.689(0.005)^{***}$	$0.312(0.004)^{***}$	$-0.120(0.006)^{***}$
a_s	$0.010(0.001)^{***}$	$0.040(0.001)^{***}$	$-0.018(0.001)^{***}$	$-0.043(0.001)^{***}$
a_m	$-0.134(0.001)^{***}$	$0.050(0.001)^{***}$	$0.066(0.001)^{***}$	$-0.020(0.001)^{***}$
r_s	$0.416(0.005)^{***}$	$-0.443(0.005)^{***}$	$-0.053(0.005)^{***}$	$0.270(0.007)^{***}$
r_m	$-1.146(0.006)^{***}$	$1.035(0.005)^{***}$	$-0.179(0.005)^{***}$	$0.139(0.007)^{***}$
f_{FN}	$-0.533(0.005)^{***}$	$0.179(0.005)^{***}$	$0.329(0.005)^{***}$	$-0.033(0.007)^{***}$
f_{LN}	$0.361(0.005)^{***}$	$-0.690(0.005)^{***}$	$0.281(0.004)^{***}$	$0.033(0.006)^{***}$
l_{FN}	$-0.175(0.001)^{***}$	$0.196(0.001)^{***}$	$-0.121(0.001)^{***}$	$0.163(0.001)^{***}$
l_{LN}	$0.089(0.001)^{***}$	$-0.097(0.001)^{***}$	$-0.033(0.001)^{***}$	$0.087(0.001)^{***}$
t	$-0.012(0.001)^{***}$	$0.014(0.001)^{***}$	$-0.015(0.001)^{***}$	$0.012(0.001)^{***}$

Table B.2: **Logistic regression model with gender-based interaction terms.**

	First name	**Last name**	**Full name**	**Alias**
Intercept	$-1.920(0.005)^{***}$	$-0.455(0.004)^{***}$	$-0.549(0.004)^{***}$	$-2.607(0.006)^{***}$
D_{sport}	$1.042(0.005)^{***}$	$-0.716(0.005)^{***}$	$-0.282(0.004)^{***}$	$0.604(0.006)^{***}$
D_{art}	$1.045(0.005)^{***}$	$-1.158(0.005)^{***}$	$-0.142(0.004)^{***}$	$0.875(0.006)^{***}$
D_{other}	$0.546(0.005)^{***}$	$-0.298(0.004)^{***}$	$-0.107(0.004)^{***}$	$0.372(0.006)^{***}$
N_{India}	$0.353(0.013)^{***}$	$-0.881(0.013)^{***}$	$0.542(0.011)^{***}$	$-0.264(0.017)^{***}$
N_{UK}	$0.303(0.005)^{***}$	$-0.693(0.006)^{***}$	$0.354(0.004)^{***}$	$-0.261(0.007)^{***}$
N_{other}	$-0.020(0.003)^{***}$	$-0.049(0.003)^{***}$	$0.070(0.003)^{***}$	$0.010(0.004)^{*}$
G_s	$0.265(0.005)^{***}$	$-0.156(0.005)^{***}$	$-0.033(0.004)^{***}$	$-0.015(0.006)^{*}$
G_m	$0.416(0.005)^{***}$	$-0.676(0.005)^{***}$	$0.325(0.005)^{***}$	$-0.117(0.007)^{***}$
G_sG_m	$-0.148(0.009)^{***}$	$-0.050(0.010)^{***}$	$-0.046(0.008)^{***}$	$-0.011(0.012)$
a_s	$0.009(0.001)^{***}$	$0.040(0.001)^{***}$	$-0.018(0.001)^{***}$	$-0.043(0.001)^{***}$
a_m	$-0.134(0.001)^{***}$	$0.050(0.001)^{***}$	$0.066(0.001)^{***}$	$-0.020(0.001)^{***}$
r_s	$0.416(0.005)^{***}$	$-0.443(0.005)^{***}$	$-0.053(0.005)^{***}$	$0.270(0.007)^{***}$
r_m	$-1.147(0.006)^{***}$	$1.035(0.005)^{***}$	$-0.179(0.005)^{***}$	$0.139(0.007)^{***}$
f_{FN}	$-0.534(0.005)^{***}$	$0.179(0.005)^{***}$	$0.329(0.005)^{***}$	$-0.033(0.007)^{***}$
f_{LN}	$0.360(0.005)^{***}$	$-0.690(0.005)^{***}$	$0.281(0.004)^{***}$	$0.033(0.006)^{***}$
l_{FN}	$-0.175(0.001)^{***}$	$0.196(0.001)^{***}$	$-0.121(0.001)^{***}$	$0.163(0.001)^{***}$
l_{LN}	$0.089(0.001)^{***}$	$-0.097(0.001)^{***}$	$-0.033(0.001)^{***}$	$0.087(0.001)^{***}$
t	$-0.012(0.001)^{***}$	$0.014(0.001)^{***}$	$-0.015(0.001)^{***}$	$0.012(0.001)^{***}$

Table B.3: **Logistic regression model with an interaction term for every occupation and gender-based feature, including the gender interaction term.**

	First name	**Last name**	**Full name**	**Alias**
Intercept	$-1.889(0.006)^{***}$	$-0.447(0.005)^{***}$	$-0.617(0.004)^{***}$	$-2.651(0.007)^{***}$
D_{sport}	$1.010(0.006)^{***}$	$-0.723(0.005)^{***}$	$-0.205(0.005)^{***}$	$0.639(0.007)^{***}$
D_{art}	$1.025(0.007)^{***}$	$-1.177(0.006)^{***}$	$-0.009(0.005)$	$0.980(0.007)^{***}$
D_{other}	$0.503(0.006)^{***}$	$-0.309(0.004)^{***}$	$-0.039(0.004)^{***}$	$0.407(0.007)^{***}$
G_s	$-0.013(0.013)$	$-0.100(0.007)^{***}$	$0.106(0.007)^{***}$	$-0.061(0.014)^{***}$
G_m	$0.471(0.011)^{***}$	$-0.813(0.009)^{***}$	$0.595(0.008)^{***}$	$0.187(0.013)^{***}$
N_{India}	$0.340(0.013)^{***}$	$-0.883(0.013)^{***}$	$0.568(0.011)^{***}$	$-0.254(0.017)^{***}$
N_{UK}	$0.302(0.005)^{***}$	$-0.687(0.006)^{***}$	$0.340(0.004)^{***}$	$-0.270(0.007)^{***}$
N_{other}	$-0.021(0.003)^{***}$	$-0.048(0.003)^{***}$	$0.072(0.003)^{***}$	$0.011(0.004)^{**}$
$D_{sport}G_s$	$0.208(0.023)^{***}$	$-0.106(0.024)^{***}$	$-0.122(0.021)^{***}$	$0.050(0.029)$
$D_{art}G_s$	$0.382(0.015)^{***}$	$-0.195(0.014)^{***}$	$-0.256(0.011)^{***}$	$0.029(0.017)$
$D_{other}G_s$	$0.295(0.015)^{***}$	$-0.043(0.010)^{***}$	$-0.163(0.010)^{***}$	$0.062(0.017)^{***}$
$D_{sport}G_m$	$-0.056(0.020)^{**}$	$0.210(0.026)^{***}$	$-0.634(0.020)^{***}$	$-0.353(0.028)^{***}$
$D_{art}G_m$	$-0.117(0.014)^{***}$	$0.300(0.015)^{***}$	$-0.475(0.012)^{***}$	$-0.522(0.017)^{***}$
$D_{other}G_m$	$-0.032(0.014)^{*}$	$0.186(0.012)^{***}$	$-0.277(0.011)^{***}$	$-0.283(0.017)^{***}$
G_sG_m	$0.087(0.024)^{***}$	$-0.037(0.019)$	$-0.050(0.017)^{**}$	$0.113(0.028)^{***}$
$D_{sport}G_sG_m$	$-0.060(0.037)$	$-0.019(0.045)$	$-0.148(0.036)^{***}$	$0.020(0.048)$
$D_{art}G_sG_m$	$-0.386(0.028)^{***}$	$0.127(0.028)^{***}$	$0.228(0.022)^{***}$	$-0.091(0.033)^{**}$
$D_{other}G_sG_m$	$-0.135(0.029)^{***}$	$-0.180(0.026)^{***}$	$0.127(0.022)^{***}$	$-0.029(0.036)$
a_s	$0.010(0.001)^{***}$	$0.039(0.001)^{***}$	$-0.019(0.001)^{***}$	$-0.043(0.001)^{***}$
a_m	$-0.135(0.001)^{***}$	$0.051(0.001)^{***}$	$0.064(0.001)^{***}$	$-0.023(0.001)^{***}$
r_s	$0.415(0.005)^{***}$	$-0.442(0.005)^{***}$	$-0.055(0.005)^{***}$	$0.272(0.007)^{***}$
r_m	$-1.142(0.006)^{***}$	$1.033(0.005)^{***}$	$-0.178(0.005)^{***}$	$0.143(0.007)^{***}$
f_{FN}	$-0.534(0.005)^{***}$	$0.178(0.005)^{***}$	$0.332(0.005)^{***}$	$-0.032(0.007)^{***}$
f_{LN}	$0.361(0.005)^{***}$	$-0.687(0.005)^{***}$	$0.276(0.004)^{***}$	$0.028(0.006)^{***}$
l_{FN}	$-0.175(0.001)^{***}$	$0.196(0.001)^{***}$	$-0.122(0.001)^{***}$	$0.162(0.001)^{***}$
l_{LN}	$0.089(0.001)^{***}$	$-0.097(0.001)^{***}$	$-0.033(0.001)^{***}$	$0.087(0.001)^{***}$
t	$-0.012(0.001)^{***}$	$0.014(0.001)^{***}$	$-0.014(0.001)^{***}$	$0.013(0.001)^{***}$

Table B.4: **Logistic regression model with an interaction term for every nationality and gender-based feature, including the gender interaction term.**

	First name	**Last name**	**Full name**	**Alias**
Intercept	$-1.915(0.006)^{***}$	$-0.455(0.004)^{***}$	$-0.578(0.004)^{***}$	$-2.576(0.007)^{***}$
N_{India}	$0.273(0.017)^{***}$	$-0.863(0.014)^{***}$	$0.663(0.012)^{***}$	$-0.258(0.021)^{***}$
N_{UK}	$0.297(0.005)^{***}$	$-0.662(0.006)^{***}$	$0.469(0.005)^{***}$	$-0.531(0.008)^{***}$
N_{other}	$-0.023(0.003)^{***}$	$-0.045(0.003)^{***}$	$0.116(0.003)^{***}$	$-0.047(0.004)^{***}$
G_s	$0.165(0.007)^{***}$	$-0.067(0.006)^{***}$	$-0.015(0.006)^{*}$	$-0.090(0.009)^{***}$
G_m	$0.439(0.007)^{***}$	$-0.736(0.007)^{***}$	$0.531(0.006)^{***}$	$-0.389(0.010)^{***}$
D_{sport}	$1.040(0.005)^{***}$	$-0.719(0.005)^{***}$	$-0.291(0.004)^{***}$	$0.623(0.006)^{***}$
D_{art}	$1.041(0.005)^{***}$	$-1.158(0.005)^{***}$	$-0.149(0.004)^{***}$	$0.892(0.006)^{***}$
D_{other}	$0.544(0.005)^{***}$	$-0.298(0.004)^{***}$	$-0.110(0.004)^{***}$	$0.381(0.006)^{***}$
$N_{India}G_s$	$0.322(0.036)^{***}$	$-0.199(0.045)^{***}$	$-0.192(0.032)^{***}$	$-0.151(0.052)^{**}$
$N_{UK}G_s$	$0.372(0.016)^{***}$	$-0.430(0.020)^{***}$	$-0.130(0.014)^{***}$	$0.384(0.023)^{***}$
$N_{other}G_s$	$0.139(0.011)^{***}$	$-0.186(0.010)^{***}$	$-0.003(0.009)$	$0.134(0.013)^{***}$
$N_{India}G_m$	$0.180(0.036)^{***}$	$0.113(0.048)^{*}$	$-0.673(0.033)^{***}$	$0.237(0.052)^{***}$
$N_{UK}G_m$	$-0.171(0.016)^{***}$	$0.095(0.018)^{***}$	$-0.656(0.014)^{***}$	$1.337(0.020)^{***}$
$N_{other}G_m$	$-0.025(0.010)^{*}$	$0.141(0.011)^{***}$	$-0.344(0.009)^{***}$	$0.340(0.014)^{***}$
G_sG_m	$-0.012(0.012)$	$-0.141(0.014)^{***}$	$-0.133(0.011)^{***}$	$0.160(0.016)^{***}$
$N_{India}G_sG_m$	$-0.476(0.070)^{***}$	$-0.619(0.154)^{***}$	$0.422(0.067)^{***}$	$-0.117(0.106)$
$N_{UK}G_sG_m$	$-0.407(0.030)^{***}$	$0.398(0.038)^{***}$	$0.244(0.027)^{***}$	$-0.673(0.038)^{***}$
$N_{Other}G_sG_m$	$-0.229(0.019)^{***}$	$0.218(0.022)^{***}$	$0.147(0.017)^{***}$	$-0.273(0.025)^{***}$
a_s	$0.009(0.001)^{***}$	$0.039(0.001)^{***}$	$-0.019(0.001)^{***}$	$-0.042(0.001)^{***}$
a_m	$-0.133(0.001)^{***}$	$0.049(0.001)^{***}$	$0.066(0.001)^{***}$	$-0.021(0.001)^{***}$
r_s	$0.416(0.005)^{***}$	$-0.443(0.005)^{***}$	$-0.060(0.005)^{***}$	$0.279(0.007)^{***}$
r_m	$-1.147(0.006)^{***}$	$1.033(0.005)^{***}$	$-0.177(0.005)^{***}$	$0.139(0.007)^{***}$
f_{FN}	$-0.535(0.005)^{***}$	$0.176(0.005)^{***}$	$0.334(0.005)^{***}$	$-0.028(0.007)^{***}$
f_{LN}	$0.359(0.005)^{***}$	$-0.689(0.005)^{***}$	$0.288(0.004)^{***}$	$0.020(0.006)^{**}$
l_{FN}	$-0.174(0.001)^{***}$	$0.196(0.001)^{***}$	$-0.121(0.001)^{***}$	$0.163(0.001)^{***}$
l_{LN}	$0.088(0.001)^{***}$	$-0.095(0.001)^{***}$	$-0.035(0.001)^{***}$	$0.090(0.001)^{***}$
t	$-0.012(0.001)^{***}$	$0.013(0.001)^{***}$	$-0.014(0.001)^{***}$	$0.012(0.001)^{***}$

Table B.5: **Logistic regression model with an interaction term gender-based feature, including the gender interaction term, and time in months.**

	First name	**Last name**	**Full name**	**Alias**
Intercept	$-1.913(0.006)^{***}$	$-0.447(0.005)^{***}$	$-0.561(0.004)^{***}$	$-2.585(0.007)^{***}$
G_s	$0.250(0.011)^{***}$	$-0.228(0.010)^{***}$	$0.031(0.009)^{***}$	$-0.012(0.014)$
G_m	$0.396(0.010)^{***}$	$-0.708(0.012)^{***}$	$0.372(0.009)^{***}$	$-0.294(0.014)^{***}$
N_{India}	$0.354(0.013)^{***}$	$-0.879(0.013)^{***}$	$0.541(0.011)^{***}$	$-0.264(0.017)^{***}$
N_{UK}	$0.303(0.005)^{***}$	$-0.693(0.006)^{***}$	$0.355(0.004)^{***}$	$-0.262(0.007)^{***}$
N_{other}	$-0.020(0.003)^{***}$	$-0.048(0.003)^{***}$	$0.070(0.003)^{***}$	$0.009(0.004)^{*}$
D_{sport}	$1.042(0.005)^{***}$	$-0.717(0.005)^{***}$	$-0.282(0.004)^{***}$	$0.603(0.006)^{***}$
D_{art}	$1.046(0.005)^{***}$	$-1.157(0.005)^{***}$	$-0.143(0.004)^{***}$	$0.875(0.006)^{***}$
D_{other}	$0.546(0.005)^{***}$	$-0.298(0.004)^{***}$	$-0.107(0.004)^{***}$	$0.371(0.006)^{***}$
G_sG_m	$-0.197(0.019)^{***}$	$0.079(0.023)^{***}$	$-0.120(0.018)^{***}$	$0.005(0.026)$
t	$-0.014(0.001)^{***}$	$0.012(0.001)^{***}$	$-0.011(0.001)^{***}$	$0.007(0.001)^{***}$
tG_s	$0.004(0.002)$	$0.018(0.002)^{***}$	$-0.017(0.002)^{***}$	$-0.000(0.003)$
tG_m	$0.006(0.002)^{*}$	$0.008(0.003)^{**}$	$-0.012(0.002)^{***}$	$0.047(0.003)^{***}$
tG_sG_m	$0.013(0.004)^{**}$	$-0.033(0.005)^{***}$	$0.019(0.004)^{***}$	$-0.005(0.006)$
a_s	$0.009(0.001)^{***}$	$0.040(0.001)^{***}$	$-0.018(0.001)^{***}$	$-0.043(0.001)^{***}$
a_m	$-0.134(0.001)^{***}$	$0.050(0.001)^{***}$	$0.066(0.001)^{***}$	$-0.020(0.001)^{***}$
r_s	$0.416(0.005)^{***}$	$-0.443(0.005)^{***}$	$-0.053(0.005)^{***}$	$0.270(0.007)^{***}$
r_m	$-1.147(0.006)^{***}$	$1.035(0.005)^{***}$	$-0.179(0.005)^{***}$	$0.140(0.007)^{***}$
f_{FN}	$-0.534(0.005)^{***}$	$0.179(0.005)^{***}$	$0.329(0.005)^{***}$	$-0.033(0.007)^{***}$
f_{LN}	$0.360(0.005)^{***}$	$-0.690(0.005)^{***}$	$0.281(0.004)^{***}$	$0.033(0.006)^{***}$
l_{FN}	$-0.175(0.001)^{***}$	$0.196(0.001)^{***}$	$-0.121(0.001)^{***}$	$0.163(0.001)^{***}$
l_{LN}	$0.089(0.001)^{***}$	$-0.097(0.001)^{***}$	$-0.033(0.001)^{***}$	$0.087(0.001)^{***}$

Table B.6: **Logistic regression model fit on the dataset used for the matched study.**

	First name	Last name	Full name	Alias
Intercept	$-2.106(0.031)^{***}$	$-0.722(0.025)^{***}$	$-0.091(0.022)^{***}$	$-2.809(0.038)^{***}$
D_{sport}	$1.025(0.045)^{***}$	$-0.667(0.053)^{***}$	$-0.317(0.040)^{***}$	$0.327(0.059)^{***}$
D_{art}	$1.099(0.026)^{***}$	$-0.942(0.025)^{***}$	$-0.421(0.020)^{***}$	$0.810(0.031)^{***}$
D_{other}	$0.498(0.026)^{***}$	$-0.033(0.018)$	$-0.252(0.017)^{***}$	$0.154(0.030)^{***}$
N_{India}	$0.550(0.063)^{***}$	$-1.006(0.087)^{***}$	$0.313(0.057)^{***}$	$-0.395(0.096)^{***}$
N_{UK}	$0.257(0.031)^{***}$	$-0.842(0.033)^{***}$	$0.312(0.024)^{***}$	$0.190(0.038)^{***}$
N_{other}	$-0.022(0.020)$	$-0.182(0.018)^{***}$	$0.100(0.016)^{***}$	$0.142(0.025)^{***}$
G_s	$0.132(0.020)^{***}$	$-0.201(0.019)^{***}$	$0.007(0.016)$	$0.124(0.024)^{***}$
G_m	$0.271(0.020)^{***}$	$-0.654(0.018)^{***}$	$0.424(0.016)^{***}$	$-0.173(0.025)^{***}$
a_s	$-0.012(0.006)^{*}$	$0.039(0.005)^{***}$	$-0.020(0.005)^{***}$	$0.012(0.007)$
a_m	$-0.140(0.006)^{***}$	$0.053(0.006)^{***}$	$0.100(0.005)^{***}$	$-0.113(0.007)^{***}$
r_s	$0.493(0.033)^{***}$	$-0.477(0.028)^{***}$	$-0.032(0.024)$	$0.285(0.040)^{***}$
r_m	$-0.896(0.035)^{***}$	$0.817(0.032)^{***}$	$-0.215(0.027)^{***}$	$0.034(0.044)$
f_{FN}	$-0.483(0.039)^{***}$	$0.020(0.033)$	$0.293(0.030)^{***}$	$-0.053(0.046)$
f_{LN}	$0.239(0.032)^{***}$	$-0.481(0.031)^{***}$	$0.133(0.026)^{***}$	$0.216(0.040)^{***}$
l_{FN}	$-0.121(0.006)^{***}$	$0.178(0.006)^{***}$	$-0.119(0.005)^{***}$	$0.157(0.007)^{***}$
l_{LN}	$0.128(0.005)^{***}$	$-0.111(0.005)^{***}$	$-0.022(0.004)^{***}$	$0.060(0.006)^{***}$
t	$-0.011(0.005)^{*}$	$0.042(0.004)^{***}$	$-0.048(0.004)^{***}$	$0.052(0.006)^{***}$

Table B.7: **Logistic regression model with political features Equation 5.2.**

	First name	Last name	Full name	Alias
Intercept	$-1.920(0.005)^{***}$	$-0.455(0.004)^{***}$	$-0.549(0.004)^{***}$	$-2.607(0.006)^{***}$
D_{sport}	$1.042(0.005)^{***}$	$-0.716(0.005)^{***}$	$-0.282(0.004)^{***}$	$0.604(0.006)^{***}$
D_{art}	$1.045(0.005)^{***}$	$-1.158(0.005)^{***}$	$-0.142(0.004)^{***}$	$0.875(0.006)^{***}$
D_{other}	$0.546(0.005)^{***}$	$-0.298(0.004)^{***}$	$-0.107(0.004)^{***}$	$0.372(0.006)^{***}$
N_{India}	$0.353(0.013)^{***}$	$-0.881(0.013)^{***}$	$0.542(0.011)^{***}$	$-0.264(0.017)^{***}$
N_{UK}	$0.303(0.005)^{***}$	$-0.693(0.006)^{***}$	$0.354(0.004)^{***}$	$-0.261(0.007)^{***}$
N_{other}	$-0.020(0.003)^{***}$	$-0.049(0.003)^{***}$	$0.070(0.003)^{***}$	$0.010(0.004)^{*}$
G_s	$0.265(0.005)^{***}$	$-0.156(0.005)^{***}$	$-0.033(0.004)^{***}$	$-0.015(0.006)^{*}$
G_m	$0.416(0.005)^{***}$	$-0.676(0.005)^{***}$	$0.325(0.005)^{***}$	$-0.117(0.007)^{***}$
G_sG_m	$-0.148(0.009)^{***}$	$-0.050(0.010)^{***}$	$-0.046(0.008)^{***}$	$-0.011(0.012)$
a_s	$0.009(0.001)^{***}$	$0.040(0.001)^{***}$	$-0.018(0.001)^{***}$	$-0.043(0.001)^{***}$
a_m	$-0.134(0.001)^{***}$	$0.050(0.001)^{***}$	$0.066(0.001)^{***}$	$-0.020(0.001)^{***}$
r_s	$0.416(0.005)^{***}$	$-0.443(0.005)^{***}$	$-0.053(0.005)^{***}$	$0.270(0.007)^{***}$
r_m	$-1.147(0.006)^{***}$	$1.035(0.005)^{***}$	$-0.179(0.005)^{***}$	$0.139(0.007)^{***}$
f_{FN}	$-0.534(0.005)^{***}$	$0.179(0.005)^{***}$	$0.329(0.005)^{***}$	$-0.033(0.007)^{***}$
f_{LN}	$0.360(0.005)^{***}$	$-0.690(0.005)^{***}$	$0.281(0.004)^{***}$	$0.033(0.006)^{***}$
l_{FN}	$-0.175(0.001)^{***}$	$0.196(0.001)^{***}$	$-0.121(0.001)^{***}$	$0.163(0.001)^{***}$
l_{LN}	$0.089(0.001)^{***}$	$-0.097(0.001)^{***}$	$-0.033(0.001)^{***}$	$0.087(0.001)^{***}$
t	$-0.012(0.001)^{***}$	$0.014(0.001)^{***}$	$-0.015(0.001)^{***}$	$0.012(0.001)^{***}$

Table B.8: **Logistic regression model with political features Equation 5.2 fit on a dataset where the mentions of Donald Trump and Hillary Clinton are excluded.**

	First name	**Last name**	**Full name**	**Alias**
Intercept	$-1.413(0.020)^{***}$	$-0.447(0.014)^{***}$	$-0.852(0.015)^{***}$	$-2.250(0.025)^{***}$
G_s	$0.128(0.016)^{***}$	$-0.105(0.011)^{***}$	$0.050(0.011)^{***}$	$0.009(0.019)$
G_m	$0.402(0.017)^{***}$	$-0.726(0.013)^{***}$	$0.533(0.012)^{***}$	$0.088(0.021)^{***}$
R_s	$-0.784(0.019)^{***}$	$0.245(0.012)^{***}$	$0.149(0.012)^{***}$	$-0.191(0.022)^{***}$
R_m	$-0.975(0.021)^{***}$	$0.292(0.013)^{***}$	$0.190(0.013)^{***}$	$-0.141(0.023)^{***}$
C_s	$-0.325(0.019)^{***}$	$0.384(0.012)^{***}$	$-0.120(0.013)^{***}$	$-0.442(0.023)^{***}$
C_m	$-0.274(0.018)^{***}$	$-0.080(0.012)^{***}$	$0.315(0.012)^{***}$	$-0.422(0.022)^{***}$
R_sR_m	$1.450(0.027)^{***}$	$-0.463(0.016)^{***}$	$-0.209(0.017)^{***}$	$0.268(0.031)^{***}$
C_sC_m	$0.526(0.026)^{***}$	$-0.192(0.017)^{***}$	$-0.163(0.017)^{***}$	$0.476(0.032)^{***}$
a_s	$0.051(0.005)^{***}$	$0.060(0.003)^{***}$	$-0.093(0.003)^{***}$	$0.026(0.006)^{***}$
a_m	$-0.072(0.005)^{***}$	$-0.033(0.003)^{***}$	$0.098(0.003)^{***}$	$-0.096(0.006)^{***}$
r_s	$0.715(0.025)^{***}$	$-0.636(0.016)^{***}$	$0.325(0.017)^{***}$	$0.083(0.030)^{**}$
r_m	$-1.216(0.027)^{***}$	$0.828(0.019)^{***}$	$-0.189(0.020)^{***}$	$-0.178(0.034)^{***}$
f_{FN}	$0.204(0.022)^{***}$	$-0.368(0.014)^{***}$	$0.333(0.014)^{***}$	$0.084(0.026)^{**}$
f_{LN}	$0.509(0.024)^{***}$	$-0.967(0.016)^{***}$	$0.781(0.016)^{***}$	$-0.115(0.030)^{***}$
l_{FN}	$-0.190(0.005)^{***}$	$0.193(0.003)^{***}$	$-0.154(0.003)^{***}$	$0.116(0.005)^{***}$
l_{LN}	$0.073(0.004)^{***}$	$-0.023(0.003)^{***}$	$-0.025(0.003)^{***}$	$0.033(0.005)^{***}$
t	$-0.008(0.003)^{*}$	$-0.003(0.002)$	$0.002(0.002)$	$0.019(0.004)^{***}$