Definite Articles Preceding Anthroponyms in Spanish: A Multi-Country Corpus Study

by

Keenan Hunt

(Under the Direction of Chad Howe)

Abstract

Through the use of corpus methodologies, this study examines the use of definite articles preceding anthroponyms in several varieties of Spanish. Several factors are investigated regarding their relation to article presence/absence, including the gender of referents, the popularity of names, and several basic syntactic factors. The data were analyzed using a mixed-effects generalized linear regression model, that included name as a random effect.

The analysis of the data found that definite articles occurred before anthroponyms in all of the countries studied, and a hypothesis stating that Chile would see the highest rate of occurrence was confirmed, with the Chile dataset seeing a significantly higher rate of article occurrence. In contrast, a hypothesis that female referents would see a higher rate of article occurrence was not confirmed by the data, as there were no significant effects from the gender of the referent on article occurrence. A third hypothesis held that modifying/specifying elements (such as *de* or *que*) following names would correspond to higher rates of article occurrence. This was confirmed by the data.

Index words: [Anthroponyms, proper names, definite articles, Spanish, corpus methodology]

Definite Articles Preceding Anthroponyms in Spanish: A Multi-Country Corpus Study

by

Keenan Hunt

B.A., Clemson University, 2019

A Thesis Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the
Degree.

Master of Arts

Athens, Georgia

2021

Definite Articles Preceding Anthroponyms in Spanish: A Multi-Country Corpus Study

by

Keenan Hunt

Major Professor: Chad Howe

Committee: Gary Baker
Diana Ranson

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
December 2021

ACKNOWLEDGMENTS

The completion of this thesis would not have been possible without the knowledge and support of many people, but I'd like to give a special thanks to Dr. Chad Howe. Beyond the generous amount of attention, energy, and feedback he has given in working with me to develop this thesis, every course of his that I've taken has sparked new linguistic fascinations that have inspired me to seek out new lines of inquiry and investigation.

Additionally, I'd like to thank Dr. Diana Ranson and Dr. Gary Baker for the feedback and myriad comments they've provided, which have not only aided me in completing this project, but have given me new perspective on what it means to conduct and discuss linguistic work.

Finally, I'd like to thank the entire Department of Romance Languages of the University of Georgia. I would not have been able to complete this thesis without the support of many individuals from every part of the department. In my time here, I've had access to a wealth of knowledge and experience provided by my mentors and colleagues, and I would not be where I am today if not for the generosity of the department.

Contents

Acknowledgments **iv**

List of Figures **vii**

List of Tables **viii**

1 Introduction **1**

2 Literature Review **5**
2.1 Semantics of Proper Names and Articles 5
2.2 The structure in other Romance varieties 7
2.3 Spanish descriptive grammar entries 9
2.4 Other studies on the structure in Spanish 13

3 Methodology **18**
3.1 Research questions 18
3.2 Data collection methods 18
3.3 Factors 23
3.4 Hypotheses 28
3.5 Data analysis methods 28

4 Results **31**
4.1 Frequency by country 31
4.2 Independent variables 34

5 Discussion **37**
5.1 Hypothesis 1: Of the countries included in this study, Chile will have the highest frequency of article presence. 37
5.2 Hypothesis 2: Female names will have a higher frequency of article presence than male names. 40

5.3 Hypothesis 3: Names followed by *que*, *de*, or by an adjective will have higher frequencies of article presence than names that are not followed by those elements. 41
5.4 Other factors 42
5.5 Methodological limitations 44

6 Conclusion **46**

Bibliography **49**

Appendices **53**

A CQP code used for data extraction **53**

B List of excluded names **54**

C Final name list by frequency **58**

List of Figures

3.1 Words-Per-Million values of the top 500 names in the *CdE* . . 30

4.1 Visualization of frequencies of article use for each country (percentage of article occurence out of total tokens) 33

List of Tables

3.1 Percentage of the *CdE* made up by each of the selected eight countries. 20
3.2 Words-per-million frequencies of the top 10 most frequent names. 22
3.3 Summary of the factors analyzed in the study 27

4.1 Article occurrence by country 32
4.2 Significance of difference in article occurrence rate between countries 32
4.3 Effects across all eight country datasets 35

C.1 All names with their frequencies by country (frequencies in words-per-million) 61

Chapter 1

Introduction

In conversational Spanish, occasionally one might hear a speaker use a definite article before the name of a person or people. On its own, this is not particularly noteworthy; many times, an article is expected, as will be discussed in the following examples.

The *Diccionario panhispánico de dudas* (Real Academia Español [RAE], 2005) provides a concise description of cases where one might expect to find a person's name (also called an anthroponym) preceded by a definite article. These two cases, illustrated by examples from the *Corpus del Español* (Davies, 2016-) are as follows:

1. The article is obligatory when the name is plural (RAE, 2005).

   (1.1) *... creo que el 12 es el Dulce Nombre de María. El día de la Virgen en general, de todas* ***las Marías****.*

   '... I think that the 12th is the Sweet Name of Maria. The day of the Virgin in general, of all **the Marias**.'

   (1.2) *En Málaga, se nos dan bien* ***los Pablos****. Desde Pablo Picasso a Pablo Pineda y Pablo Alborán.*

   'In Malaga, **the Pablos** have done well by us. From Pablo Picasso to Pablo Pineda and Pablo Alborán.'

2. The article is obligatory when the name is singular and either followed by a specifier or preceded by a qualifier (RAE, 2005).

   (1.3) *Cada uno de esos hombres quiere mucho a una mujer imaginaria y muy poco a* ***la Isabel****, a* ***la Gloria*** *o a* ***la Paula de carne y hueso*** *que tienen a su lado.*

'Every one of those men loves an imaginary woman yet very few love **the Isabel**, **the Gloria**, or **the Paula of flesh and blood** that they have at their side.'

(1.4) *¡Cuánto más grande y más conmovedora es la historia* ***del José real****!*

'How grand and moving is the story of **the real José**!'

Examples from the *Corpus del español: Web/dialects* (Davies, 2016-).

Apart from these contexts, however, there are other contexts where one can find use of the definite article before an anthroponym. In the same section describing the situations where use of an article is obligatory, the *DPD* states the following regarding cases where neither condition applies:

> The anteposition of the article, in these cases, tends to be typical of common speech… Notwithstanding, there are zones of the Hispanic world, for example in Chile, where the anteposition is also found in cultured speech, commonly in colloquial registers, and especially before the names of women (RAE, 2005, translation mine).

This possible use of definite articles with anthroponyms is not something that one will find in the average textbook of Spanish for second language learners, and, in the anecdotal experience of the author, is not considered to be acceptable under prescriptive norms. At the same time, these uses are well-documented, as we will discuss in section 2.3. Following are several examples of cases where non-modified, non-plural names are seen with articles:

(1.5) *Con* ***la Andrea*** *no paramos de reírnos: yo de vergüenza, ella de vergüenza ajena.*

'With **the Andrea** we didn't stop laughing: Myself from shame, her from embarrassment.'

(1.6) *… El que no creo que consiga quedarse es* ***el Felipe*** *porque la situación ahora mismo es muy distinta…*

'… The one I don't think will stay is **the Felipe**, because the situation now is very different…'

(1.7) *A punto de coger el macuto, camino de nuestros lares, que va* ***el Emilio*** *y dice que quieto, parao,* ***la Astrid*** *y la banda nos obsequian con un extra, una selección de música peliculera sobre el guaperas James Bond…*

> 'About to pick up the bag, on the way to our homes, **the Emilio** went and said quiet, be still, **the Astrid** and the band are going to give us an extra, a selection of film music about the heart-throb James Bond...'
>
> Examples from the *Corpus del español: Web/dialects* (Davies, 2016-).

Unfortunately, even with various sources describing the phenomenon, and various studies into the semantic and pragmatic uses of this structure, there is a lack of studies that take a quantitative approach to measuring the presence of the structure in contemporary Spanish.

Some key questions remain unanswered regarding the use of definite articles with anthroponyms in Spanish:

- Exactly how geographically widespread is this phenomenon in Spanish?
- Do the patterns of usage of this structure vary geographically?

Finding answers to these questions would help in determining whether this structure could be considered a feature of Spanish as a whole, or a more dialect-specific feature. Additionally, other Ibero-romance languages such as Catalan and Portuguese have similar structures, which are much better understood and more studied than the structure in Spanish (Alves, 2008; Alves, 2017; Amaral, 2016; Christodulelis, 2013; Coromina i Pou, 2001; GenCat, 2020a; GenCat, 2020b; Mendes, 2015). Gaining a more complete understanding of the nature of this structure in Spanish could help in drawing comparisons among Ibero-Romance languages.

The present thesis seeks to analyze the extent of the use of definite articles with anthroponyms in the Spanish-speaking world. Specifically, this study looks at the rates of appearance of the structure in various country sections of the *Corpus del español* (*CdE*), an online corpus which contains over two billion words from various web-pages from 21 Spanish-speaking countries (Davies, 2016-). Certain linguistic and extra-linguistic factors have been included in the analysis in order to tease out any patterns of use that may be similar or different across countries in the corpus.

Although there are limitations to a corpus study, the large amount of data gathered can serve as some basis for a wider understanding of which Spanish countries tend to see more use of the structure. To date, past studies (DeMello, 1992; Tieperman, 2020) that have investigated the structure have been small in sample size or have only focused on specific dialects. While the limitations of the *Corpus del Español* are such that it is impossible to easily divide countries into smaller subdivisions/dialectal regions, the large size of the corpus across

multiple countries should allow for some wider understanding of the relative frequencies of the structure across the Spanish-speaking world, at least at the level of different countries.

To further argue the benefits of conducting a corpus study, there have been calls from the field of onomastics for corpus-based studies of names. One article describes the potential benefit of using corpus linguistics for the study of names:

> Corpus linguistics provides powerful empirical methods for studying names in actual language use through frequency-based evidence. It can, therefore, serve to check and refine more traditional, normative descriptions of name usage as found in grammars and other reference works (Motschenbacher, 2020:98).

As will be discussed in the following chapter, the use of definite articles with anthroponyms is well documented in Spanish reference works (Demonte and Bosque, 1999, RAE, 2005), but there is a lack of corpus-based research on the phenomenon. As discussed by Motschenbacher in his article, such corpus-based studies can be very useful in order to verify and refine our understanding of certain usage patterns that, as in the case of this feature, are largely only documented in passing by reference works. Beyond the findings discussed in this study regarding the specific phenomenon at hand, it is my hope that this study will further affirm the benefits of corpus-based studies, especially with regards to uncommon linguistic phenomena.

This study will be divided into six chapters; the current introduction is followed by a review of past literature concerning the structure being studied here, which will then be followed by an overview of the methodology used in the corpus study, including a section discussing the efficacy of the mixed-effects statistical model used. After that, there will be an overview of the results, followed by a discussion of the results, and finally a conclusion with final remarks regarding the study.

Chapter 2

Literature Review

2.1 Semantics of Proper Names and Articles

Previous studies of the semantics and syntax of proper names provide valuable information that helps to guide this study.

In a 2006 article, Matushansky argues that proper names largely function the same as common names, except with the definite article:

> …What I claim is that proper names enter syntax with essentially the same semantics as common nouns (modulo an additional argument slot for the naming convention). This means that we expect them to have the same syntax as common nouns – which is in fact the case, with every determiner other than the definite article (Matushansky, 2006: 289).

Matushansky essentially argues that instead of viewing the absence of definite articles with proper names as the rule, perhaps it makes more sense to view this absence as a special opt-in rule. In other words, when it comes to proper names that appear with an article, instead of asking "why is this definite article here?" we should be asking "why isn't this definite article allowed to be omitted?" Matushansky points this out:

> One possibility is to reconsider where the default lies – it might be that the general view is incorrect: it is not that proper names of ships, rivers, etc., are exceptional in that they block m-merger[1] of the definite article, but rather that proper names of people, (in English) countries, etc., are special in that they are subject to preproprial m-merger (Matushansky, 2006: 299).

[1] M-merger is a term utilized by Matushansky to describe the merging of two syntactic heads; in this case, m-merger is the merging of the determiner with a proper name, yielding only the proper name in cases where m-merger occurs (Matushansky, 2006: 296-300).

Going forward with this perspective, Matushansky describes the contexts where definite articles must be overt (i.e., where they are not allowed to opt in

to what Matushansky calls the m-merger). The article must be overt if 1) the proper name is (restrictively) modified, 2) the proper name is pluralized, or 3) the proper name is lexically marked as requiring an overt definite article (Matushansky, 2006:295). It is important to note that while Matushansky (2006) focuses primarily on English, she recognizes that the lexically marked categories of types of names that require an overt article can and will differ between languages.

Matushansky provides examples of all three types of requirements:

- Restrictively modified proper names

  (2.1) This is not *(the) Elisabeth I know.

  (2.2) The gifts were sent by *(the) charitable Miss Murray.

  (Adapted from Matushansky, 2006:291)

- Pluralized proper names

  (2.3) The Clintons

  (2.4) The Alps

  (2.5) The Hebrides

  (Adapted from Matushansky, 2006:295)

- Lexically marked proper names

  (2.6) The Ukraine

  (2.7) The Matterhorn

  (Adapted from Matushansky, 2006:290)

Although Matushansky focuses on English and a few other languages, in section 2.3 we will see how these general rules hold true in the description of Spanish as well. We can already see the similarity between what Matushansky proposes and the rules referenced by the *DPD* in the introduction of this study, which state that definite articles can occur with anthroponyms that are plural or modified (RAE, 2005).

2.2 The structure in other Romance varieties

Although the use of definite articles with anthroponyms has received some attention in Spanish (Calderón Campos, 2015; Calderón Campos, 2018; Carranza Brito, 2008; Christodulelis, 2017; DeMello, 1992; Tieperman, 2020), the structure also exists in other related languages, although the amount of pre-existing sociolinguistic literature varies greatly from one language to the next. In Catalan, the definite articles *el* or *la* can appear with anthroponyms. Additionally, Catalan has personal articles (*en/n'* for masculine, *na/n'* for femenine) that are specifically reserved for use before names of people (Generalitat de Catalunya [GenCat], 2020aa).

These articles and their use vary throughout areas that speak varieties of Catalan. In general, these uses are not found as often in formal registers, and are most common in oral, colloquial speech (Coromina i Pou, 2001).

In an entry on the linguistic inquiry service Optimot, the Generalitat de Catalunya describes the use of the articles in Catalunya:

> Among the speakers of the Principality of Catalonia, generally the definite article is used with feminine names (whether they begin with a vowel or with a consonant) and with masculine names that begin with a vowel (la Maria, la Isabel, l'Antònia, l'Antoni); On the other hand, either the personal article or the definite are used with masculine names beginning with a consonant (en/el Pere). The personal feminine article (na) is not used commonly (GenCat, 2020a, Translation mine).[2]

On the other hand, the Generalitat then goes on to describe how in Balearic speech, the personal articles are used regularly with all names. In Valencian speech, on the other hand, the use of articles with names of people is less common (GenCat, 2020a). The following examples taken from a separate reference entry show sentences both with and without a definite/personal article.

(2.8) ***La Maria*** *m'ha dit que vindrà.*

'**The Maria** told me that she'll come.'

(2.9) ***Maria*** *m'ha dit que vindrà.*

'**Maria** told me that she'll come.'

(2.10) ***En Josep*** *sempre es queixa.*

'**The Josep** always complains.'

[2] Original text: En els parlars del Principat de Catalunya, generalment s'usa l'article definit amb els noms femenins (tant si comencen per vocal com per consonant) i amb els masculins començats per vocal (la Maria, la Isabel, l'Antònia, l'Antoni); en canvi, es fa servir l'article personal o el definit amb els noms masculins començats per consonant (en/el Pere). No se sol usar l'article personal femení (na).

(2.11) ***Josep*** *sempre es queixa.*

'**Josep** always complains.'

(Adapted from GenCat 2020b)

There is a variety of previous work done on the use of definite articles with anthroponyms in Brazilian Portuguese (Alves, 2008; Alves, 2017; Amaral, 2016; Christodulelis, 2013), much of it sociolinguistic in nature. Alves (2008) and Alves (2017) investigate the patterns of usage of this structure in the speech of young speakers from the city of Barra Longa who live in Belo Horizonte. In these studies, she found that the level of intimacy between the speaker and the referent was one factor that correlated with higher frequency of article appearance, which concords with studies of Spanish that will be discussed in section 2.4. Additionally, other relevant factors included the social networks of the participants and their attitude towards the distinct patterns of usage in the 2 cities. Participants that had more social connection to Barra Longa (where appearance of the article is less common) tended to maintain that pattern of usage (Alves, 2017: 86). The following examples show utterances where the definite article was used in their data.

(2.12) *Ela já tem um filho... **o Lucas** né? um menino...*

'She already has a son... **the Lucas**, right? A boy...'

(2.13) *Meu primo... **o Reinaldo** estudou comigo e continua lá...*

'My cousin... **the Reinaldo** studied with me and continues there...'

(Adapted from Alves, 2017)

Mendes (2015) takes a similar approach to Alves, studying the differing patterns of usage in two towns that are close in proximity; Abre Campo, where there is frequent use of the definite article with anthroponyms, and Matipó, where there is less frequent use of the article (Mendes, 2015: 357). Mendes also provides information regarding how this difference came to be through the use of historical documents from both towns, and finds that over an extended period of time, the use of articles with anthroponyms decreased in Abre Campo, and that both towns had fairly equal rates of usage before this change.

It is important to note (as does Mendes) that this structure is one that is expected to occur with a much higher frequency in oral language, rather than in written registers (Mendes, 2015: 343).

Somewhat in contrast to the findings of Alves, Amaral (2016) found that in three distinct regions of Brazil, the appearance or absence of an article was related to intimacy, but also tied to the dominant pattern of usage within the region. Speakers tended to use the dominant pattern (appearance vs. absence) for referents who were known and with whom they had a higher level of intimacy, such as "friends, parents, and neighbors," which Amaral provides examples for, such as example 2.14 (Amaral, 2016:119). On the other hand, speakers tended to use the non-dominant option when referencing people that they did not know well. Amaral argues that these data shows that these elements have pragmatic use, an idea which will be further discussed in section 2.4 (Amaral, 2016).

(2.14) ... ***a Andréia*** *faz poco tempo que ela tá aqui.*

'... **the Andréia** has only been here a short time.'

(Adapted from Amaral, 2016)

Christodulelis (2013) provides more data regarding the pragmatic function of definite article use with anthroponyms in Brazilian Portuguese. They found that, similar to what Matushansky (2006) described, articles are required when an anthroponym is modified. They also found that for their participants, the use of articles was "grammatically optional but pragmatically preferred" when the referent was mutually known by the interlocutors (Christodulelis, 2013: 49). In section 2.4.1 there is more discussion on the pragmatic uses of the structure in Spanish, which show a similarity between its use in Spanish and Brazilian Portuguese.

While the present study is concerned with the structure in Spanish and its similarities with other Ibero-Romance languages especially, it is worth pointing out that there is also work describing article use with anthroponyms in various dialects of Italian as well. While the structure varies in its use regionally, it is interesting to note that multiple sources describe a general preference for using articles with names of females over males (Marcato and Thüne, 2002; Maiden and Robustelli, 2007). This aspect of the structure's use will be discussed in section 2.3.2, as Spanish has also been described as having a similar pattern of usage.

2.3 Spanish descriptive grammar entries

Although there is a general lack of academic studies concerning the specific phenomenon of definite articles appearing with anthroponyms in Spanish, there

are mentions of the phenomenon in the *Gramática descriptiva de la lengua Española* (henceforth the *Gramática descriptiva*) (Demonte and Bosque, 1999).

2.3.1 Null articles with proper names in Spanish

Several entries in the *Gramática descriptiva* deal with larger categories of proper names, including more than just anthroponyms: names of mountains, bodies of water, boats, islands, companies, and organizations, and more. We can look to a few entries that describe why these names might appear with or without a determiner, to see whether it supports what both Matushansky (2006) and the *DPD* say.

In a chapter largely describing concordance in Spanish, Martínez García (1999) describes criteria for predicting article appearance that largely match with the criteria described in Matushansky (2006). He says that last names with articles normally appear as lexicalized units (*Del Campo*), or pluralized (*los López*, *los Martínez*), while given names normally have an article only when they appear with "un adjetivo o unidades análogas," such as *la famosa Reyes*, or *el antedicho Natalio* (Martínez García, 1999: 2717). These examples line up with the rules described by Matushansky (2006); We can expect overt articles to appear with restrictively modified proper names (*la famosa Reyes*), with pluralized proper names (*los López*), or when the plural name is a member of a specific class.

For the final criterion regarding specific lexical categories, Martínez García does recognize that patterns of article occurrence seem to be less strict when it comes to certain groups of names, such as toponyms. They give many examples where the article can be considered optional: *(el) Japón*, *(el) Perú*, *(el) Cuzco*, *(la China)*, *(la) Florida*, etc. (Martínez García, 1999: 2717-2718). These lexical items may vary in their appearance with an article from one speaker to the next.

Martínez García also describes how many times these lexical items with optional articles, as well as many with obligatory articles, are part of classes that are connected to a common noun. For instance, in the case of *El Sahara*, we could assume that there is ellipsis of a noun *desierto*; *El (desierto) Sahara* (Martínez García, 1999: 2718). This is also mentioned as a possible explanation by Matushansky (2006), with examples such as *the Thames (river)*, or *the Titanic (ship)* (Matushansky, 2006: 300).

In a chapter regarding the presence or absence of a determiner, Laca (1999) also describes how certain lexical categories of proper names take an article, while others do not, and still others vary. For instance, bodies of water always have a determiner (*el Atlántico*, *el Nilo*), as do mountain chains (*los Andes*, *los Alpes*). On the other hand, cities and towns normally do not have an article,

unless it is a lexicalized part of the name (*La Coruña*, *La Habana*). Laca, like Martínez García, also describes categories that can vary; certain countries at times may appear in speech with an article, such as *(la) Argentina*. Interestingly but perhaps unsurprisingly, Laca points out that many manuals in both Latin America and Spain discourage the use of articles with states and countries (Laca, 1999).

In these chapters, it becomes apparent that the rules proposed by Matushansky (2006) seem to largely reflect the situation in Spanish as described in the *Gramática descriptiva* with regards to when a proper name will have an overt or omitted article; restrictively modified proper names and pluralized names will generally appear with an article, and certain types of lexical classes (such as bodies of water and mountain chains) will also tend to appear with an article. There is variation with many specific lexical items, especially with toponyms, which could be related to dialectal variation.

2.3.2 Anthroponyms with articles in Spanish

Having seen that the aforementioned rules of overt article appearance seem to largely apply as described to Spanish in general, we can narrow our focus to personal names. If we assume that the first two rules of article requirement (of being restrictively modified or of being pluralized) apply to anthroponyms, we can assume that in general, anthroponyms will have an overt definite article when restrictively modified or when pluralized. We could inversely assume that we would not expect to see an overt article with singular, unmodified anthroponyms.

However, the basis and inspiration for the present study is the fact that unmodified, singular anthroponyms can appear with definite articles. Multiple entries in the *Gramática descriptiva* confirm this. In a footnote in their chapter on concordance, Martínez García writes:

> Coming from rural speech and extending into the vulgar register of the language, the article accompanies first names, both male and female, contributing a value that is understood as derogatory and discourteous for the named person... that, in contrast, is not expressed with a last name, where the gender of the article is reduced to solely indicating the sex of the named person (Martínez García, 1999:2717, translation mine).[3]

[3] Original text: Proveniente del habla rural y extendido en el registro vulgar de la lengua, el artículo acompaña a los nombres de pila, tanto de mujer como de varón, aportando un valor que se entiende como despectivo y descortés para con la persona nombrada… que, en cambio, no se expresa con un apellido, en el cual el género del artículo se reduce a indicar el sexo de la persona nombrada…

In the following section, I will briefly discuss other studies that contradict the idea that such uses are always disrespectful in nature, but for the purposes

of this study, the key section of the previous citation is the reference to "habla rural y extendido en el registro vulgar." This is an idea echoed in other entries.

Beyond describing other types of proper names and their occurrence with articles, Laca (1999) describes the same phenomenon of articles appearing with personal names:

> The use of articles has strong connotations with spoken language, especially of a lower socio-cultural level, except in regions influenced by Catalan and Portuguese, where it is more general (Laca, 1999:924, translation mine).

In this case, Laca reaffirms that this phenomenon is linked to spoken language, and that it is associated with speech of a lower socio-cultural level. Interestingly, they also put forward that the phenomenon is more widespread in areas with contact with Catalan and Portuguese. Unfortunately, they do not give additional detail regarding this geographical variation.

Other entries in the *Gramática descriptiva* make mention of the register in which the phenomenon occurs, which is, in general, spoken and colloquial speech. One mention of this comes from Rigau (1999):

> In colloquial contexts it is not rare to see the presence of a definite article before the proper name of a person (la Lola, el Pepe). It is treated as an expletive use of the article... In other Romance languages this element becomes obligatory and in some cases a special article is used for these types of proper names, for example, in Catalan (En Pere «Pedro» , Na Caterina «Catalina») (Rigau, 1999:321, translation mine).[4]

Other Romance languages were discussed previously in section 2.2, but as far as "colloquial contexts," there are other entries that corroborate this, such as in Fernández Leborans (1999):

> ... It is necessary to consider... that the article with an unmodifed NNPP [proper name] is a case of an 'expletive' article, as in Italian: *Il Gianni/Gianni mi ha telefonato*. In Spanish, the presence of the definite article with a NNPP [proper name] has been classified as familiar or colloquial (*La María*; *el Antonio*). More generalized is the use of the definite article with last names of women (*La Garbo*; *la Thatcher*) (Fernández Leborans, 1999:112-113, translation mine).[5]

[4] Original text: En contextos coloquiales no es rara la presencia del artículo definido ante los nombres propios de persona (la Lola, el Pepe). Se trata de un uso expletivo del artículo… En otras lenguas románicas este recurso llega a ser obligatorio y en algunas incluso se utiliza un artículo especial para este tipo de nombres propios, por ejemplo, en catalán (En Pere «Pedro» , Na Caterina «Catalina»).

[5] Original text: … Habrá que considerar… que el artículo con NNPP [proper name] no modificados es un caso de artículo 'expletivo,' como en italiano: *Il Gianni/Gianni* mi ha telefonato. En español, la presencia del artículo definido con NNPP se ha calificado de familiar o coloquial (*La María*; *el Antonio*). Más generalizado está el uso del artículo definido con apellidos de mujer (*La Garbo*; *la Thatcher*).

In line with what other authors have mentioned and what is discussed in sections 2.1 and 2.3.1, Fernández Leborans (1999) does explicitly mention "NNPP no modificados" (unmodified proper names) in order to distinguish these uses from those cases where the proper names are modified (or pluralized). While the latter is fully expected in standard Spanish, the colloquial use of articles with unmodified anthroponyms (which will be referred to primarily as "colloquial article use") is the primary focus of this study, and sources such as Fernández Leborans (1999) and Rigau (1999) show that this use is recognized, but considered to be distinct from other expected uses. It is also important to note the mention that Fernández Leborans makes of the last names of women appearing more generally with articles. The *DPD* section cited in the introduction of this study also makes mention of this pattern, stating that "la anteposición del artículo, en estos casos, suele ser propia del habla popular… especialmente ante nombres de mujer…" (RAE, 2005). Additionally, As discussed briefly at the end of section 2.2, there are mentions of a similar pattern of use in some dialects of Italy (Maiden and Robustelli, 2007). Unfortunately, in the entries of the *Gramática descriptiva* there is not much discussion of the basis for these claims, and the current study aims to provide quantitative data to better understand the gender aspect of the structure.

Even with the number of chapters in the *Gramática descriptiva* that recognize the existence of the phenomenon of definite article use with (unmodified) anthroponyms, there is a lack of information regarding some aspects of the structure, such as its geographic distribution. There are mentions of it being widespread in colloquial speech, but the only mention of a specific region with more use of the feature in these entries of the *Gramática descriptiva* is the singular mention of more use in regions influenced by Catalan and Portuguese (Laca, 1999: 924).

This is a large part of the basis of the current study; filling in this gap in our knowledge could help in providing future paths to research on this structure of colloquial Spanish.

2.4 Other studies on the structure in Spanish

Beyond the various entries of the *Gramática descriptiva* that discuss the use of definite articles with anthroponyms, there are other sources that have analyzed the structure.

2.4.1 Pragmatic uses of the structure

Carranza Brito (2008) proposes a "continuum of individuation" to describe the various situations in which an article may or may not appear. Relevant to the present study, Carranza Brito (2008) argues that article use with anthroponyms (at the very least with given names) can be read in three distinct ways, depending on context. These are:

- To make the name in question more "eligible"; to specify a known individual for the purpose of the discourse.

  (2.15) A. *Quien ya cayó de mi gracia fue* ***la Merced****.*
  B. *¿Cuál Merced?*
  A. ***La Merced*** *chica. No la hermana, sino la hija.*

  A. 'The one who fell out of my grace was **the Merced**.'
  B. 'Which Merced?'
  C. '**The Merced** girl. Not the sister, but rather the daughter.'

- To show a derogatory attitude towards the referent.

  (2.16) *En eso que llega* ***la Gabriela****, muy sangrona, como siempre, creyéndose dueña y señora del lugar.*

  'In that moment **the Gabriela** arrived, unpleasant as always, thinking herself owner and mistress of the place.'

- To show an appreciative attitude towards the referent.

  (2.17) *En eso que llega* ***la Rosy****, vino a felicitarme por mi cumpleaños, me trajo mi regalito. Ya después nos fuimos todos a cenar.*

  'In that moment **the Rosy** arrived, come to congratulate me on my birthday, and bringing a gift. After that we all went out for dinner.'

(Examples adapted from Carranza Brito, 2008)

With these examples, taken from colloquial conversations collected in Mexico City, it becomes clear to see how a definite article can appear even with a singular, unmodified first name. Moreover, the proposed readings of these uses of the article go beyond just mentioning the existence of this structure.

Calderón Campos (2015) argues for an interpretation similar to that of Carranza Brito (2008). In this article, he discusses many aspects of the structure,

including its history and uses in literature, and, most relevant to the current study, the uses of the structure in written texts that imitate spoken language. He points to three uses of the definite article with personal names:

> The first and most general presupposes intimacy or closeness towards the designated individual; to this can be added subjective tones of disdain or praise, or connotations that label the person as belonging to a low socio-cultural group (Calderón Campos, 2015:90-91, translation mine).[6]

[6] Original text: El primero y más general presupone intimidad o cercanía respecto del individuo designado; a este valor se pueden añadir matices subjetivos de desprecio o elogio, o connotaciones que etiquetan al personaje como perteneciente a un nivel sociocultural bajo.

These uses of the article with anthroponyms are very similar to the uses described by Carranza Brito (2008), with the additional cited use of showing that the referenced individual belongs to a low sociocultural class.

From a historical perspective, some important data can be gleaned from this study. Using the corpus *CORDE*, Calderón Campos (2015) shows that the structure had its highest occurrence in the 16th century. Although the number of cases that he found decreases over the next centuries, Calderón Campos proposes that this has to do with the change over time in the popularity of the literature where the structure appears, and, relevant to the current study, Calderón Campos states that over the centuries we should expect that the structure has survived in everyday speech (Calderón Campos, 2015: 85-86).

In a presentation on her work concerning the acceptability of this structure in Spanish, Christodulelis (2017) describes her findings that article use was perceived by a variety of native Spanish speakers to be more "intense," and additionally found that there was a correlation between acceptance of the article and emotionally charged contexts (Christodulelis, 2017: 39). Although this does largely support the proposed negative and positive functions of the structure proposed by Carranza Brito (2008), the findings of Christodulelis (2017) bring attention to the perception of the structure's use, especially when it comes to intensity and other factors that have not been discussed in other studies.

While the current study is not concerned with attempting to classify or analyze the pragmatic or discursive functions of article use with anthroponyms, it is important to note that these past studies have shown that this structure appears many times with pragmatic value, even if the article could otherwise be removed from an already grammatically acceptable sentence.

While the current study is more concerned with the general patterns of use at a very large geographical scale, there is also a lack of socio-pragmatic understanding of the structure in Spanish. The present study hopes to provide information that would serve as important background or justification for more narrowly focused (socio-)pragmatic studies in the future.

2.4.2 Corpus Studies

One of the earlier studies that takes a dialectological approach to studying the use of definite articles with anthroponyms is in De Mello's 1992 study, in which they use data collected from "habla urbana culta" from 10 cities (DeMello, 1992: 222). In his data, even with only a relatively small sample size (135 tokens), some clear patterns emerge; the Santiago (Chile) set contained a large majority (98 of 135) of the cases of article use, and a majority (104) were also found with female referents. Unfortunately, and as noted by De Mello, the expectation is that the use of articles with anthroponyms will be much lower in higher-class speech, and De Mello's tokens were extracted from this type of data (DeMello, 1992: 228). This means that, while certainly giving some insight into the frequency of use in Chile, De Mello's data may not accurately represent overall usage patterns in the Spanish-speaking world. Even so, his data does show that there are unexpected (pragmatic) uses of the article in multiple countries, as seen in the following example from De Mello's Buenos Aires data.

(2.18) ***El Lolo*** *estaba tan contento que...*

'**The Lolo** was so happy that...'

(Example adapted from DeMello, 1992)

One of the studies that most closely resembles the current study is a socio-pragmatic study carried out in an unpublished master's thesis by Tieperman (2020). Their study had two components: a corpus-based study, and a survey-based study. In both components, they used data from three cities: Granada (Spain), Mexico City (Mexico), and Santiago (Chile). Using the *PRESEEA* corpus, they included several factors as part of the corpus-based study; city, referent gender, participant gender, social class, and age (Tieperman, 2020: 23).

In the results of their corpus-based study, Tieperman found with a sample size of approximately 1,000 total tokens (including those with and without a definite article), that the significant factors correlating with a higher use of definite article appearance were city, referent gender, and class. For city, Santiago had the highest rates of usage. For referent gender, the use was higher for female referents. For class, the use was higher among middle- and lower-class participants. Importantly, Tieperman also points out that in Santiago, a majority of tokens appeared with a realized article, whereas it was disfavored in Mexico City and Granada (Tieperman, 2020: 39).

The results of Tieperman's (2020) study add more evidence that article use before (unmodified, singular) anthroponyms is higher among lower-class

speakers, that it is higher with female referents, and that it is higher in certain areas, in this case Chile.

Chile and areas of contact with Catalan or Portuguese are often cited as areas with higher usage of definite articles with anthroponyms (DeMello, 1992, RAE, 2005, Laca, 1999), but the results of Tieperman (2020) show that even in parts of Mexico and Spain the structure occurs, albeit with a much lower frequency. This should not be entirely surprising, as other studies have investigated the feature in these areas (such as Carranza Brito in Mexico City), but Tieperman's study is one of few recent studies that take an approach that looks at wider dialectal variation.

The current study aims to provide an even larger picture than the one provided by Tieperman (2020); while Tieperman's corpus data contained approximately 1,000 tokens, the current study includes analysis of a much larger sample size (over a quarter of a million total tokens for 98 names) with the hope of shedding light on patterns of usage across many countries in the Spanish-speaking world.

At a broader level, there have been calls for furthering the use of corpus linguistics methods in the study of names. Motschenbacher propounds the benefits of corpus-based studies in onomastics. He briefly discusses some previous corpus-based studies that have examined patterns of usage with names, and then argues that those studies have given us more nuanced views of name usage, in contexts that reflect actual language use (Motschenbacher, 2020:97). One study discussed in the article (also conducted by Motschenbacher), even focused on a similar topic to the present study: the co-occurrence of definite articles with English country names (Motschenbacher, 2020:97).

Motschenbacher also argues that due to the large size of many corpora, an additional benefit of corpus-based investigation of names is the ability to find and study unexpected or infrequent structures. For that reason, he argues that "Corpus linguistics is, therefore, especially useful for the investigation of names and name categories that are less prototypical in their grammatical behavior" (Motschenbacher, 2020:98). The use of definite articles with anthroponyms in Spanish is one such structure, and the large quantity of data contained in the *CdE* will help in both understanding broader usage patterns of the structure, while also providing a more complete picture than what we currently know from Spanish grammars and reference works.

Chapter 3

Methodology

3.1 Research questions

With the current dearth of studies that take a dialectological approach in studying patterns of usage with the definite article and its appearance with anthroponyms, the current study is intended to build a wider foundational knowledge of the general patterns of usage across several Spanish-speaking countries. The following questions present the general focus of the current corpus study:

> Question 1: How much does the frequency of definite article use with anthroponyms vary between different Spanish-speaking countries, and which countries have the most/least use?
>
> Question 2: Are there specific linguistic factors that correlate with more frequent use of definite articles with anthroponyms? Additionally, do these factors differ from country to country?

As seen in previous studies (DeMello, 1992, Tieperman, 2020), there is evidence of variation, but in these studies the number of dialects and number of tokens were limited. The current study takes a significantly larger number of tokens in order to find consistent large-scale patterns. Data for this study were collected from the *Corpus del Español* (*CdE*), specifically from the Web/Dialect corpus, which contains over 2 billion words.

3.2 Data collection methods

The data for this study were extracted from the *Corpus del Español: Web/Dialects* (Davies, 2016-) by using the computer program *PuTTY* (Tatham et al., 1997-) and the *CQP Query Language* (Evert and Hardie, 2021-) to conduct queries on

a version of the *CdE* hosted on a University of Georgia server. The exact code used appears in appendix A and is discussed specifically in section 3.2.3.

3.2.1 Information about and limitations of the *Corpus del Español*

The *Corpus del Español* was originally created by Mark Davies in 2001, utilizing a tagger called *Palabras*, which was created by Eckhard Bick and based off his earlier tagger *Palavras* for Portuguese (Bick, 2000). This system of automated tagging was then hand-corrected at BYU by Davies in order to create the *CdE*. The system of tags incorporated into the *CdE* assists linguists in searching for different words, parts of speech, and collocates, among other functionalities.

This study utilizes the *Corpus del Español: Web/Dialects*, a newer corpus added in 2016 that has over 2 billion words taken from 21 Spanish-speaking countries (Davies, 2016-). The data are sorted by country using Google's country information provided on web pages and includes text from blogs and other types of web pages.

One limitation of this corpus is that the texts do not necessarily closely represent spoken Spanish, and the structure at question here is mostly reported to occur in colloquial, conversational Spanish. That being said, with the relatively non-formal nature of many of the sources, such as blogs, it is expected that there will still be occurrences of article use with anthroponyms. It is believed here that the relative frequencies seen in the data sets from each country will be generally representative of overall trends, even if those frequencies are lower overall than they might be in oral data.

On the other hand, an advantage of using the *CdE* is the large amount of data that it contains. The smallest country data set extracted for this study (Paraguay) had over eight thousand tokens, and the largest (Spain) had well over one hundred thousand tokens. With this amount of data, it becomes infeasible to analyze each case fully on a qualitative level as done in past studies (such as DeMello, 1992), but it is possible to carry out large quantitative analyses of the overall trends in order to glean a general picture of the structure, at least as far as this specific register (blogs and other web pages) is concerned.

To give an idea of the distribution and quantity of all words contained in the *CdE* for each country, table 3.1 shows the percentages for each country, in terms of total words from each country compared to the totality of all approximately 2 billion words of the *CdE*.

Relating to the distribution of words from each country arises another limitation of the *CdE*. As several of the cited studies of Brazilian Portuguese (Alves,

Table 3.1: Percentage of the *CdE* made up by each of the selected eight countries.

Country	Number of words (*CdE* total = 2,100,761,228)	Percentage of *CdE*
Argentina	182,704,898	8.70%
Chile	70,598,279	3.36%
Colombia	180,145,658	8.58%
Cuba	67,655,690	3.22%
Mexico	260,598,272	12.40%
Paraguay	32,990,144	1.57%
Peru	115,324,436	5.49%
Spain	459,312,821	21.86%

2008, Alves, 2017, Mendes, 2015) show, even cities or towns that are relatively close in distance to each other may have distinct patterns of usage when it comes to this structure. In most previous research done in Spanish with the structure, the focus has been only on speakers from singular cities in different countries (DeMello, 1992 Tieperman, 2020). Little is known about what variation of the structure could exist in different countries at the level of regions, cities, or even neighborhoods. It is possible that even Chile, which is pointed to as having a high frequency of the structure, could have significant variation between specific regions or cities.

This study unfortunately does not include analysis of specific regional varieties at a level more detailed than country, due to both the scale of the data and the nature of data tagging utilized by the *CdE*. This is an important limitation to recognize, and future studies should examine variation between regions or towns, for example.

3.2.2 Other limitations concerning data extraction

The *CdE*'s system of POS (part of speech) annotation allows for sorting and searching the corpus for various attributes, at both a linguistic and extra-linguistic level. This includes a tag that allows for searches of proper names (Bick, 2000).

Unfortunately, the proper name annotating for the *CdE* includes many types of proper names beyond anthroponyms, and there is not a more specific type of POS annotation for anthroponyms only. As a result, executing a normal search limited to proper names yields thousands of token types that are not

anthroponyms, and sorting these out efficiently turned out to be infeasible using the built-in POS system of the *CdE*.

Another limitation of the *CdE*'s annotations for this study involves the distinction between given names, surnames, and nicknames. In the *CdE*, there is no way using the built-in annotations to automate the process of distinguishing between a given name, a surname, or even a nickname.

As a result of the above limitations to the otherwise robust *CdE*, the data extraction was limited to searches using a set of manually curated first names that are popular in Spanish. Ideally, data could be used to compare the rates of article appearance with different types of names, but for the purposes of this study, the data were limited to a more controllable set of names, in the hopes of keeping data consistent throughout the larger amount of data extracted.

How the list of names was created

The curated list of first names was created using a large data-set of names created by a user of an online data forum (Cruz, 2017). The original list of names contained approximately 1,600 different female names and 1,500 different male names. These names were then used in a general search of the entire corpus to rank all 3,100 names in order of their frequency, measured in words-per-million (WPM).

The final list of names used for the queries in data extraction consisted of 98 of the most popular given names. Exclusions were made of several dozen names and corrections were made for some inaccuracies involving diacritics. The entire list of excluded names can be found in Appendix B, and the final list of the 100 names used can be found in Appendix C. These names were excluded because they were not always used as a Spanish first name: some are borrowings from other languages (such as *Edward*, *Jackson*, and *Washington*), some also share a common noun meaning (*Trinidad*, *Niño*) and likely were present near the top of the list due to incorrect tagging, and others were excluded for also sharing the name of a well-known geographical location (*Europa*, *Galicia*, *Valencia*, *Siria*, and *Washington* once again). There are undoubtedly some cases where these names are used as given names in Spanish (and thus were included in the original data set), but they present ambiguities and in many cases are over-represented in a data set that aims to narrow data down to the given names of people. For instance, the name with the highest frequency above all others was *Dios* (even more than quasi-shibboleths in the Spanish-speaking world such as *María* or *Juan*), and this is almost certainly due to references to God, rather than to individuals with the given name of *Dios*. These names were excluded for the sake of keeping the large amount of data more consistent.

It is important to note that this list of popular names does not necessarily reflect the actual popularity of these names in the various countries included here. However, it became clear in the process of searching for sources regarding names that it would be impossible to find a single, methodologically consistent source of names for all the countries included in this study. As a result, this study instead has opted to use the frequency of the names within the corpus itself as a general representation of the popularity of these names. An unfortunate limitation of the current study is that it is likely that the frequencies in the corpus differ from the actual frequencies of given names, especially among countries, and future studies should attempt to rectify this issue. A related issue is the fact that this method may exclude names which are culturally relevant in certain regions but uncommon otherwise, such as the name *Montse* in Catalonia.

The selection of the names was based on the overall frequency of the names in the entire *CdE*, but after the selection, the list of names was applied separately to the section of each country studied here to determine the words-per-million (WPM) frequency of the names in each country. These frequencies were used in the analysis of the data for the factor of name frequency, and the frequencies used were based on the frequencies of the names in each country. The full table of names and their frequencies can be found in Appendix C, but the top 10 names of the overall corpus and their corresponding WPM values are represented in table 3.2.

Table 3.2: Words-per-million frequencies of the top 10 most frequent names.

Name	*CdE* Total	Argentina	Chile	Colombia	Cuba	Mexico	Paraguay	Peru	Spain
Juan	174.08	195.36	140.87	229.02	155.51	148.72	241.71	178.91	113.19
José	153.01	112.12	107.19	139.72	288.43	129.27	187.78	162.64	124.5
María	118.06	121.68	84.82	130.08	128.1	116.08	154.69	169.25	95.49
Carlos	115.38	142.4	94.92	146.28	147.32	93.88	171.41	138.7	71.86
Luis	96.45	94.08	68.82	117.59	136.32	95.98	119.05	126.55	69.4
Pablo	71.69	88.7	102.31	87.71	64.01	74.93	68.88	55.13	43.82
Pedro	69.87	53.02	65.34	57.48	76.96	68.15	133.73	67.82	49.98
Francisco	69.26	74.36	63.17	62.21	62.19	60.37	93.92	61.32	55.21
Manuel	61.45	47.95	49.95	87.37	88.93	61.63	66.81	60.37	51.67
Miguel	60.62	60.63	41.85	44.36	72.34	71.29	63.64	73.42	51.41

3.2.3 Choices made for data extraction

After creating the curated list of popular names, the names were then used as the query for a search to extract all relevant occurrences of those names occurring

in the given countries. The *CQP* code used to extract the data can be found in Appendix A.

Data were extracted from 8 countries: Argentina, Chile, Colombia, Cuba, Mexico, Paraguay, Peru, and Spain. These countries were chosen to represent dialects from a variety of regions, and also for many due to their large corpus sections.

Importantly, the code used for extraction limited data to cases where a name appeared with no other name preceding it or following it. This is a limitation, as it disallows inclusion of complex names (such as *María Jesus* or *José María*), but unfortunately the current methodology is limited due to the structure of the *CdE* when it comes to distinguishing between names that are complex given names (*el José María*) or full names (*el Juan Rodríguez*). If such a distinction were to be included in the analysis, there would need to be a way to discern between cases of complex names and full names, and without manually checking every example, this would not be possible. Because of the lack of methods to do this using the *CdE*'s built-in system of annotation, the data were limited only to names from the name list appearing with no other names preceding or following them.

It also merits mention that, initially, name type was intended to be included as a linguistic factor, with the hope of discovering any potential significant correlation between article presence/absence and name type (given name, complex name, full name, nickname, etc.). However, in initial searches there were problems with including this factor, arising from the aforementioned limitations of the *CdE*'s annotation system, as well as other ambiguities, such as not having a way to discern the referent's gender for some nicknames (*Dani* for *Daniel*/*Daniela*, *Gabi* for *Gabriel*/*Gabriela*, etc.). Because of these issues, the factor of name type was excluded, and names were limited to the curated list of given names. Future studies should include name type in their analysis where possible, as it is very possible that names appearing in less formal contexts (such as nicknames) would see higher article presence.

3.3 Factors

Besides limiting the data extraction to cases where the names appeared in their singular form (not pluralized) and were not preceded or followed by any other name, no other limitations to part of speech (or other tags) were applied to preceding or following words.

After the data were extracted, the tokens were checked for preceding and following positions for each instance, looking for elements that matched the factors used.

The independent factors were chosen with the intent of finding patterns of variation pertaining to different linguistic aspects of names and their appearing contexts, as discussed in the rest of this section. These factors were coded by making use of the POS tags of each word in the *CdE*'s annotation system.

Three of the factors chosen relate to the nature of the names used: the frequency (in WPM) of the names, the gender of the names, and whether the names are consonant or vowel initial.

The gender of the names was chosen as a factor due to the mention of gender (specifically female names) in several sources as being a predictor of higher article presence (Fernández Leborans, 1999, RAE, 2005).

The frequency of the names was chosen in order to gain some understanding about whether or not the prevalence of a name affects its likelihood of appearing with an article. This factor was coded by using WPM data for each name in each country that was extracted from the *CdE* separately from the main dataset. The WPM values calculated for the names within each country's corpus section can be found in appendix C. This factor was chosen not for any specific reason pertaining to prior research, but rather because it seems plausible that there could be an effect of markedness at play, and measuring the statistical significance of name frequency with article occurrence could capture such an effect.

The following two examples are taken from the Spain dataset, and show an example of article use with a more common name (*Juan*), and a less common name (*Lara*).

(3.1) *Si haces una ampliación **del Juan** que capturé con mi cámara fotográfica, acaba siendo un pixel.*

'If you enlarge **the Juan** in the picture I captured with my camera, it ends up being a pixel.'

(3.2) *Se va a ver a Gina todo el tiempo, prepara la comida favorita de Gina a Lara y guarda una parte para Gina, celebra su fiesta en casa de Gina en lugar de la suya o **la Lara**...*

'They got to see Gina all the time, preparing Gina's favorite food for Lara and saving part for Gina, having their party at Gina's house instead of their own or at **the Lara's**...'

Examples from the *Corpus del español: Web/dialects* (Davies, 2016-).

The factor of consonant or vowel initial was included to see whether any phonotactic effects are at play. There is no mention of this as having an effect in Spanish in any of the grammar entries cited in this study, although interestingly Catalan is noted as having some phonotactic preferences (GenCat, 2020).

Five factors were chosen dealing with syntactic elements appearing before or after the queried names. These factors include: preceding *a* or *de*, following 3rd person singular verb form, following *que*, following *de*, and an adjective following the name.

(3.3) *Yo siento que perdí **a la Claudia que** era antes y quiero recuperarla....*

'I feel that I lost **the Claudia that** was before and I want to get her back...'

(3.4) *... Es la primera vez que me gustan TODOS los temas; es algo raro, es **el Ricardo de** siempre pero renovado.*

'... It's the first time that I liked ALL of the topics; it's strange, it's **the same Richard as always** but renewed.'

(3.5) *Y entonces, me volví a sentir cono [sic] **la Patricia adolescente**, mal en su piel...*

'And then, I returned to feeling like **the adolescent Patricia**, uncomfortable in her skin...'

(3.6) *En los últimos carnavales, cuando **la Isabel estaba jugando** carnaval con todos nosotros, oyó que mi tía y su esposo gritaban.*

'In the last carnivals, when **the Isabel was playing** carnival with all of us, she heard my aunt and her spouse screaming.'

(3.7) *Primeramente, la respuesta **del Guillermo que** usted menciona fue borrada porque no se ajustó a la simple regla de este foro.*

'First, the response **from the Guillermo that** you mentioned was erased because it didn't comply with the simple rule of this forum.'

(3.8) *Era tan distinto **al Daniel de** siempre, yo estaba sorprendida.*

'He was so different **from the normal Daniel**, I was surprised.'

Examples from the *Corpus del español: Web/dialects* (Davies, 2016-).

Some of these factors were chosen to capture some of the common syntactic configurations that are expected to yield higher frequencies of article presence;

names followed by *que* (such as in examples 3.3 and 3.7), *de* (such as in examples 3.4 and 3.8), or an adjective (such as in example 3.5) are likely modified, and thus, according to past literature, would be expected to appear with an article (Matushansky, 2006, Martínez García, 1999, RAE, 2005).

The factor of whether or not the names were followed by a conjugated verb in the 3rd person was included to gather cases where the name is acting as a subject (such as in example 3.6); if the name is directly followed by a conjugated 3rd person, it is likely that it is the subject of the verb, and it's possible that whether or not the name is a subject in the sentence could have a correlation with article presence.

The factor of *a* or *de* preceding the name or article (such as in examples 3.3, 3.7, and 3.8) was included in case there are any unexpected effects involving the conjoined forms of the two prepositions (*al* and *del*).

Unfortunately, the annotations currently implemented in the *CdE* do not allow for the analysis of a broader set of syntactic and semantic factors that one might normally use in a more focused and qualitative study; the *CdE* does not have tagging for phrases or semantic roles, thus any analysis must rely on specific lemmas or part-of-speech elements. This is an important limitation to recognize. While the *Corpus del Español* has many benefits (not the least of which is its immense size), it lacks certain levels of annotation that would allow for more complex syntactic and semantic analysis.

Table 3.3: Summary of the factors analyzed in the study

Name-specific factors	Name frequency	The frequency of the name in the country section the token was taken from, in words-per-million. This is a continuous variable, with values ranging from low single digits to around 200 WPM, varying from country to country.
	Gender	The gender traditionally associated with the name, either male or female.
	Consonant or vowel initial	Whether the name begins with a vowel or a consonant.
Context-specific (syntactic) factors	Preceding preposition	Whether the name (or the article, in cases where an article is present) is preceded by the preposition *a* or *de* (or *al* and *del* in the case of article appearance). The 3 possible values are absence, presence of *a*, and presence of *de*.
	Following 3ps Verb	Whether the name is followed by a verb conjugated in a 3rd person singular form. The two possible values are absence or presence.
	Following *que*	Whether the name is followed by the word *que*. The two possible values are absence or presence.
	Following *de*	Whether the name is followed by the word *de*. The two possible values are absence or presence.
	Following adjective	Whether the name is followed by an adjective. The two possible values are absence or presence.

3.4 Hypotheses

Below are the three hypotheses formulated based on the literature discussed in Chapter 2 and on the discussion of the factors in the previous section:

- Hypothesis 1: Of the countries included in this study, Chile will have the highest frequency of article presence.
- Hypothesis 2: Female names will have a higher frequency of article presence than male names.
- Hypothesis 3: Names followed by *que*, *de*, or by an adjective will have higher frequencies of article presence than names that are not followed by those elements.

3.5 Data analysis methods

Data analysis was conducted using R (R Core Team, 2021) with RStudio (R Studio Team, 2021). After using Excel to create spreadsheets housing all the tokens with the various values for each factor, these spreadsheets were imported as CSV files into R.

Using the Lme4 package for R, various generalized linear models for analysis were created for each dataset. During the early stages of analysis, this included multiple fixed-effects models, as well as mixed-effects models that included name as a random effect.

While comparing these models, there were surprising differences in the factors that proved to be significant in predicting article presence. For instance, in the original fixed-effect models, the factor of name frequency showed that in five of the eight countries higher frequency of a name disfavored article use. However, in the mixed-effects model (with name as a random effect), there was no significance of name frequency in any of the eight country's datasets. Similar differences appeared with other factors as well, most notably gender, which will be discussed in the following section.

These differences have the potential to drastically change the analysis of the results; some factors that are significant in most countries with one model are not significant in any with another model.

After examining the reason behind these differences, the mixed-effects model was selected as the model of choice for all of the eight datasets. In the following section, I explain why I believe the mixed-effects model (with name as a random effect) is a better fit for the analysis in this study.

3.5.1 Fixed-effects model vs. mixed-effects model

Random effects are often used in linguistics for the variable of "speaker" in corpora created from interviews with various participants. In such cases, accounting for speaker as a random effect assumes that the patterns of variation of a particular feature within a dataset might be more related to the grammar of a particular speaker or group of speakers, as opposed to the grammar of all speakers of that language.

I argue that a similar situation can be seen in this study: particular names might have greater tendency to favor or disfavor article presence, not due to any other factors (such as the associated gender of the name) but simply because certain names might have unique idiosyncrasies in regard to the feature. Including name as a random effect is intended to account for name-specific variation and thus reveal significant predictive factors that are not name-specific, but rather specific to each factor.

Additionally, as noted by Barth and Kapatrinski (2018), random effects can be implemented in mixed-effects models to better fit datasets that include unbalanced groups, such as words in a Zipfian distribution. In their chapter on mixed-effects in corpus linguistics, they note:

> In corpus data, the items that are more frequent in the language will be observed more frequently… To the extent that this is true, high-frequency items are not just more likely to deviate from the sample mean than low-frequency items; they are also likely to deviate from the mean in different directions than low-frequency items. Thus in a corpus study, the sample is biased to oversample items that are likely to be exceptional (Barth and Kapatrinski, 2018:100).

In the case of the present study, a large number (98) of names is used, and the frequency of cases for each of the names does represent a Zipfian distribution; a few names occur with very high frequencies, but the majority occur with much lower frequencies that quickly reach the single digits for their WPM value (as opposed to the 100-200 WPM values that the top names have in every country). A graph showing the frequencies of the top 500 names by frequency in the *CdE* can be seen in figure 3.1. The Zipfian distribution of the names makes a case for including name as a random effect.

Finally, the regression functions used in R provide an Akaike information criterion (AIC) number, which represents the estimated quality of a model (Levshina, 2015: 149). This number can be compared between different models

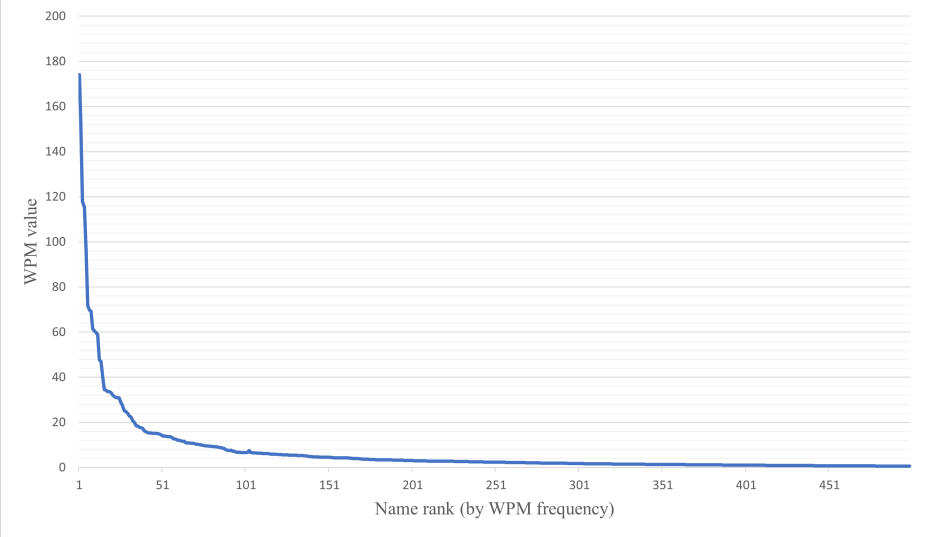

Figure 3.1: Words-Per-Million values of the top 500 names in the *CdE*

to analyze the relative quality of their fit, with lower numbers indicating a better fit.

In conducting the analysis done here, I created both fixed-effects only models for the data from each country which did not include name as a factor, as well as mixed-effects models for each country that did include name as a random effect, with all other effects unchanged (fixed effects). In comparing the models, in the case of every single country, the mixed-effects model had a lower AIC number.

For the previously listed reasons, the mixed-effects model was chosen as the model of choice for the analyses here. Using a random effect for a specific proper first name of a person may be a novel application of a random effect, but the arguments for its use line up strongly with similar types of factors where random effects are generally accepted, such as with specific words, speakers, or texts (Geeraerts et al., 2018:2-3). Additionally, the AIC number predicted a better fit of model for the mixed-effects model for the datasets of all eight countries.

Chapter 4

Results

In this chapter, the results of the corpus study are presented, with tables and figures containing many of the numbers that were found as a result of the analyses.

Section 4.1 describes the overall frequency of the occurrence of definite articles with names across the eight countries' datasets. Section 4.2 describes the results of the mixed-effects regression analysis, with subsections going over the significance (or lack thereof) of the effects for each coded factor. The sections of this chapter lay out the patterns found in the analysis, which will then be discussed in greater detail in the following chapter.

4.1 Frequency by country

In terms of raw frequencies, the data collected in this study show that across all eight country datasets, the occurrence of definite articles with names is low. Notably, in the Chile dataset, 1.49% of all tokens (extracted names) appeared with a definite article, which is more than double the rate at which definite articles appears with any other data set. The lowest frequency of article appearance is in the Paraguay data set. These numbers can be seen in 4.1.

After conducting a Pearson's Chi-squared (χ^2) test using RStudio (R Studio Team, 2021), country was a significant factor (p-value < 2.2e-16). Further χ^2 tests were conducted comparing each country to each other country to determine which countries had differences that were statistically significant, and the results of those tests can be seen in table 4.2. Notably, Chile (which had the highest proportion of article occurrence) had a statistically significant difference when compared to all of the other countries. Paraguay (which had the lowest proportion of article occurrence) was statistically significant when compared to every country except Mexico (which had the second lowest proportion).

Table 4.1: Article occurrence by country

Country	Argentina	Chile	Colombia	Cuba	Mexico	Paraguay	Peru	Spain
Article	384	217	224	100	304	29	170	583
No Article	61427	14312	43392	14336	67632	8789	28518	118103
Total	61811	14529	43616	14436	67936	8818	28688	118686
% Occurring with Article	0.62	1.49	0.51	0.69	0.45	0.33	0.59	0.49

Table 4.2: Significance of difference in article occurrence rate between countries

Country	Chile	Colombia	Cuba	Mexico	Paraguay	Peru	Spain
Argentina	***	*		***	***		***
Chile		***	***	***	***	***	***
Colombia			*		*		
Cuba				***	***		***
Mexico						***	
Paraguay						***	*
Peru							*

*** = P<.005 ** = P<.01 * = P<.05

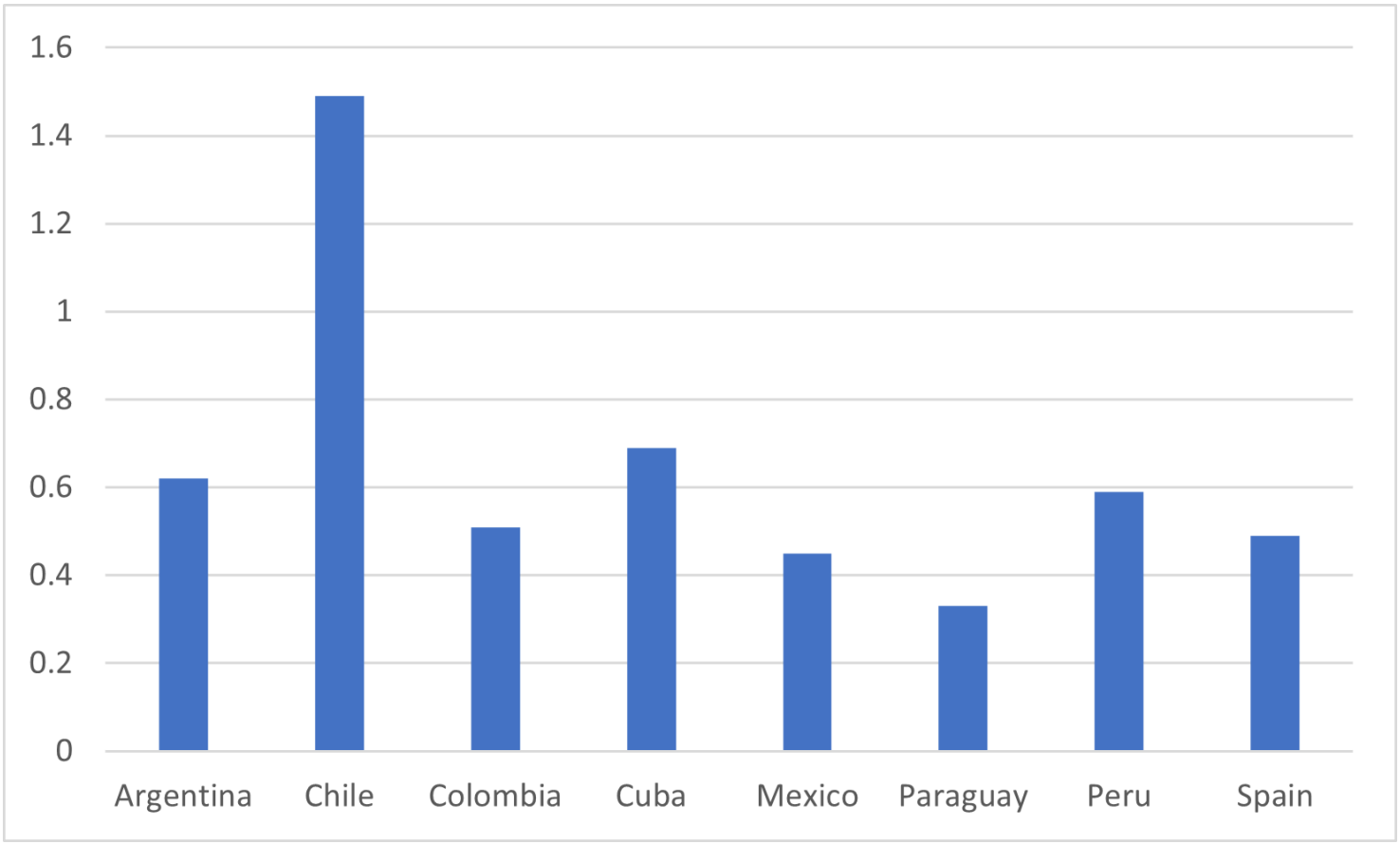

Figure 4.1: Visualization of frequencies of article use for each country (percentage of article occurence out of total tokens)

4.2 Independent variables

In this section, the results of the main analysis will be discussed. Table 4.3 shows the results of the mixed-effects generalized linear model discussed in section 3.5.1, with respect to the significance of factors in their favoring or disfavoring of definite article presence.

4.2.1 Name-specific variables

Name-specific variables are variables that relate to characteristics of the names used in the analysis, as opposed to characteristics relating to the context of a specific token. Gender, name frequency, and whether the name is consonant or vowel initial are the three coded variables that this applies to, as for each of these variables the possible values are inherent to each name, and for each case of a given name these values will be constant within a dataset.

Using the mixed-effects model for the data obtained in this study, gender had no significant effect on definite article occurrence in any of the countries' datasets; neither male nor female names correlated with a significant increase or decrease in article presence.

Similarly, as the words-per-million values associated with different names increased (based on values unique to each country), there was no significant increase or decrease of definite article occurrence in any of the countries' datasets. As can be seen in table 4.3, in 7 of the 8 countries (all except Mexico), an increase in WPM value (i.e. with names that appear more frequently) correlated with less article occurrence, but this correlation was never statistically significant.

Finally, names being vowel-initial (as opposed to consonant-initial) likewise had no significant effect on definite article occurrence in any of the countries' datasets.

Table 4.3: Effects across all eight country datasets

	Argentina		Chile		Colombia		Cuba		Mexico		Paraguay		Peru		Spain	
	Significance	Coefficient	Significance	Coefficient	Significance	Coefficient	Significance	Coefficient	Significance	Coefficient	Significance	Coefficient	Significance	Coefficient	Significance	Coefficient
Male		-0.225		-0.63		-0.531		-0.382		-0.489		0.2882		0.3		0.101
Increase in WPM value		-0.004		-0.005		-0.005		-0.005		0.001		-0.003		-0.004		-0.009
Vowel Initial		0.165		-0.103		-0.529		-0.256		-0.403		-1.272		0.473		-0.214
Preceded by "a"	***	0.753	***	1.188	**	0.675		0.488	*	0.433	*	1.694		0.282	**	0.459
Preceded by "de"	*	0.325	**	0.584	***	0.723		0.372	*	0.342	***	1.636	***	1.182	***	0.814
Followed by 3ps verb	*	-0.369		-0.318		0.343		-14.023		-0.135		-6.356		-0.149		-0.286
Followed by "que"	***	1.165		0.181	***	2.334	***	2.871	***	1.539		-23.73	***	2.28	***	1.785
Followed by "de"	***	1.03	***	1.103	***	2.114	***	1.887	***	1.061	***	1.759	**	0.676	***	1.653
Followed by adj.	***	2.026	***	1.618	***	3.003	***	3.12	***	1.977	**	3.37	***	2.169	***	2.542

Green indicates a significant positive effect, while red indicates a significant negative effect. *** = $P<.005$ ** = $P<.01$ * = $P<.05$

4.2.2 Context-specific variables

Context-specific variables are those variables that relate to syntactic features that surround any given data token. They are not related to the inherent characteristics of any given name, but rather to the context in which a token is found in the data.

There were significant positive effects on definite article occurrence in cases where a name (and article, if present) were directly preceded by *a* or *de* (contracting to *al* or *del* in cases where a following definite masculine article was present).

In cases where there was a preceding *a* (ex. *lo di a la Juana*), there was a significant positive effect on definite article occurrence in all of the countries except Cuba and Peru.

In cases where there was a preceding *de* (ex. *es un cuadro del Pablo*), there was a significant positive effect on definite article occurrence in all of the countries except Cuba.

The presence of a verb conjugated in a singular 3rd person tense (ex. *la María llegó* or *el Santiago habla con ella*) directly following a name had a significant negative effect on article occurrence only in the data from Argentina. For the 7 other countries, there was no significant positive or negative effect in cases where the name was followed by a verb conjugated in a singular 3rd person tense.

The presence of *que* directly following a name (ex. *no es el Pedro que conocía*) had a significant positive effect on article occurrence for every country except for Chile and Paraguay.

The presence of *de* directly following a name (ex. *es la Paula de la escuela*) had a significant positive effect on article occurrence in all 8 countries' datasets.

Finally, the presence of an adjective directly following a name (ex. *fue la Ana mayor que me dijo eso*) had a significant positive effect on article occurrence in all 8 countries' datasets.

Chapter 5

Discussion

In this chapter, the results of the corpus study are discussed, beginning with discussions of the hypotheses. Following, there is a brief section discussing other notable findings regarding factors that were not included as part of any hypothesis, but rather to try capture any unforeseen patterns. After that, there will be a discussion of the limitations of the methodology used for this study.

5.1 Hypothesis 1: Of the countries included in this study, Chile will have the highest frequency of article presence.

Regarding the first hypothesis, the raw frequency data confirm that Chile had a much higher frequency of article occurrence in the data, with more than double the frequency rate of any other country. When Chi-squared tests were applied, Chile was shown to have a statistically significantly higher article occurrence rate than any other countries in the study (see table 4.2 and section 4.1).

Importantly, all countries had cases of article presence, even in cases where there were none of the expected predictors (which is further discussed in section 5.3). Even in the case of Paraguay, which had only 29 article occurrences total, there were still uses that contained none of the expected predictors, as seen in the following examples taken from the Paraguay corpus data used in the analysis.

(5.1) *El despido se dio alegando que querían cambiar el formato del noticiero poniendo una figura femenina al lado* ***del Roberto****, pero sorpresivamente el que reemplazó esta siesta a Andrés fue Diego…*

'They gave the dismissal alleging that they wanted to change the format of the news by putting a feminine figure next to **the Roberto**, but surprisingly the one who replaced Andrés this break was Diego...'

(5.2) *... Le costó una feroz arremetida de **la Juana**; primero soy mujer, después abuela...*

'It cost him a ferocious onslaught from **the Juana**; I am first woman, then grandmother...'

Examples from the *Corpus del español: Web/dialects* (Davies, 2016-).

These are clear examples of colloquial article use, as opposed to an expected use; the articles here are not necessitated by any syntactic restrictions, as the names they precede are not modified, plural, or part of a lexical category that would otherwise require an article (Matushansky, 2006, Martínez García, 1999, RAE, 2005).

There were similar examples in all of the datasets, as shown in the following examples from Cuba, Peru, and Argentina.

Cuba:

(5.3) *¿No le dije que una vez **el Ernesto**, en frente de todos, me quitó una vieja en El farol?*

'I didn't tell you that one time **the Ernesto**, in front of everyone, took an old lady from me in El farol?'

Peru:

(5.4) *... Pues hubieran invertido papeles: **la Claudia** en una de esas dos.*

'... Well they had inverted roles: **the Claudia** in one of the two.'

Argentina:

(5.5) *Je, como lo hizo **el Alberto** en su momento, como lo hizo también masita.*

'Ha, just as **the Alberto** did it at the time, just like he made the dough as well.'

Examples from the *Corpus del español: Web/dialects* (Davies, 2016-).

Although Chile had the highest frequency of article use out of the 8 countries, the colloquial use of articles with anthroponyms certainly exists in all the 8 countries. This supports the claims made by some authors (see Martínez García, 1999; Laca, 1999; Rigau, 1999; Fernández Leborans, 1999) that this structure is not uncommon in colloquial Spanish in general. Although the frequencies vary among countries, the results here support the idea that the colloquial use of articles with anthroponyms at the very least exists as a feature of many dialects of Spanish, not just in the Spanish of Chile. At the same time, this study shows the need for more research investigating how this feature varies across the Spanish-speaking world, especially across different dialects, sociolects, and many other extra-linguistic categories.

It is important to note that the low frequencies across all 8 datasets could reflect the methodology used here rather than the true extent of article use in each of the countries. The data in the Corpus del Español is taken primarily from various webpages, and although many of these are blogs, other webpages are taken from news sites, webpages from professional organizations, and other such web-hosted content. It is presumed in this study that many of these webpages reflect more formal writing, which may not accurately reflect the colloquial speech where colloquial article use is most expected. It is possible that frequencies in the same countries could be found to be higher (or even lower) in spoken data, which is beyond the scope of this study but should be considered in the future.

One recent related study that did use conversational sociolinguistic data was Tieperman's (2020) study, in which she found an exceedingly high frequency of article occurrence in the Chilean data (in her case, taken from the *PRESEEA* corpus). In fact, the majority of analyzed tokens in the Santiago portion of her corpus study appeared with an overt definite article (Tieperman, 2020: 34). Compared to the results of this study (in which the Chilean dataset had an article appearance rate of 1.5%), this difference is extreme. A possible explanation for this different could be due to the type of data; sociolinguistic interviews with oral data would be expected to yield a higher rate of article occurrence. Even so, the large discrepancy (and lack of any similar patterns in Tieperman's Granada or Mexico City datasets) is strange. Two additional explanations are relating to the small number of speakers in the Santiago data (18 speakers), or perhaps relating to the fact that Tieperman's data only came from specific cities (in the case of Chile, Santiago de Chile). It is possible the patterns seen in Tieperman's study are more specific to the specific speakers or the regional variety/varieties that exist in Santiago, rather than in the whole of Chile.

5.2 Hypothesis 2: Female names will have a higher frequency of article presence than male names.

Gender of the names was not found to be a significant factor in any of the 8 countries' datasets, which refutes the second hypothesis. Although female names have been pointed to in past works as occurring with articles more often (DeMello, 1992, Fernández Leborans, 1999, RAE, 2005, Tieperman, 2020), the mixed-effects model used here did not show any significant effect of female names when compared to male names.

One of the most notable differences between the two potential models that were considered for these analyses (discussed in section 3.5.1) is that gender did have a significant effect on article occurrence in the fixed-effects model, with 4 out of the 8 countries showing a significant correlation between female names and article presence. This does reflect similar findings of prior studies.

When compared to past studies that have shown a correlation between female names and article occurrence, the lack of any significant effects when accounting for name as a random effect, as done here, could suggest that certain anthroponyms (perhaps certain female names) may be associated with more article occurrence. Future studies could benefit from a greater focus on a few specific names (both male and female) to see whether more definitive patterns of use could be discovered. If done in that manner, the Zipfian nature of name distribution (discussed previously in section 3.5.1) could mean that studying a few high-frequency names might yield more definitive results when it comes to individual names and gender.

Once again, the results of Tieperman's (2020) corpus study are quite different than the results found here. In her corpus study, her analysis showed the gender of the referent to be a significant predictor of article usage. A possible explanation for this difference could once again relate to the small amount of speakers that make up the *PRESEEA* data used by Tieperman; it is possible that for some specific speakers there is an effect that would not hold true for all speakers (although Tieperman did include speaker as a random effect). Additionally, it is possible that the large amount of names appearing in the current study across thousands of tokens is one reason for a lack of significant gender effect. It is possible that with a smaller number of names (limited to the most frequent names) there would be a significant effect of gender. However, even if that were the case, it would not be conclusive; I have already put forward that I believe the inclusion of Name as a random effect accounts for variation between individual names that might otherwise appear to be a larger trend. In other words, using a smaller amount of names would not necessarily give us

better insight into all female or male names, but rather the specific names that are analyzed in such a case. I argue that the inclusion of many names of varying frequency better reflects the overall patterns of usage when it comes to effects such as gender of the referent.

5.3 Hypothesis 3: Names followed by *que*, *de*, or by an adjective will have higher frequencies of article presence than names that are not followed by those elements.

The third hypothesis was made based on observations by previous studies on definite article occurrence with proper names (see Matushansky, 2006, Martínez García, 1999, RAE, 2005), specifically regarding cases that contain one of the following syntactic elements that traditionally predict article usage; having a following *que* that specifies a referent, a *de* that specifies the referent, or an adjective that modifies the name. In cases where those elements are present, they restrictively modify the proper name in question, thus requiring an overt article with the name (Matushansky, 2006; Martínez García, 1999; RAE, 2005).

Such cases are expected across all dialects. On the other hand, cases of article use without any of those syntactic elements would better fit the description of colloquial article use with anthroponyms, described in a plethora of past literature (Martínez García, 1999, Laca, 1999, Rigau, 1999, Fernández Leborans, 1999, Carranza Brito, 2008, Calderón Campos, 2015). This colloquial but optional use has been noted as being more prominent in certain places, such as Chile, which is confirmed by both previous work (DeMello, 1992; RAE, 2005; Tieperman, 2020) as well as the results of this study (see table 4.1).

The third hypothesis and the factors relating to it were included in this study in order to test the validity of claims regarding the normal, expected usage of articles with anthroponyms (not the colloquial usage). The results, seen in table 4.3, show definitively that the third hypothesis and the underlying claims are supported in the data collected in this study. Overwhelmingly, *que*, *de*, and adjectives that followed anthroponyms correlated significantly with an increase in article occurrence. Out of these 3 factors applied to 8 countries, all were found to have a significant positive effect except for *que* in two countries, Paraguay and Chile.

In Paraguay, the country's dataset was the smallest, and the sampling (in terms of what is included in the CDE) could thus explain why there was no significance for the factor of a following *que*. In the case of Chile, it is possible

that the lack of significant effect of the factor could be due to a similar amount of natural variability in the content of the dataset. Alternatively, it could be possible that, due to the relative widespread nature of the structure overall in the country, the pertinent underlying restrictions on article omittance could be different from those in other countries, where article use with anthroponyms is not nearly as ubiquitous. More research would be necessary to confirm whether that might be the case.

Outside of those two exceptions, however, the data and analyses in this study overwhelmingly support the notion that the linguistic factors most correlating with article occurrence are syntactic elements that modify or further specify a named referent.

5.4 Other factors

In this section I discuss the name-specific and context-specific factors that were not included in the three hypotheses. Besides gender, the two factors related to properties of the name (name frequency and whether the name in question is consonant or vowel initial) had no significant effect in any of the datasets, but these factors are briefly discussed here.

In the fixed-effects model that was eventually set aside in favor of a mixed-effects model, higher WPM value correlated negatively with article presence (i.e., low frequency names had more article occurrence than high frequency names), with significance in 6 out of 8 countries. However, the fact that these effects disappear in the mixed-effects model when name was included as a random effect suggests that specific high or low frequency names were affecting the fixed-effects analysis disproportionally.

That the frequency of names did not have any significant effect on the occurrence of definite articles suggests the possibility that the use of definite articles before anthroponyms is equally productive for both high and low frequency names[7].

[7] It should be noted that it is possible there could be significant effects when looking at only expected canonical cases or only non-expected colloquial cases.

The factor relating to whether a name was consonant or vowel initial was included to catch possible unforeseen patterns of phonotactic preference for speakers' use of articles. Although there were no previous findings or claims in the literature on Spanish regarding this factor, there is attested variation between certain Catalan varieties regarding names that begin with a consonant or a vowel (GenCat, 2020). The factor was not significant in this study for any of the datasets, but I find it plausible that future studies (especially those with oral data) could find a significant effect for this factor, as I am inclined to believe

that speakers may be more sensitive to phonotactic tendencies in oral data as opposed to written data.

For the factor of names being followed by a verb conjugated in a 3rd person singular form, there was only a significant effect for one country: a negative effect on article occurrence in the Argentina dataset. Because this was the only country with any significance for this factor, it is possible that this was an anomaly due to the sampling/data extraction methods used. However, it is also possible that in Argentine Spanish, the patterns of use could be slightly different than in other countries, and this warrants further investigation.

Perhaps the most interesting factor was that of names (and articles, if present) being preceded by the prepositions *a* and *de*, which contract to *al* and *del* in cases where they precede masculine articles. This factor yielded unexpected results: the presence of preceding "de" had a significant positive effect on article occurrence in 7 of 8 country datasets (all except Cuba). Similarly, the presence of preceding *a* had a significant positive effect in 6 of 8 countries (all except Cuba and Peru). A few examples of both types follow.

(5.6) *Ahí, además, estaba su mamá... y en los comerciales ella desde el switch le hablaba* ***a la Mónica*** *por el sono interno…*

'Here, moreover, was her mom... and in the commercials she from the switch talked to **the Mónica** through the internal sound.'

(5.7) *Como los eventos sociales se cancelaron (la mayoría, no así la boda* ***de la Carola*** *y* ***el Rodrigo****, jejeje), teníamos que conseguir coberturas en casas...*[8]

'Since the social events were cancelled (the majority, not so the wedding **of the Carola** and **the Rodrigo**, hehehe), we had to look for reservations at homes...'

Examples from the *Corpus del español: Web/dialects* (Davies, 2016-).

[8] Although Carola wasn't one of the names included in the corpus searches, this example came up due to the presence of "el Rodrigo," and the example has several interesting aspects: the use of article for both Carola and Rodrigo, and the preceding "de" for which it is included here.

It is also important to note that some of the cases where the preceding element was *a* are not cases of prepositional *a*, but rather the so-called "personal *a*," which simply marks a direct object as being a person (or personified entity), as seen in the following example.

(5.8) *Al menos cuando ves gente que no va a ver la Fórmula 1? sino a ver* ***al Alonso****.*

'At least when you see people that aren't going to watch Formula 1? Rather to watch **the Alonso**.'

Examples from the *Corpus del español: Web/dialects* (Davies, 2016-).

No prior sources on the structure made any specific mention of a preceding preposition as having any effect on the presence of definite articles with anthroponyms. As far as possible explanations for this consistent positive correlation with *a* and *de*, there are a few that I put forward.

The first specifically deals with *a*. It is possible that, due to the common occurrence of *a* and *al* as "personal *a*" and not as a preposition, there might be a separate effect at play. This possibility would need to be accounted for in future studies that examine the factor of preceding prepositions. It's likely that such an effect would be linked to the nature of sentences that would require "personal *a*". If there is a positive effect, perhaps it could be explained with relation to the types of verbs that commonly appear with anthroponyms as an accusative object of the sentence; it could be that such verb configurations lend themselves more readily to definite article presence.

The second explanation for the significant effect of prepositional *a* and *de* deals with a presupposition that, if a preposition such as *a* or *de* is used before a person's name, it's probable that whoever the preposition describes is already known by at least the speaker, and possibly the listener(s). In other words, it seems unlikely that a speaker would say *Lo doy a (la) María* in the first place if neither they nor the listener had any knowledge of who María is. Presupposing that there is prior knowledge of the referent, the higher use of articles finds a potential explanation in the research of Carranza Brito (2008) and Calderón Campos (2015), who both argue that one of the primary functions of non-required definite articles was to recover a referent in discourse and imply a level of intimacy with the referent. In other words, if we assume that preposition use occurs more often in contexts where the referent is known by the speaker, it stands to reason that there would also be higher article use, as definite articles can be used to imply intimacy/familiarity.

Whether the assumptions made regarding preceding prepositions hold true is a question that should be explored in future studies.

5.5 Methodological limitations

Using corpora for a study such as this one comes with benefits and drawbacks. On the positive side, the vast quantity of data allows for comparisons at a scale much larger than non-corpus methods could normally accommodate. in De Mello's 1992 study, the number of tokens found was in the low hundreds, while the current study has over two thousand instances of article use taken from

an overall pool of over three hundred thousand total tokens. Additionally, the methods used here could be replicated with more names or extended to other countries to easily multiply those numbers several times over.

On the other hand, using a corpus such as the *Corpus del Español* has distinct disadvantages. For instance, although the data seem to show consistent trends of article occurrence across the 8 datasets, the trends we see might not necessarily reflect the patterns we would see in speech.

Additionally, there are ambiguities that arise from these methods, some of which are inherent to Spanish and would not be easy to filter out without manually checking every pertinent case (such as ambiguities between the 1st and 3rd person singular forms of subjunctive verbs), but others which are inherent to the sometimes faulty or ambiguous tagging systems used in the *Corpus del Español* (or any other corpus, for that matter).

One minor issue related to this study in particular deals with the topic of finding and extracting data for specific names. In the case of many names, they double as common nouns (such as Rosa, or Ángel), or have other associated meanings, which can make finding relevant cases more toilsome. This was one reason why article use was analyzed with fewer than a hundred hand-picked names. Even after removing as many ambiguous or double-meaning names as possible, the list was based on the relative frequencies from within the corpus itself, and there is no guarantee that all of those names are proportional to the frequency that those names might have in day-to-day speech. However, the *CdE* does not have specific tags for anthroponyms, and to my knowledge no other corpus does, so any other methods not using a specific list would have to sort through hundreds of irrelevant proper names, some of which may be common nouns that were mistagged.

Overall, I believe that using the *Corpus del Español* in this study provided more benefits than drawbacks; due to the large amount of data and the relative ease in accessing and searching through it, it was possible to find significant amounts of relative cases for a very specific feature. This allowed for comparisons at a country level done in a quantitative manner, and although there are improvements to be made both to the data extraction methods and to the analyses conducted, this study is, to my knowledge, one of the broadest studies conducted on the topic in Spanish, and it also lays the groundwork for future work on the topic. Showing that definite articles can occur with anthroponyms (both in the expected cases and the so-called colloquial use cases) across all parts of the Spanish-speaking world signals that the topic warrants further research, especially since some of the findings of others are contradicted by the findings here (especially with regards to gender).

Chapter 6

Conclusion

Over the past several decades, corpus studies have become an integral part of linguistic investigation. The large amount of data available in corpora can be used to find patterns that might otherwise be considered statistically insignificant due to their low frequency.

One such low-frequency feature of Spanish language is the use of definite articles with anthroponyms, especially in unexpected cases where there are no modifications or specifications done to the name in question. Although such unexpected cases have been studied in the past, most descriptions we have of this feature come from brief mentions in descriptive grammars, or studies more focused on the pragmatics of the construction. Even so, we know that this structure is not extremely uncommon in the colloquial speech of many Spanish-speakers.

Notably, much of the previous literature (Calderón Campos, 2015; Calderón Campos, 2018; DeMello, 1992; RAE, 2005; Tieperman, 2020) gave very little information about how widespread the feature is in the Spanish-speaking world, beyond occasional mentions of more frequent usage in Chile and areas of contact with Catalan or Portuguese.

This study has determined the extent of definite article use with first names in 8 Spanish-speaking countries by analyzing large datasets taken from the *Corpus del Español*. Beyond examining the extent of the feature in terms of raw frequencies, different syntactic factors and name-specific factors were included, meant to serve as a comparison with the findings of past studies (DeMello, 1992; Tieperman, 2020). These factors and their significance were compared among the 8 country-specific datasets.

It was hypothesized that Chile would see a higher rate of article occurrence with anthroponyms than any of the other countries studied. This was confirmed, as Chile's data had more than double the rate of occurrence of the next

highest. These numbers were all low (>2%), but every country had a variety of cases of article occurrence, some of which fit expected criteria, and others of which did not.

A second hypothesis was that female names would see a higher rate of article occurrence with anthroponyms, as this had been attested to by other sources (RAE, 2005; Tieperman, 2020). When accounting for names as a random effect, however, there was no significant effect from either male or female names regarding article occurrence.

The third and final hypothesis was that syntactic elements that modify or specify names (such as *de* or *que* following a name and further specifying a referent) would correlate with higher rates of article occurrence. This was confirmed by the data, which show that these factors were significantly positively correlated with article occurrence.

Beyond the scope of the three hypotheses, one of the crucial intended purposes of this study was to show that this structure does exist across many (if not all) varieties of Spanish. The data here show unequivocally that, at least for the eight countries examined, there exist both canonical (expected) uses of the definite article with anthroponyms, but also unexpected cases that seem to call for further investigation into their pragmatic and discursive functions. In the future, more corpus studies should be conducted on this topic, either refining the methods used here or with other corpora that fit large-scale investigation. Additionally, socio-pragmatic and pragmatic studies (Calderón Campos, 2015; Calderón Campos, 2018; Carranza Brito, 2008; Tieperman, 2020) have already provided much information, and presumably in the future could reveal even more about the qualitative nature of the feature in Spanish.

I argue that the results of this study show potential for more research on the topic of definite article use with personal names in Spanish, especially in the area of socio-pramatics. Even though it is a rather niche topic within Spanish linguistics, gaining a more complete understanding of the nature of proper names has implications for semantics, pragmatics, and other areas of linguistic study, especially when the patterns are very similar across dialects, as seen here.

One aspect of this study that is especially noteworthy is the lack of a (statistically significant) effect of gender in the data. This contradicts claims made in the past, especially in the *Gramática descriptiva* (Demonte and Bosque, 1999). However, the fact that such findings contradict the findings of past descriptions of the structure does reinforce the claims made by Motschenbacher: corpus-based studies can improve our understanding of structures that might otherwise be relegated to short mentions in passing made in reference works (Motschen-

bacher, 2020:98). Motschenbacher concludes his article with the following related statement:

> ...The instruments that corpus linguistics offers have the potential to substantially improve our knowledge about how names are actually used. They are likely to give us a better understanding of how the use of names is shaped by the lexical, grammatical and extralinguistic context, which constructions they engage in, and which meanings they are associated with beside their basic denotation (Motschenbacher, 2020:100).

In the case of this study, the use of a corpus allowed for a broader understanding of patterns than would be feasible through most other data collection methods, due to the low frequency nature of the structure. It also has provided evidence of patterns that contradict previous claims regarding gender, and provided context and fuel for further study of those aspects of the feature. Whether those patterns are refuted or confirmed in future studies on the topic, it is my belief that this study has shown the value of corpora as a linguist's tool for research topics such as this one, affirming Motschenbacher's propounding of the effectiveness of corpora in name-related linguistic studies.

On a final note, it is important to point out that the current study was not intended to serve as a replacement for qualitative studies on the topic, but rather it was meant to provide large-scale quantitative data on the subject. Hopefully the findings here will complement new studies, both qualitative and quantitative. I believe that such studies based on and expanding upon the groundwork laid here will be able to reveal more about this niche yet fascinating feature of Spanish language.

Bibliography

Alves, A. P. M. (2008). *Um estudo sócio-linguístico da variação sintática ausência/presença de artigo definido diante de antropônimos na fala dos jovens de barra longa-mg que residem em belo horizonte* (Master's thesis). Universidade Federal de Minas Gerais.

Alves, A. P. M. (2017). O comportamento linguístico dos jovens de barra longa/mg em relação ao uso do artigo definido diante de antropónimos. *Caletroscópio*, *5*(8). https://periodicos.ufop.br:8082/pp/index.php/caletroscopio/article/view/3769

Amaral, E. T. R. (2016). A importância do fator intimidade na variação ausência/presença de artigo definido diante de antropônimos. *Veredas – Revista de Estudos Linguísticos*, *11*(1), 116–127. https://periodicos.ufjf.br/index.php/veredas/article/view/25220

Barth, D., & Kapatrinski, V. (2018). Evaluating logistic mixed-effects models of corpus-linguistic data in light of lexical diffusion. In D. Geeraerts, K. Heylen, & D. Speelman (Eds.), *Mixed-effects regression models in linguistics* (pp. 99–116). Springer International Publishing.

Bick, E. (2000). Palabras [Tagging Software]. https://visl.sdu.dk/~eckhard/pdf/PLP20-amilo.ps.pdf

Calderón Campos, M. (2015). El antropónimo precedido de artículo en la historia del español. *Hispania*, *98*(1), 79–93. http://www.jstor.org/stable/24368853

Calderón Campos, M. (2018). Intersubjectification and textual emphasis in the use of definite article + proper name in spanish. In M. Bouzouita, I. Sitaridou, & E. Pato (Eds.), *Studies in historical ibero-romance morphosyntax* (pp. 75–98). John Benjamins Publishing Company.

Carranza Brito, R. (2008). Artículo ante nombre propio: Matices de significado. *LL Journal*, *3*(2). https://lljournal.commons.gc.cuny.edu/2008-2-carranza-brito-texto/

Christodulelis, E. (2013). *Pragmatic constraints on definite article use with anthroponyms in BRAZILIAN PORTUGUESE* (Master's thesis). The Ohio State University. https://www.academia.edu/24671619/

Pragmatic_Constraints_on_Definite_Article_Use_with_Anthroponyms_in_Brazilian_Portuguese

Christodulelis, E. (2017). Definite articles with anthroponyms across varieties of Spanish: Quantitative support of conventional conversational implicature [Presentation of paper]. https://www.academia.edu/35001008/Definite_Articles_with_Anthroponyms_Across_Varieties_of_Spanish_Quantitative_Support_of_Conventional_and_Conversational_Implicature

Coromina i Pou, E. (2001). *L'article personal en català. marca d'oralitat en l'escriptura* (Doctoral dissertation). Universitat Autònoma de Barcelona. Tesis Doctorals en Xarxa.

Cruz, A. (2017). Spanish first name [Data set]. https://data.world/axtscz/spanish-first-name

Davies, M. (2016-). Corpus del español: Web/dialects [Corpus]. https://www.corpusdelespanol.org/web-dial/

DeMello, G. (1992). El artículo definido con nombre propio de persona en el español hablado culto contemporáneo. *Studia Neophilologica*, *64*(2), 221–234. https://doi.org/10.1080/00393279208588100

Demonte, V., & Bosque, I. (Eds.). (1999). *Gramática descriptiva de la lengua española*. Espasa-Calpe. https://www.rae.es/obras-academicas/obras-linguisticas/gramatica-descriptiva-de-la-lengua-espanola

Evert, S., & Hardie, A. (2021). The ims open corpus workbench, version 3.4 [Computer Software]. http://cwb.sourceforge.net/

Fernández Leborans, M. J. (1999). El nombre propio. In I. Bosque & V. Demonte (Eds.), *Gramática descriptiva de la lengua española* (pp. 77–128). Espasa.

Geeraerts, D., Heylen, K., & Speelman, D. (2018). *Mixed-effects regression models in linguistics*. Springer International Publishing. https://doi.org/10.1007/978-3-319-69830-4

Generalitat de Catalunya [GenCat]. (2020a). L'article davant de noms de persona [Reference Work].

Generalitat de Catalunya [GenCat]. (2020b). Ús de l'article davant dels noms de persona, d'animals i d'objectes singulars [Reference Work].

Harris, J. (1991). The exponence of gender in Spanish. *Linguistic Inquiry*, *22*(1), 27–62. http://www.jstor.org/stable/4178707

Laca, B. (1999). Presencia y ausencia de determinante. In I. Bosque & V. Demonte (Eds.), *Gramática descriptiva de la lengua española* (pp. 891–928). Espasa.

Levshina, N. (2015). *How to do linguistics with r: Data exploration and statistical analysis*. John Benjamins Publishing Company. https://doi.org/10.1075/z.195

Maiden, M., & Robustelli, C. (2007). *A reference grammar of Modern Italian*. Routledge. https://doi.org/10.4324/9780203783504

Marcato, G., & Thüne, E. (2002). Italian. gender and female visibility in Italian. In M. Hellinger & H. Bußmann (Eds.), *Gender across languages: The linguistic representation of women and men.* (pp. 187–217). John Benjamins Publishing Company.

Martínez García, J. A. (1999). La concordancia. In I. Bosque & V. Demonte (Eds.), *Gramática descriptiva de la lengua española* (pp. 2695–2786). Espasa.

Matushansky, O. (2006). Why rose is the rose: On the use of definite articles in proper names. In O. Bonami & P. Cabredo Hofherr (Eds.), *Empirical issues in syntax and semantics 6* (pp. 285–307). Colloque de syntaxe et sémantique à Paris. http://www.cssp.cnrs.fr/eiss6/matushansky-eiss6.pdf

Mendes, A. A. (2015). *Ausênsia e/ou presença de artigo definido diante de antropônimos na fala dos moradores das cidades de abre campo e matipó: Um estudo sociolinguistico* (Doctoral dissertation). Universidade Federal de Minas Gerais.

Motschenbacher, H. (2020). Corpus linguistic onomastics: A plea for a corpus-based investigation of names. *Names*, *68*(2), 88–103. https://doi.org/10.1080/00277738.2020.1731240

R Core Team. (2021). R: A language and environment for statistical computing [Computer Software]. https://www.R-project.org/

R Studio Team. (2021). Rstudio: Integrated development for R [Computer Software]. http://www.rstudio.com/

Real Academia Español [RAE]. (2005). *El Diccionario panhispánico de dudas*. Real Academia Española. https://www.rae.es/dpd/el

Rigau, G. (1999). La estructura del sintagma nominal. In I. Bosque & V. Demonte (Eds.), *Gramática descriptiva de la lengua española* (pp. 311–362). Espasa.

Tatham, S., Lanes, A., Harris, B., & Nevins, J. (1997-). Putty: A free ssh and telnet client [Computer Software]. https://www.chiark.greenend.org.uk/~sgtatham/putty/

Tieperman, R. E. (2020).

¿Cómo está el Pablo?: Examining definite article use with personal proper names in three regional varieties of Spanish (Master's thesis). Texas Tech University.

Appendix A

CQP code used for data extraction

This appendix contains the CQP query language code used to extract the data for each country's dataset. For the example, the names used were substituted with placeholders. This example reflects the code used to extract the data for the Spain dataset.

```
SpainData = [pos!= "o|np|nms|nmp|nfs|nfp"] [word = "Name1|Name2|
Name3|Name4|Name5"%cd pos="o"] [pos!="o|np|nms|nmp|nfs|nfp"]
:: match.source_country_code = "ES" ;
```

The query essentially is made up of three components. The first is [pos!= "o|np|nms|nmp|nfs|nfp"], which is an element that is meant to exclude any proper names that might come before the target names. Although the POS tag "o" is the primary tag used to define proper nouns, initial searches showed that many anthroponyms have been improperly annotated using the other tags ("np", "nms", etc.), and thus those tags were included as well.

The second component is the large section of code, which contains all of the names used in the final set of extractions used to create the eight datasets. This element of the search query looks for any of those names, is case-insensitive, and also checks to make sure that the name is also annotated with a proper name tag in the corpus.

The final component of the code is equivalent to the first, and insures that a complex name (such as *Ana María*) does not show up in cases where the initial name might be part of the list, but followed by a less common name.

Additionally, after these three components of the query, the section with "match.source_country_code" is used to narrow the search to only the tokens coming from a specific country's corpus (Spain, in the case of the example). This was changed for each of the countries studied.

Appendix B

List of excluded names

This appendix includes several lists of names that were excluded the corpus study. Many of these names were frequent enough to show up in the top (approximately) two hundred names when the entire name list was run through the corpus. However, many of these names were likely over-represented, due to having multiple meanings and being mis-annotated with the corpus POS tags.

The names have been divided into groups based on the reason for their exclusion, although several names fall into multiple categories. For instance, "Cruz" is excluded because it is primarily a last name (the focus of this study was only on first names), and it also can be used as a common noun, meaning many hundreds or possibly thousands of tokens could be flagged as false positive or false negatives due to the word being used in a context where "Cruz" has nothing to do with anyone's name (or where it is a last name).

Place names:

- Asunción
- Carolina
- Europa
- Galicia
- León
- Siria
- Valencia
- Washington (also borrowed, also a last name)

Religious names:

- Concepción
- Cruz (also a last name)
- Domingo
- Espíritu
- Jesus
- Navidad
- Trinidad
- Salvador (also a common noun)
- Santa
- Santo
- Santos

Common noun/adjective/has other meanings:

- Justo
- Las
- Luna
- Luz
- Amor
- Ángel (also religious)
- Ángeles (could also be marked for location due to Los Angeles, or religious)
- Castillo
- Corazón

- Fidel
- Flores
- Rico
- Río
- Ríos
- Mercedes
- Niño
- Oscar
- Paz (also could be a false positive with La Paz)
- Salud
- Vírgen (also religious)
- Villa

Borrowed names:

- Edward
- George
- Jackson
- Paul
- Richard
- Thomas
- Walter

Last names:

- Aguilar

- Belén
- Castro
- García
- Peña
- Platón
- Ramos
- Reyes
- Silva
- Soto

Appendix C

Final name list by frequency

	CDE (Total)	Argentina	Chile	Colombia	Cuba	Mexico	Paraguay	Peru	Spain
Abraham	10.90	8.03	8.9	16.07	5.14	14.99	5.48	8.56	5.46
Adolfo	7.55	8.31	8.18	5.82	6.95	6.58	9.24	6.39	6.04
Agustín	13.88	20.94	10.46	12.74	26.3	15.72	25.19	15.17	10.84
Alberto	33.27	41.39	30.34	39.79	58.39	30.49	34.99	56.95	24.69
Alejandro	30.94	43.46	29.82	39.54	32.53	34.06	24.45	42.52	17.9
Alfonso	14.80	6.58	9.34	26.99	29.2	12.86	11.19	16.8	17.09
Alfredo	15.12	18.92	14.71	14.44	22.62	13.53	33.63	18.23	11.26
Alonso	15.02	11.15	7.74	9.75	32.72	10.87	22.53	11.52	24.05
Álvaro	15.35	5.19	20.72	40.68	5.98	7.25	8.26	10.2	11.93
Ana	33.69	28.46	17.13	25.61	29.58	26.86	26.31	24.62	38.28
Andrea	8.45	11.22	16.84	10.79	7.05	6.07	8.91	7.58	7.5
Andrés	31.04	22.49	64.48	76.45	24.54	33.52	30.69	26.36	15.43
Antonio	58.93	38.35	47.36	53.65	114.76	51.27	81.01	53.19	67.41
Beatriz	6.66	8.56	6.17	6.76	8.45	4.79	6.51	4.92	6.64
Benedicto	10.14	8.63	6.87	7.21	13.07	8.04	12.46	14.8	7.86
Camilo	6.76	4.4	12.4	24.75	17.38	2.5	6.51	3.73	1.86
Carlos	115.38	142.4	94.92	146.28	147.32	93.88	171.41	138.7	71.86
Clara	6.64	3.75	1.58	10.56	64.05	2.21	6.1	4.27	3.59
Claudia	11.60	14.06	21.43	15.28	5.3	7.9	6.04	11.6	5.65
Claudio	7.67	20.54	33.53	3.62	3.94	4.64	7.4	9.25	3.74
Cristina	25.34	139.82	10.93	9.24	12.04	7.22	24.75	9.33	15.71
Cristóbal	9.32	5.88	11.1	6.13	14.51	8.24	8.26	6.71	6.12
Daniel	40.50	77.94	34.9	43.39	23.77	32.93	28.45	38.3	23.24
Darío	8.50	12.99	5.05	13.92	8.26	3.5	14.86	3.95	3.26
Diego	31.40	55.69	34.68	35.02	24.6	28.73	32.86	29.47	22.21
Eduardo	28.78	45.39	42.35	30.2	33.79	23.41	28.83	32.62	15.07
Elena	10.86	9.46	7.4	7.27	12.75	8.9	7.04	10.14	11.63
Elías	9.14	5.84	5.44	11.76	7.37	11.01	8.44	10.04	3.55
Emilio	10.83	10.43	11.39	9.15	19.22	14.42	11.43	7.31	8.37
Enrique	32.26	25.41	19.96	34.97	46.23	70.19	31.61	29.22	20.01
Ernesto	12.99	13.66	12.26	13.81	53.72	14.52	9.83	13.95	4.07
Esteban	9.53	13.51	11.38	8.35	16.95	6.59	10.09	7.39	7.36
Federico	10.59	24.58	7.87	7.74	9.35	6.35	54.58	9.29	6.33
Felipe	23.98	14.11	61.11	33.11	18.92	31.69	17.32	19.33	20.32
Félix	10.31	9.27	4.92	8.09	29.16	9.26	23.71	8.01	7.43
Fernando	47.07	50.24	46.2	59.58	58.45	34.24	117.36	46.54	43.41
Francisco	69.26	74.36	63.17	62.21	62.19	60.37	93.92	61.32	55.21
Franco	14.04	15.92	11.78	11.68	9.96	8.39	102	12.62	16.45
Gabriel	18.49	25.15	19.52	28.26	20.12	19.59	15.81	18.05	8.7
Gerardo	9.69	14.21	4.86	6.82	46.7	12.89	16.25	7.38	3.19
Gonzalo	10.78	11.68	19.72	11.48	7.96	5.69	7.73	23.7	7.94
Guillermo	22.36	42.97	24.27	27.72	30.23	20.33	19	20.19	12.88
Ignacio	15.11	19.94	22.9	11.47	19.44	13.45	29.84	10.51	16.29

Isabel	15.52	10.2	23.48	11.95	23.93	11.83	10	11.32	19.15
Javier	33.68	32.17	21.31	26.87	20.85	36.24	30.72	54.1	43.39
Joaquín	12.1	8.44	15	9.7	12.1	11.72	8.67	7.72	12.46
Jorge	59.77	110	68.51	80.4	80.2	49.7	66.04	77.62	25.7
José	153.01	112.15	107.19	139.72	288.43	129.27	187.78	162.64	124.5
Juan	174.08	195.36	140.87	229.02	155.51	148.72	241.71	178.91	113.19
Juana	7.25	7.9	5.81	5.52	8.44	6.69	8.82	7.54	6.02
Julián	9.97	13.11	6.11	15.58	13.82	6.44	7.37	12.4	7.96
Julio	20.77	33.83	15.33	22.04	30.51	12.07	29.87	31.65	10.49
Lara	6.69	4.14	3	6.17	6.02	7.71	7.52	4.08	7.32
Laura	15.15	18.67	9.09	18.19	16.95	16.62	11.9	12.02	14.29
Lázaro	7.63	8.82	3	3.66	33.86	18.01	4.47	4.97	3.62
Leonardo	9.45	16.31	7.25	10.9	13.29	7.25	6.6	6.32	6.21
Lorenzo	11.75	29.8	4.51	5.72	10.39	6.08	48.48	6.27	9.91
Lucas	17.68	24.64	9.47	19.59	9.82	21.16	19.24	10	11.87
Lucía	10.26	11.97	6.33	10.72	10.3	6.62	8.26	6.89	8.84
Luis	96.45	94.08	68.82	117.59	136.32	95.98	119.05	126.55	69.4
Manuel	61.45	47.95	49.95	87.37	88.93	61.63	66.81	60.37	51.67
Marcelo	12.02	42.7	23.75	5.88	3.82	7.18	28.45	6.71	4.84
Marco	10.72	5.22	15.54	12.75	7.28	12.04	11.51	20.41	6.26
Marcos	18.40	19.83	12.97	14.15	10.01	17.46	17.88	36.12	9.95
María	118.06	121.68	84.82	130.08	128.1	116.08	154.69	169.25	95.49
Mariano	12.47	25.21	5.67	5.97	8.85	6.38	16.63	8.72	19.38
Mario	34.49	37.42	35.24	34.35	35.12	36.39	49.55	42.05	28.23
Marta	11.58	11.19	7.59	26.12	18.93	5.82	12.73	4.89	13.28
Martín	47.81	127.25	29.98	32.42	40.55	31.28	45.38	82.49	37.59
Mateo	15.31	7.45	6.7	19.46	4.84	20.88	14.47	11.04	7.63
Mauricio	13.69	20.54	21.6	24.78	4.53	7.99	9.56	8.91	2.8
Miguel	60.62	60.63	41.85	44.36	72.34	71.29	63.64	73.42	51.41
Miranda	8.82	6.68	9.58	11.08	11.17	7.75	14.06	8.83	4.36
Mónica	9.59	10.57	13.42	8.76	2.75	7.77	6.16	9.76	12.38
Néstor	8.89	50.86	3.4	7.04	8.26	3.33	11.93	6.41	1.75
Nicolás	17.41	33.62	24.93	19.53	35.65	13.45	27.76	18.05	9.39
Pablo	71.69	88.7	102.31	87.71	64.01	74.93	68.88	55.13	43.82
Patricia	9.33	14.55	15.23	10	6.02	7.51	8.2	13.09	6.21
Paula	9.36	12.62	15	6.65	6.4	6.65	3.05	2.86	12.87
Paulo	6.76	6.67	9.47	7.9	6.21	5.24	15.16	11.2	3.61
Pedro	69.87	53.02	65.34	57.48	76.96	68.15	133.73	67.82	49.98
Rafael	34.33	17.61	17.45	43.28	54.92	25.62	33.12	26.49	18.41
Ramón	22.85	27.99	14.13	10.95	60.68	14.71	32.17	18.9	18.67
Ricardo	30.93	56.89	38.48	29.4	34.24	28.33	35.05	38.24	13.64
Roberto	27.43	37.47	29.36	23.7	43.02	23.08	28.12	26.81	11.15
Rodrigo	13.51	13.05	45.44	16.79	6.1	8.49	12.55	9.38	11.26
Romero	14.56	15.72	5.94	13.57	12.13	13.93	29.19	17.29	7.22

Rosa	13.76	11.27	8.41	11.55	21.89	10.17	18.06	26.71	10.31
Samuel	9.16	5.3	6.39	12.88	7.65	10.69	6.07	6.99	5.68
Sara	7.78	5.11	6.17	8.56	12.89	7.52	5.65	5.34	9.77
Sebastián	17.83	25.19	59.14	17.93	12.04	9.92	16.07	15.97	15.54
Simón	14.01	11.72	7.93	24.35	14.15	11.72	10.86	26.31	8.47
Susana	9.84	22.62	3.93	8.14	6.01	5.4	6.01	29.67	9.4
Teresa	12.73	12.01	10.57	9.66	13.72	11.94	20.72	12.26	13.48
Tito	7.28	5.55	4.48	9.23	5.57	5.05	6.99	9.68	7.22
Tomás	15.94	17.61	21.93	18.01	21.22	13.98	16.22	13.2	12.47
Vicente	16.18	13.96	13.39	13.32	27.86	18.01	16.84	11.36	15.33
Víctor	24.68	27.47	29.09	21.63	35.62	21.96	42.06	37.58	17.29

Table C.1: All names with their frequencies by country (frequencies in words-per-million)