The Dissertation Committee for Jacqueline Rae Larsen Serigos certifies that this is the approved version of the following dissertation:

Applying Corpus and Computational Methods to Loanword Research: New Approaches to Anglicisms in Spanish

Committee:

Almeida Jacqueline Toribio, Supervisor

Barbara E. Bullock, Co-Supervisor

Dale Koike

Katrin Erk

Stefan Gries

Applying Corpus and Computational Methods to Loanword Research: New Approaches to Anglicisms in Spanish

by

Jacqueline Rae Larsen Serigos

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2017

Acknowledgements

I would like to acknowledge all of the people who have helped me get to this point. First and foremost I would like to express my immense gratitude to dissertation co-supervisors and friends, Almeida Jacqueline Toribio and Barbara Bullock, for their guidance and support on both an academic and personal level. Working with you both has been truly inspiring. Your research, dedication and never-ending exploration of the field and into new areas have provided me a model to aspire to. It has been a true privilege to work with you.

I also owe a big thanks to Katrin Erk, whose teaching opened up a new field of research to me, as well as Stefan Gries, whose insightful comments and suggestions greatly strengthen the quality of the analysis here. I would also like to thank Dale Koike for her insights and perspective throughout this process. Funding for this project came from the Office of Graduate Studies. Other closely related research, which led me to this dissertation topic, was made possible by the Argentine Travel grant from Teresa Lozano Long Institute of Latin American Studies (LLILAS) at The University of Texas at Austin.

Lastly I would like to mention and thank my family, without whom, none of this would have been possible. To my husband, Pedro: las palabras me quedan cortas, in addition to your invaluable statistical consulting and brainstorming sessions, you have been my rock. Your believing in me, even when I did not believe in myself, gave me the strength to continue. To my parents, Christian and Clotilde, you have been there every step of the way, providing me help in every way possible. To Nicole, my favorite sister, even though we are hundred of miles apart, it has always felt like you are right by my side. To my daughter, Delfina Rae, you bring me so much love and joy everyday, give me much needed balance in my life, and most importantly you can always bring a smile to my face.

Applying Corpus and Computational Methods to Loanword Research: New Approaches to Anglicisms in Spanish

Jacqueline Rae Larsen Serigos, Ph.D.
The University of Texas at Austin, 2017

Supervisors: Almeida Jacqueline Toribio and Barbara E. Bullock

Understanding both the linguistic and social roles of loanwords is becoming more relevant as globalization has brought loanwords into new settings, often previously viewed as monolingual. Their occurrence has the potential to impact speech communities, in that they have the capacity to alter the semantic relationships and social values ascribed to individual elements within the existing lexicon. In order to identify broad patterns, we must turn towards large and varied sources of data, specifically corpora. This dissertation aims to tackle some of the practical issues involved in the use of corpora, while addressing two conceptual issues in the field of loanword research – the social distribution and semantic nature of loanwords. In this dissertation, I propose two methods, adapted from advances in computational linguistics, which will contribute to two different stages of loanword research: processing corpora to find tokens of interest and semantically analyzing tokens of interest. These methods will be employed in two case studies. The first seeks to explore the social stratification of loanwords in Argentine Spanish. The second measures the semantic specificity of loanwords relative to their native equivalents.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Understanding both the linguistic and social roles of loanwords is becoming more relevant as globalization has brought loanwords into new settings that were previously considered monolingual. Loanwords' occurrence has the potential to impact speech communities in that these words can alter the semantic relationships and social values ascribed to individual elements within the existing lexicon. Loanwords are found in the vast majority of, if not in all, languages across the globe, even in those that show no other signs of language contact, such as code-switching or language attrition, and thus expand the contexts in which we may study the outcomes of language contact. Their growing use in 'monolingual' varieties has also made them relevant computationally, as their appearance challenges automated natural language tools, such as part-of-speech taggers and speech recognition software, which are often designed to handle only one linguistic system at a time.

Within the field of loanword research, linguists most often rely on naturally occurring data, such as sociolinguistic interviews or spontaneous speech, as a means of answering some long-standing questions. For example, how can we distinguish between a loanword and a code-switch? How and to what extent are loanwords integrated into the recipient language? When and why do speakers choose to use certain loanwords? These questions are intended to identify major patterns across language communities, but in practice they have been addressed mostly using highly localized data (e.g. sociolinguistic interviews) and/or sporadic samples; extrapolating the results from these studies to the larger population risks exaggerated or inaccurate conclusions (Hovy, Johannsen, & Søgaard, 2015). In order to identify broader patterns, we must turn towards larger and more varied sources of data.

The trend towards ‘big data’ is appreciable across disciplines, ranging from the STEM fields, such as computer science and engineering, to the liberal arts, like the humanities and social sciences. For the latter, social media sites like Twitter, Flickr, and Facebook have provided particularly rich, user-generated data, often with various levels of accompanying metadata (e.g. information on social network, geo-location, gender, age) that can be effectively utilized to contribute valuable insights in diverse disciplines (see Bamman, Eisenstein & Schnoebelen, 2014; Cho, Myers & Leskovec 2011; De Longueville, Smith, & Luraschi 2009; Eisenstein, 2017). Though big data has much to offer academic research, it also raises basic questions: What type of information is captured in big data? What does the data represent? Is it authentic? How accessible is it? What methods and/or tools are needed to process it? (Manovich, 2011). With respect to loanword research in particular, we can distinguish conceptual issues –e.g., What types of language are represented in a corpus, in terms of contact features and content? How can we compare across corpora, given the inherent variability of language? – from practical ones – e.g., What types of corpora can we access or create that would be well-suited for loanword research? What tools can we create to identify loanwords automatically? How do automated methods compare to manual ones?

This dissertation aims to tackle some of the practical issues involved in the use of corpora while addressing two conceptual issues in field of loanword research: the social distribution and semantic nature of loanwords. The remainder of this chapter discusses the role of loanwords and anglicisms within the context of the globalization of English, defines the term loanword using Haugen’s (1950) framework, examines the use of corpora in loanword research and, lastly, presents guiding questions for this dissertation, along with the dissertation outline.

THE IMPACT OF LOANWORDS ACROSS THE GLOBE

Lexical borrowing is so ubiquitous that loanwords are present in the vast majority of languages, if not in all languages (Winford, 2003). For example, the English lexicon contains approximately 75% borrowed material from numerous other languages, such as *table, forest,* and *soup* from French, *algebra, apricot,* and *tariff* from Arabic, and *breeze, embargo,* and *oregano* from Spanish. These longstanding borrowings often become so integrated into the source language that they are no longer recognized as foreign elements by native speakers.

The degree of borrowing that arises between language pairs is determined in part by type of contact (see Thomason and Kaufman's (1988) borrowing scale). Intense direct contact can result in heavy structural borrowing, along side lexical borrowing. On the other hand, indirect contact, also referred to as distant (Loveday, 1996), causal (Winford, 2003) and weak contact (Zenner, Speelman & Geeraerts, 2014), is mostly limited to lexical borrowings. Direct contact often implies a certain level of bilingualism, which may be equal or unequal, within the community. Unequal bilingualism commonly arises in situations of urban segregation, established ethnic enclaves and geographic isolation (Winford, 2003). The asymmetry in power and prestige between the languages in contact often leads to large numbers of borrowings from the language of prestige into the minority language and sometimes ends in a complete shift to the dominant language. In contrast, situations of more or less equal bilingualism tend to be somewhat more limited in their borrowings and more bidirectional, such as in the case of the borrowing between Flemish and French found in Brussels (Treffers-Daller, 1999). Indirect contact, which often appears without significant levels of bilingualism, may arise through travel, foreign language instruction or the media. An example of indirect contact, European exploration in the Americas brought borrowings from the Algonquian languages, like *skunk moccasin, teepee,* and *wigwam,* into American English (Winford, 2003).

More recently, the globalization of English has created new sites of both direct and indirect contact. Since the 18th century, English has grown in importance due to colonial expansion and political and social prestige. Japan provides an early example of the growing presence of English; during the late 18th and early 19th century a great influx of anglicisms entered the Japanese language and now account for 7.29 % of the lexicon, though many of these borrowings may be unrecognizable as English due to intense phonological and morphological nativization, e.g. *wa-pro < word processor* (Loveday, 1996). English spread even more rapidly after the end of World War II due to the growing political, cultural and economic power of the United States. English is now recognized as an official language in over 50 countries and is spoken by around 2 billion people[1] (Crystal, 2012). Its overwhelming presence in the spheres of business and technology is illuminated by the fact that 85% of supranational organizations use English as their official language and that English makes up 80% of the language content on the Internet (Crystal, 1999; Johnson, 2009).

The degrees of English diffusion across the globe have been classified in Kachru's (1985) seminal work where he proposes the three concentric circles model. This model offers an alternative to the native/nonnative speaker dichotomy, which is insufficient to describe the variety of functions and degrees of penetration English has come to display across cultures and languages. The model comprises the Inner Circle, the Outer Circle and the Expanding Circle. The Inner Circle refers to the traditional bases of English, countries where it serves as a primary language, such as the USA, the UK,

[1] This number represents 400 million native speakers and 1.6 million L2 learners.

Canada, Australia and New Zealand. The Outer Circle refers to countries where English arrived by means of earlier colonial expansion and has remained institutionalized within the country, mostly former colonies of the UK or the USA, such as Malaysia, Singapore, India, Ghana, and Kenya. In these contexts, English is normally one of multiple languages speakers use. The Expanding Circle includes the rest of the world, where English is normally restricted to specific contexts, such as foreign language education, tourism or business. The Expanding Circle is where we would expect to find instances of indirect contact, as opposed to the Outer Circle, where higher levels of bilingualism lead to direct or intense contact.

The indirect contact found in the Expanding Circle often results in various forms of lexical borrowings, including phrasal units, such as collocations, idioms and proverbs, and single word borrowings (Furiassi, Pulcini & González, 2012). The expansion of anglicisms into European languages has been particularly well documented as their presence often well outweighs the presence of other foreign language borrowings (see Chesley, 2010; Graedler, 2004). Within some languages, anglicism rates as high as 6% and 7% have been documented (Alex, 2008; Loveday, 1996). Higher rates of anglicism usage have been found to correlate with higher levels of English proficiency (MacKenzie, 2012). The undeniable presence of anglicisms has led to a variety of responses, from efforts to suppress them via language policies to a more general acceptance reflected in the inclusion of some anglicisms in dictionaries and in the creation of anglicism dictionaries.

WHAT IS A LOANWORD?

Frameworks for classifying contact-induced lexical changes appear as early as the 19th century. The fundamental framework upon which most current classification schemes are built is presented in Haugen (1950). He bases his classification system on two processes: importation and substitution. Importation refers to the incorporation of a foreign form, which may or may not include the meaning, into a recipient language. Substitution describes the nativization of a foreign form by replacing foreign phonemes or morphemes with those from the recipient language. Both processes can take place within the same borrowing. For example, the anglicism *ciberespacio* ‘cyberspace’, which appears in Spanish, imports the English form *cyber* and substitutes the English morpheme *space* with the Spanish *espacio*. Within the broad category of borrowing, which he defines as “the attempted reproduction in one language of patterns previously found in another”, Haugen delineates two subcategories: loanwords and loanshifts (212). Loanwords are borrowings that import the form, and possibly the meaning, from a recipient language. Loanshifts, in contrast, import only the meaning and substitute the form.

Loanwords may be composed completely of foreign morphemes, such as the anglicism *software* in Spanish, or they may include foreign morphemes combined with native morphemes, as seen in the example *ciberespacio*; these combinations are referred to as loanblends. While loanwords are frequently imported with their meaning, in some cases the semantic of loanwords are limited or altered in the recipient language. For example, the anglicism *court* in Spanish only refers to the space where sports, such as tennis or basketball, are played, whereas its usage related to royalty and law is not extended from English. Some loanwords take on meanings that do not exist in the source language. For example, the anglicism *crack* in Spanish refers to someone who is exceptionally talented.

Loanshifts fall into two categories: extensions (semantic loans) and loan translations (calques). An extension is a native word that has adopted or extended its meaning due to the influence of a foreign word. The verb *agarrar* 'to grasp', as used in the Spanish spoken in Texas, has been semantically extended by the English light verb *to get* when used in expressions such as *agarrar una beca* 'to get a scholarship' (Bullock, Serigos, & Toribio, n.d.). Loan translations result from the combination of native forms based on a foreign pattern, such the term *brainwashing* from the Mandarin *xǐ nǎo*, the phrase *blue blood* from the Spanish *sangre azul* or the phrase *último nombre*[2] found in New York Spanish from the English *last name*.

Many studies have made adjustments to the definition of loanwords offered in the general framework presented by Haugen. Some studies take a more inclusive approach, considering both pure loanwords and loanblends as loanwords (Poplack, Sankoff & Miller, 1988). Other studies limit the definition of loanwords by adding criteria not mentioned by Haugen, such as the need for a word to be used frequently and known to the speech community (Mackey, 1970; Poplack & Sankoff, 1984) or the need for the word to be recognizable as foreign to native speakers (Zenner, Speelman & Geeraerts, 2012). Additionally, Thomason and Kaufman (1988) stipulate that, for a word to be considered a borrowing, it must be adopted by native or quasi-native speakers of the recipient language. Within their framework, those borrowings adopted by non-native speakers only are considered interference.

CORPORA FOR LOANWORD RESEARCH

Much research analyzing contact features, such loanwords, calques, and code-switches, suffers from sparsity of data. In the past, studies have made use of selected examples to draw conclusions about contact phenomena in general. This is problematic in that these examples may not be representative of the phenomena they seek to explain.

[2] In Standard Spanish *last name* would be translated as *apellido*.

More recently, however, there has been a push towards ‘big’ data in contact linguistics, as has occurred in numerous other disciplines within the STEM fields, social sciences, and even humanities. For linguistics in particular, this push has sparked a return to the use of corpora accompanied by increasingly sophisticated statistical analyses. In the 1950’s, Chomsky’s harsh criticism of corpora as a source of linguistic data was highly influential, leaving corpus work considerably marginalized in favor of introspection (McEnery & Wilson, 2003). The 1980s witnessed a boom in corpus studies in part thanks to the linking of corpora and computers. Corpora have become valued for their real-world data not accessible via other traditional linguistic methodologies, such as intuition or experimentation, and, depending on their size, offer access to a high number of tokens. As a benefit of the seemingly limitless number and ease of accessibility of texts on the Internet, corpora have become readily available.

Despite the massive stream of linguistic data, creating large corpora that contain contact features, like code-switching and calques, has been cited as a major challenge for the field, due to the fact that these features often do not appear in the sources used for monolingual corpora, such as newspapers, Wikipedia pages, books, and movie subtitles, etc. Code-switching and calques tend to occur in informal settings and are often limited to multilingual communities. There have been efforts to create these contact and bilingual corpora, such as Spanish in Texas Corpus (Bullock & Toribio, 2013), New Mexico Spanish-English Bilingual Corpus (Torres Cacoullos & Travis, 2015a), and the Siarad, Patagonia, and Miami corpora from BangorTalk (Deuchar, Davies, Herring, Couto & Carter, 2014). Due to the time-consuming nature of the data collection and manual transcription, these mixed corpora are often small in size, with most well under one million words.

Unlike other forms of language contact, anglicisms generally do not suffer from the scarcity problem, as they are becoming ubiquitous in many well-documented languages and occur in both formal and informal settings. As a result, studies on anglicisms have begun utilizing corpora with word counts in the millions (Andersen, 2014; Balteiro, 2011; Barrs, 2014; Onysko & Winter-Froemel, 2011; Varga, Orešković Dvorski & Bjelobaba, 2012; Zenner et al., 2012). The generous quantities of data available for anglicism research pose methodological challenges for data processing. The field of contact linguistics has exhibited a bias towards manual annotation, in part due to a strong desire to create complete inventories, listing every contact feature used in a corpus (Zenner et al., 2012). However, an exclusive dependence on manual methods significantly hinders, if not outright excludes, the possibility of using million- to billion-word corpora. Some studies have overcome these challenges by moving away from exhaustive lists and instead choose to analyze selected sets of anglicisms, while other studies have begun to utilize automated methods.

Automated methods from the neighboring field of computational linguistics hold the potential to greatly expand the possibilities of anglicism research and contact linguistics in general. While much of computational linguistics has focused on monolingual texts, a body of research into multilingual tools is emerging as the universality of language mixing, even in varieties considered monolingual, becomes acknowledged. In this sense, both contact linguistics and computational linguistics stand to greatly benefit from increased communication between the two fields.

THIS DISSERTATION

Considering the growing importance of big data across multiple disciplines and the challenges that multilingual data presents, this dissertation aims to look towards computational linguistics from the vantage point of contact linguistics with the hope of expanding the possibilities for loanword research and eventually for other contact phenomena. The guiding question for this dissertation is: What is the role of corpora in loanword research and, more precisely, what methods can enhance the way we process and analyze corpora to make loanword research more efficient and accountable? This question is accompanied by the two research questions, each explored in two separate case studies: What is the social distribution of anglicisms in Argentine Spanish? What is the semantic role of loanwords relative to the existing linguistic system in Argentine Spanish?

In this dissertation, I propose two methods, adapted from advances in computational linguistics that will contribute to two stages of loanword research: processing corpora and analyzing tokens of interest. In the second chapter of this dissertation, I present a method to automatically identify loanword tokens within a corpus, which is utilized in a case study that explores the social stratification of loanwords in Argentine Spanish. In the third chapter, I present a procedure to measure the semantic specificity of a word. This method is utilized in a case study that investigates the semantic dimension of loanwords relative to their native equivalents. In the final chapter, I present a summary of the work, discuss the theoretical implications, acknowledge the limitations and, finally, propose ideas for future work.

Chapter 2: Social Stratification of Anglicisms

INTRODUCTION

As a highly conspicuous feature of language contact, loanwords are often imbued with social meaning in addition to their semantic significance. For example, in many Spanish-speaking communities in the US, English loanwords are often devalued and stigmatized as Spanglish (Otheguy & Stern, 2011). In contrast, in the Netherlands, Dutch speakers may employ English loanwords to favorably portray themselves as youthful and modern (Zenner et al., 2014). In France, English has a growing presence in advertisements, as it becomes associated with positive attributes such reliability, business efficiency, and sophistication (Ruellot, 2011). Like much of Europe, Argentina too has seen a large influx of anglicisms, covering a range of semantic domains, as seen in the examples below.

(1) Nos casamos el año que viene ... todo al revés. De noche, en el centro, con sillones y <finger food> y me encantaría que haya una banda tocando en vivo porque el <disc jockey> no va más.

‘We will get married next year, with everything the complete opposite. At night, in downtown, with comfy chairs and finger food, and I would love to have a band play live because disc jockeys are over now.’

(From the film Mi primera boda ‘My First Wedding’ 2011)

(2) Luego del <shock>, las víctimas hicieron la denuncia ante la policía.

‘After the shock, the victims filed a report with the police’

(From the newspaper Clarín 2013)

(3) Aunque te vistas casual nunca te olvides de acompañar tus <outfits> con buenos accesorios.

‘Even if you dress casually, never forget to accompany your outfits with good accessories.’

(From the newspaper La Nación 2013)

(4) Se trató de tres <rounds> de dos minutos en un <ring> profesional

‘It was three rounds of two minutes in a professional [boxing] ring’

(From the newspaper Crónica 2013)

(5) La radiación cósmica de fondo es la luz que quedó del <Big Bang>

‘The cosmic background radiation is the light that remains from the Big Bang’

(From the newspaper La Nación 2010)

(6) A los pocos minutos, fue <trending topic> (el tema del momento) en Twitter.

‘In just a few minutes, it was a trending topic (topic of the moment) on Twitter”

(From the newspaper Clarín 2013)

However, Argentina, along with South America in general, has received less attention, resulting in a paucity of empirical research on the usage of anglicisms in this region. Argentina presents an interesting site for the study of loanwords in that it has incorporated English lexical items without having much direct contact with the language itself; it does not share a border or belong to a political union with an English-speaking country, and there is not a significant number of English-speaking immigrants. In spite of this lack of direct language contact, Argentina has been cited as one of the countries with the heaviest use of anglicisms in the Spanish-speaking world (Bordelois, 2011) and this development has not gone unnoticed. In the public domain, anglicisms have generated both positive and negative responses, ranging from dismay over the unnecessary “contamination” of Spanish to praise of these foreignisms’ utility to more precisely express ideas or concepts.

This wealth of opinion in the absence of empirical studies leaves questions as to actual usage of anglicisms. How prevalent are anglicisms in Argentine Spanish? Are they equally prevalent across all settings, registers, and topics or are they concentrated in particular pockets? This chapter aims to answer these questions, using two corpora, one composed of newspaper articles and the other of film dialogue, to analyze the distribution of anglicisms across written and oral discourse and across social groups in Argentina. Additionally, it presents and evaluates a method for automatically detecting anglicisms within a Spanish text. This method builds on previously existing models, adding additional layers of annotation to account for Named Entities and lemmas as well as identifying loanword phrases. The remainder of the chapter proceeds as follows. First, I will present a brief overview of sociolinguistic work on loanwords, a linguistic portrait of Argentina, and previous loanword detection models. Secondly, I will describe the data and methods used in this study, with particular attention paid to the automated method used to identify anglicisms. Subsequently, I will present the two sets of results: (1) the performance of the anglicism detection model and (2) the distribution of anglicisms throughout the corpora. Finally, I will discuss the implications of these findings and issues to be tackled in future work.

LITERATURE REVIEW

Sociolinguistics and Loanwords

Loanwords hold interest for a wide variety of disciplines within linguistics and, as such, are studied from numerous perspectives. Within semantics, lexicology, and lexicography, the arrival of loanwords into recipient languages brings new concepts and shifting of meanings. The levels of phonology, morphology, and syntax are all subject to adaptation as loanwords become integrated. From a sociolinguistic perspective, loanwords, which result from interaction between communities, may hold social value in addition to their denotational significance. This section discusses existing sociolinguistic literature on loanwords, which has explored attitudes towards them and the social factors that affect their usage.

Several studies have examined language attitudes towards loanwords, using a variety of techniques, including media discourse analysis, surveys and matched guises to capture both conscious and unconscious perception of loanwords. Graedler (2014) uses a corpus of Norwegian newspapers to analyze attitudes towards the presence of English in Norwegian. She finds a predominantly negative view in which English is described as an invading force, modified with adjectives such as *contagious, forcible,* and *undermining,* and the resulting lexical borrowings are considered concerning for the future of the Norwegian language.

More favorable perceptions of borrowings were found in Indonesia; Hassall, Murtisari, Donnelly & Wood (2008) explore the perception of western loanwords (those borrowed from English and Dutch) among young tertiary educated Indonesians, using both matched guises and surveys. These borrowings, such as *organisasi* 'organisation', *identitas* 'identity', *favorit* 'favorite', are commonly found in domains associated with modern life. Their matched guise results were mostly inconclusive, with the only significant trend being that participants who agreed with the statement "English is a beautiful language" were less likely to view western-guised speakers as displaying ''modesty'' or ''kindness''. However, this finding was deemed inconsistent with the other results, and the authors concluded that, due to its artificial nature, this method was ill suited for purely lexical variation tasks within a single speech variety. Their surveys revealed a generally favorable attitude towards loanwords, which was positively correlated with understanding loanwords, liking English, and an absence of liking

Indonesian. They parallel their findings with those of Loveday (1996) and Fisherman (1990), whose studies of loanwords in Japan and Israel, respectively, find mostly positive reactions and argue that negative reactions expressed in media outlets merely reflect a vocal minority rather than the general opinion. The cause of this disdain is attributed to notions of linguistic purism, incomprehension of loanwords, lack of education, or loyalty to one's country. The issue of (in)comprehension of loanwords in Japan is further explored in Daulton (2004). In testing a set of 1231 loanwords found in the Mainichi corpus, such as *tero* 'terrorism', *biru* 'building' and *doru* 'dollar', Daulton finds that up to a quarter of loanwords tested were not understood by young adults in Japan, indicating overly liberal use of loans by the media.

Other studies have analyzed the usage of loanwords to better understand their social meaning. Ngom (2003) looks at the usage of English, French and Arabic loanwords in Wolof. His sociolinguistic interviews show how loanwords are used to index group membership or to mark social status in the Senegalese speech community. The use of English, French and Arabic borrowings varies across age and topic and reflects changes within the community. English loanwords, such as *boy-town* and *guy*, hold a growing covert prestige and are most associated with younger males, yet rejected by older participants as rebellious and counter to their culture. Arabic enjoys a higher status due to its connection to Islam. Arabic borrowings, such as *Kilifas* from *qalifa* 'spiritual ruler or leader' and *malaaka* from *malak* 'angel', are frequently employed when discussing religion and are primarily used by older members of society. Finally, French borrowings, such as *marse* from *marché* 'market' and *lekol* from *l'ecole* 'the school', are not differentiated across age groups and maintain a general level of prestige and prevalence in political discussions. However, French's influence appears to be waning due to the emergence of English within the community as the language of socioeconomic success for younger generations.

Expanding beyond the community level, Varga et al. (2012) compare and contrast countries by analyzing English loanwords in French and Italian daily newspapers. They find that differences in each country's language policy result in different types of loanword usage. French resists connotative loanwords –loanwords that have an existing equivalent in the receiving language – while Italian makes liberal use of them, as exemplified the French/Italian dichotomy *le Conseil de stabilité nancière* versus *il Financial Stability Forum*. Yet both languages widely employ denotative loanwords – loanwords that are used to denote new products or new concepts – such as *smartphone* and *PC*.

Zenner et al. (2014) applies a sociolinguistic approach to English loanwords in Dutch, joining a growing collection of loanword research focused on indirect or "weak" contact situation, primarily in Western Europe. This study explores social factors – both speaker related and situational – to see which factors are correlated with increased loanword use. Zenner et al.'s study is one of the first to use oral data, dialogue from a reality TV show, and to make use of complex statistical modeling, specifically mixed logistic regressions. The results show that English is employed by younger, more highly educated male participants, and by those from the core provinces, often to express negative emotions and to show "a high degree of speaker involvement" (11). Zenner et al. conclude that English serves as a tool for individuals to highlight their identities as young and modern.

While many of the above-mentioned studies analyze loanwords in weak-contact situations, places where contact occurs mainly through mass media (i.e. Internet, Hollywood, pop music, television, etc.), a few studies have also attended to lexical borrowings resulting from more direct contact, such as in Spanish within the United States. Matus-Mendoza (2002) finds that many working-class Mexican migrants to the U.S., referred to as *norteños,* make greater use of English borrowings relative to other members of their community in Mexico. English borrowings serve as means of distinguishing *norteños* from nonmigrants, though the former are met with mixed sentiments, including rejection – one community member describes them as "despicable people [who] believe that they have made it" (334) – and admiration – another community member confesses he "want[s] to work on the other side of the border to come back wearing my cowboy boots, my Levis jeans and shirt, and my Texan hat" (333-334). Poplack et al. (1988) document similar findings among French speakers in Canada; working-class French speakers use more English borrowings overall (but not nonce borrowings) than middle-class speakers. A contrasting result was found in the Spanish of New York City, where Varra (2013) reports that members of the upper class present a higher proportion of English borrowings than other social groups, which is explained as a result of the prestige that English holds in the community.

Linguistic Portrait of Argentina

Argentine Spanish is well suited for loanword research, as it has received a large influx of Anglicisms within a mostly monolingual society. Indigenous populations, waves of immigration, and foreign education have enriched Argentina's linguistic profile. This section will provide a brief account of Argentina's linguistic history and a thumbnail contemporary portrait of its linguistic situation.

While Spanish is the country's only official language to date, there are 13 indigenous and 20 European and Asian languages spoken in Argentina (Messineo & Cúneo, 2006). Before the arrival of the Spanish conquistadors in the 1500s, over 35 indigenous languages were spoken in the region. Brutal assimilation and militaristic campaigns in the 1800s displaced and eradicated many of the indigenous populations and, with them, a great source of linguistic diversity. European languages, such as French and Italian, accompanied large waves of immigrants that arrived during the 19th and 20th centuries. Most immigrants quickly adopted the Spanish language within one to two generations, due to the strong nationalistic culture they encountered in Argentina (Bordelois, 2011; Nielsen, 2003).

This legacy of assimilation is attested in the results from a national survey in 2006, which found that less than 10% of those surveyed spoke a language other than Spanish at home (Bein, n.d.). In spite of the small number of heritage-language speakers, almost 50% of those surveyed had some knowledge of a second language. English is overwhelmingly the foreign language of choice; 85% of those who speak a second language speak English, followed by 8.3% who speak Portuguese (Bein, n.d.). This population of L2 English speakers is more highly concentrated among the youth and the middle and upper classes, as knowledge of English is inversely correlated with age and positively correlated with socioeconomic status (Albistur, 2006).

English is more predominant within these groups because most learn foreign languages through the educational system, which over the years has given increasingly more attention to English. In the 19th and early 20th century, both English and French were taught in the public school system. In some cases, Italian, Portuguese and German were also taught, though much more rarely. In 1942, Italian was introduced as an alternative to French, which reduced the presence of French, and thus relatively speaking increased the presence of English. The presence of foreign languages in the public education system continued to grow; in the 1990's, they were introduced in primary school. Due to social, political, and economic pressures, English has prevailed as the foreign language of choice in the public school system, as the languages of immigrant populations, such as Italian and German, have taken a back seat, often found only in private schools. According to Friedrich (2003: 174), “English has been given more emphasis in the [Argentine] educational system than in many other South American countries.”

The growing presence of English in Argentina is visible in the job market as well (Gall & Hobby, 2007). Friedrich’s (2003) meta-analysis of advertisements for managerial positions revealed that over half of the jobs required English and for many other positions, it was considered desirable. English is also perceived to be important; among Argentine MBA students, over 90% believed that people who knew English had greater employment opportunities (Friechrich 2003:181).

The increase of English as a foreign language has been accompanied by a growing presence of English borrowings, embedded within day-to-day language. English appears in the names of “consumer goods, businesses, advertising and fashionable expressions”, which hints at social prestige of these borrowings (Nielsen, 2003:204), though this use of English is not uncontroversial. Bordelois (2011) has named the capital, Buenos Aires, as the Latin American city that has imported the greatest number English terms “unnecessarily”. The prevalence of loanwords is regularly commented on both positively and negatively. Opinion articles in mainstream newspapers have described anglicisms as an invading force that should be avoided except when “necessary”, similar to descriptions of English found in Norwegian news outlets (Graedler 2014, Roffo 2016). Some articles do acknowledge positive attributes of anglicisms, such as their relative brevity – consider the monosyllabic loanword *tip* versus the Spanish equivalent *consejo* (Melgarejo, 2011) – and their association with youthfulness (Pagano, 2013). Regardless of their reception, English-origin loanwords are highly visible across Argentine society.

Anglicism Detection Methods

Since loanwords are often infrequent or sporadic in a given text, large datasets are needed in order to obtain enough tokens of loanwords for statistical analysis. When working with large datasets, which can range from millions to billions of words, automated methods are necessary, as manual processing would be prohibitively time-consuming. Automated methods for language detection have largely focused on the document level, using features such as character encoding (Kikui, 1996), document similarity (Aslam & Frost, 2003), characteristic letter sequences (Dunning, 1994), correlation between word and part of speech (Grefenstette, 1995), syntactic structure (Lins & Gonçalves, 2004) and character n-gram statistics (Cavnar & Trenkle, 1994; McNamee & Mayfield, 2004), as summarized in Hughes, Baldwin, Bird, Nicholson & Mackinaly (2006). These techniques are best suited to larger amounts of textual input, ranging from a few sentences to thousands of words. They are not as effective with mixed-language texts where the language detection needs to occur at a finer granularity (e.g. a word or phrase) (Hughes et al. 2006). Texts with code-switching and/or foreign word insertions require greater precision, as there is much less input (one word versus a document of words). Of the models created for finer-grained language detection, most use one or a combination of the following approaches: character n-gram, pattern matching, and lookup.

Character N-Grams

Character n-grams can be used to guess the language of a word, based on the assumption that common character sequences differ from language to language. For example, the sequence *th* is much more likely to appear in English than in Spanish, while the opposite holds true for the sequence *qu*[3]. Thus if a Spanish-English bilingual were asked to guess to which language the words *queche* ‘a type of sailing ship’ and *thamin* ‘a type of deer’ belong, most likely they would be able to correctly identify *queche* as Spanish and *thamin* as English, even without prior knowledge of these words, simply due to their character sequences.

This is how character n-gram models guess the language of a given word. Just as a bilingual has prior knowledge of both languages through exposure to them, character n-gram models acquire “knowledge” of a language through large bodies of text, called training corpora. The words in the training corpora are segmented into character sequences, called n-grams. The relative frequencies of n-grams in the corpora are used to approximate their probability of appearing a given language, resulting in a statistical profile of n-grams for each language. These statistical profiles are used in language classifiers to compare probabilities between languages. The probabilities of a token being language A or language B are compared, and the token is assigned to whichever language produces the higher probability, as seen in Solorio & Lui (2008a) and Guzman, Serigos, Bullock & Toribio (2016) for Spanish-English code-switched data.

N-gram probabilities can also be used to distinguish native from foreign words by setting a threshold. To identify foreign words appearing within Finnish, Mansikkaniemi & Kurimo (2012) calculate the average perplexity[4] of the character sequence of a word, using a character n-gram model trained on Finnish words and normalized for word

[3] This difference was demonstrated using the Subtlex corpus and ACTIV-ES corpus as representative of English and Spanish respectively, where the sequence *qu* represented 1% of all bigrams in Spanish, yet only 0.03% in English. Even more distinct, the sequence *th* represented 1.5% of all bigrams in English, essentially 0 % in Spanish (5.5e-05).

[4] Perplexity is a measurement of how well a probability distribution predicts a sample. A low perplexity indicates the probability distribution predicted the sample well, while a high perplexity indicates that the probability distribution poorly predicted the sample.

length; words within 30% of the highest perplexity values were considered potential foreign entity names. Koo (2015) uses character n-gram models to identify adapted loanwords from any source language in Korean. Many loanwords undergo vowel insertion in order to repair consonant clusters unacceptable in Korean. The language classifier was trained on a corpus in an unsupervised manner, using native and foreign seed words extracted from the corpus. The model was evaluated on a corpus of 9.2 million words and achieved an F-score of 94.77. The vast majority loanword identifiers are evaluated using the F-score or precision and recall (Leidig, Schlipee & Schultz, 2014).

The F-score is a measurement of accuracy, calculated from precision and recall. Evaluations of loanword identifiers typically calculate the F-score such that equal weight is given to precision and recall, as reflected in the formula below. The more general form of the F-score includes the parameter β, which can alter the amount of weight given to precision versus recall.

$$Fscore = 2 \times \frac{Precision \times Recall}{Precision\ + Recall}$$

Precision indicates, of all tokens identified as loanwords, how many are actual loanwords. It is calculated by dividing the number of correctly identified loanwords (true positives) by the total number of tokens labeled as loanwords (true positives + false positives).

$$precision\ = \frac{true\ positives}{true\ positives\ +\ false\ positives}$$

Recall indicates, of all loanwords in the text, how many are captured by the model. It is calculated by dividing the number of correctly identified loanwords (true positives) by the total number of loanwords present (true positives + false negatives).

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

Pattern Matching

Similar to n-gram models in that it is based on the assumption that languages have particular character sequences, pattern matching is used in language identification tasks, though less frequently. Rather than compare the statistical probability of a given character sequence, pattern matching systems search for specific sets of character sequences, determined a priori. For example, Chesley (2010) proposes an identification system using knowledge of character patterns and foreign morphology to identify all foreign lexical items found within French. She looked at letter sequences not common in French, such as *w, qi,* and *ö,* along with derivational and inflectional morphological paradigms of other languages, such as the Spanish *-ismo* '-ism'. This system achieved a recall of 64.29%, tested on just 14 borrowings, of which 9 were identified. Andersen (2005) also tries pattern matching with low success, along with a series of other techniques to identify anglicisms in a list of neologisms extracted from Norwegian text.

Look-up

Look-up techniques, as the name implies, simply look up a given token in a reference, often a dictionary or word list. Alex (2006, 2008) proposes a lookup system to identify English loanwords in German, using two sources: monolingual lexicons and back-off search engines, followed with a post-processing heuristics module. The model first looks up tokens in German and English CELEX lexicons. Tokens found in both lexicons are classified in post-processing heuristic modules. Tokens not found in either lexicon are looked up in the search engine Yahoo. Alex finds the search engine outperforms a fixed corpus; a comparison of various fixed corpus sizes and a Yahoo web search revealed that a corpus considerably larger than 40m tokens would be required for the corpus module to perform as well as the search engine module. This finding reflects the inherent limitations of fixed-size corpora; they are unlikely to contain all possible lexical items and, because languages constantly evolve, they will never be truly up-to-date. The model achieves an F-score of 83.18 on unseen test data. In spite of the benefits of the search engine approach, the author admits three drawbacks: it is more computationally costly, time-consuming and limited by the search engine's search limit.

Combination

Several recent studies have sought to exploit the benefits of each model, using a combination of both n-grams and lookup modules. Andersen (2005) tested a series of techniques, finding that a combination of looking up characteristically English character n-grams and pattern-matching outperformed single-technique systems, with a recall of 49% and precision 75%. It should be noted that this method was not tested on running text, but rather on a word list. Rosner & Farrugia (2005) proposed identifying English-Maltese code-switching in text messages using a system that combines Hidden Markov Model language tagging with dictionary lookup and character-based n-gram modeling. Hidden Markov Models, frequently used in part-of-speech taggers, are sequence classifiers that use the probability distribution over possible sequences of tags to determine the best tag for a given token. While this method yields accuracies of 95%, there is no information provided about the amount of code-switching present, so it is difficult to compare this system to those that work on loanword identification.

Leidig et al. (2014) developed an English loanword detection system comprising 5 features. The features were based on character patterns, (n-gram perplexity feature and G2Pconfidence) and look-up techniques (Hunspell, Wiktionary, Google hit count). The n-gram perplexity feature is determined by comparing the scores generated from caseinsensitive 5-gram models for English and the matrix language. G2Pconfidence is similar to the n-gram perplexity feature, differing only in that it compares scores at graphone-level instead of grapheme-level. The graphones were created using Phonetisaurus, an open Grapheme-to-Phoneme conversion toolkit. The look-up features make use of three sources to compare a token's presence in English and the matrix language: Hundspell, an open source spell-checker and morphological analyzer, Wiktionary, a community-driven online dictionary and Google hit count, the normalized count in language-restricted Google searches.

The features were first evaluated separately, then combined using three approaches: Voting, Decision Trees, Support Vector Machines (SVM). Voting is the most basic approach, in which each feature's classification counts as one vote in determining the final language tag. Decision Trees and Support Vector Machines are machine-learning methods that are commonly used for classification in a variety of natural language processing tasks. The highest performing single feature was the G2Pconfidence, followed by n-gram perplexity score, thus the character pattern features were most effective. When combining the features, the simple voting technique worked best on two of three tested datasets. They tested these models on three different word lists, which varied in their quantity of Anglicisms: 15% anglicisms, 4% anglicisms, and 2% anglicisms. All models performed much better on the high-percentage anglicisms, with F-scores of 75%, than on the other two, with F-scores of 62% and 52%, respectively.

METHODS

In an effort to explore the social dimensions of anglicisms in Argentina, this study makes use of two corpora: ACTIV-ES and NACC. This section describes the corpora and explains the algorithm created to identify anglicisms.

Data

The two corpora used in this study represent different types of language: informal spoken (film dialogue in the ACTIV-ES corpus) and formal written (newspaper articles in the NACC corpus). Using both corpora affords a more complete picture of anglicisms in Argentina, accounting for differences across speech types and registers.

ACTIV-ES

ACTIV-ES is an open corpus, created from an online repository of TV and film subtitles from Argentina, Spain, and Mexico, to represent 'everyday' Spanish (Francom, Hulden & Ussishkin, 2014). This corpus provides free access to data that can be costly to create through traditional means such as transcribing sociolinguistic interviews. While most film dialogue is scripted and therefore not technically naturalistic data, Brysbaert and New (2009) have demonstrated that language data from TV/film subtitles can provide a closer proxy to everyday speech than more formal naturalistic data, such as the parliamentary proceedings found in the Europarl corpus. This finding is intuitive, as most films and TV shows strive to mimic everyday natural speech. Francom et al. (2014) tested the lexical content of ACTIV-ES to see if it represented well each of the three Spanish varieties. A visual word recognition task revealed that the lexical behavior of native speakers from each of the three populations correlated significantly with lexical content (frequency and dispersion of words) in their respective sub-corpora and not with that in the other sub-corpora. This finding leads the authors to suggest that, "at the word level, lexical variation found in the ACTIV-ES corpus approximates particular usage patterns of the respective native populations and provide support for the representativeness of these sub-corpora" (1737).

Though in its entirety the ACTIV-ES corpus represents three Spanish varieties, for this study only the more recent films (after 1990) from the Argentine sub-corpus are considered; this limited timeframe was chosen in order to provide a more accurate comparison with NACC data, which is from 2012-2013. As the Argentine sub-corpus is quite modern, with 86% of the movies being from 1990 or later, not much data was lost. This modern portion of the ACTIV-ES corpus contains 955,789 words from the scripts of 106 movies. The corpus offers several layers of metadata, which makes it particularly amenable to sociolinguistic study: country, year, genre, and Internet Movie Database (IMDb) ID. The distribution of movies across genres and time can be seen in Figure 1, where each circle represents a film transcript; the size of the circle corresponds to the length of transcript. Not surprisingly, the most represented genres are drama and comedy, with 58 and 32 movies respectively.

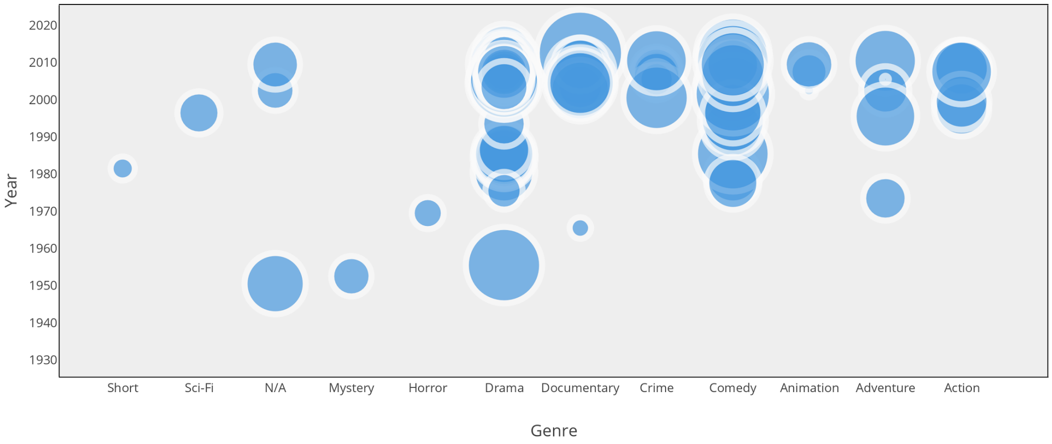

Figure 1: ACTIV-ES Argentine Films by Year, Genre, and Word Count

NACC Corpus

The NACC Corpus was created by the researcher, from three national newspapers: La Nación, Clarín and Crónica. The 1.8 million-word corpus comprises the text from all of the articles appearing on each newspaper's online home page from November 2012 to March 2013. The articles were downloaded, compiled, and stripped of their html format, using the programming language Python and the html library BeautifulSoup. These articles cover a broad range of topics, including politics, social issues, opinion, business, sports and fashion. Each newspaper represents around one third of the corpus: La Nación (32.9%), Clarín (32.4%), and Crónica (34.6%). Issues with downloading and text encoding caused a small percentage of articles to be eliminated, thus not all articles from this time period are represented.

The three newspapers were chosen to represent different segments of the population, as each newspaper is known to have a distinct target audience. The connection between social status and newspaper readership has been supported by Chan & Goldthorpe (2007), who found that a reader's status affected their choice of newspaper, independently of the effects of education. La Nación, the second most read newspaper in the country, targets a high socioeconomic-status clientele. It has earned a politically conservative reputation and was in long-standing conflict with the Kirchner leftist government, which was in power at the time of data collection (Boczkowski & de Santos, 2007). In contrast to La Nación, Clarín initially had a more favorable outlook towards the Kirchner government, as the newspaper is considered to be more centrist. However, the relationship began to sour in 2008, amid a conflict between the government and the agricultural sector, when Clarín published articles considered to be favorable toward the agricultural sector and thus against the government. Tensions continued to mount throughout the remaining years of the Kirchner government over a series of media laws and court cases. Clarín boasts the largest readership in the country and is oriented towards the middle class. As of 2004, over 5.3 and 1.4 million unique users accessed Clarín.com and Lanación.com monthly, establishing both websites as the top two general interest online newspapers in the country (Boczkowski & de Santos, 2007). Crónica, though not as widely read as Clarín and La Nación, is still a major newspaper distributed throughout the country and is targeted to a lower socio-economic readership. In contrast to Clarín and Nación, it was considered to be pro-Kirchner at the time of data collection.

These distinct target audiences for the newspapers allow for a uniform though imperfect manner of dividing the data into social groups. One challenge of using corpora is that many lack extensive metadata on speakers. This lack often limits researchers' ability to include extra-linguistic variables, such as age, social group, etc., in their analysis. Information about newspaper writers can also be challenging to access. However, target audiences of newspapers are often well known. This information is crucial for newspapers, as they carefully consider their audience when choosing content. Because presumably aspects of the writing (including sentence structure, syntax, and word choice), like the content, are tailored to the readership, the social characteristics of a given readership can serve to form the extra-linguistic factor: social group.

Loanwords

Before identifying English loanwords in the corpus, it was necessary to set the parameters as to what constitutes a loanword, as there are numerous definitions for the term. This study adheres to the definition provided by Haugen (1950), as discussed in the first chapter of this dissertation: Loanwords are words whose phonemic shape and meaning have been imported into a recipient language with no morphemic substitution. This definition includes only loanwords that are identifiable as English, thus loanwords that have been orthographically or morphemically modified will be excluded. For example, the adapted loanword *fútbol* 'football/soccer' or *chatear* 'to chat via an electronic messaging platform' are excluded from analysis, whereas unadapted loanwords, such as *gadget, backstage,* and *make up,* are included.

This narrow definition of loanwords suits this Spanish variety and the goals of this research well; a manual examination of segments of the corpus and the researcher's prior knowledge of the Argentine dialect reveal that the vast majority of loanwords remain unadapted upon initial adoption to the Argentine dialect. Those that are adapted are most often long-standing loanwords, which are commonly viewed by native speakers as part of the Spanish language, rather than foreignisms, for example *gol* 'goal', *éstres* 'stress', *esmoquin* 'smoking jacket', *bol* 'bowl'. This study is interested in newer borrowings that are more identifiably foreign, as the focus is on their social function. It should be noted that the adapted/unadapted distinction does not provide a fail-proof test between new and established borrowings. Some long-standing loanwords remain unadapted, such as *bar* and *club*, likely due to the fact that their existing form is already amenable to the recipient language. Similar to many linguistic categorization schemes, the boundary between loanwords and native words is blurry. The perception of these two categories differs across time, regions, and speakers and most likely is more of a continuum than two discrete categories. This issue of perception and categorization is further addressed in the discussion section of this chapter.

Identification System

To find all tokens of English origin in the Argentine corpus, an Anglicism Identifier was developed by the researcher. Building on the previous language classifiers discussed in the literature review, the Anglicism Identifier uses a combination of n-gram and look-up methods to distinguish English from Spanish tokens. The classifier adds layers of annotation to address the role of Named Entities (NEs), lemmas, and loan phrases. Several parameters of the classifier were tested to optimize the performance and are presented in the results section.

The first step in the classifier is tokenization of the corpus. This is done using regular expressions to produce a chronological list of words and punctuation. The Anglicism Identifier then evaluates each token, producing four tiers of annotation: lemma, language (Spanish, English, Punctuation, or Number), NE, Anglicism. A sample output for the sentence *Romina Renom dijo,* "*los tapados oversized son un must*" is seen in Table 1.

Table 1: English Identifier Sample Annotation

Token	**Lemma**	**Language**	**Named Entity**	**Anglicism**
Romina	Romina	Spn	Yes	No
Renom	Renom	Spn	Yes	No
dijo	decir	Spn	No	No
"	"	Punct	No	No
los	el	Spn	No	No
tapados	tapado	Spn	No	No
oversized	oversized	Eng	No	Yes
son	ser	Spn	No	No
un	un	Spn	No	No
must	must	Eng	No	Yes
"	"	Punct	No	No

The overall architecture of the Anglicism Identifier is presented in Figure 2. For the language tag, a simple look-up approach is applied to identify punctuation and numbers. For tokens not identified as either, a language identifier module with a character n-gram and look-up components, visualized in Figure 3, is employed to determine whether the token is Spanish or English.

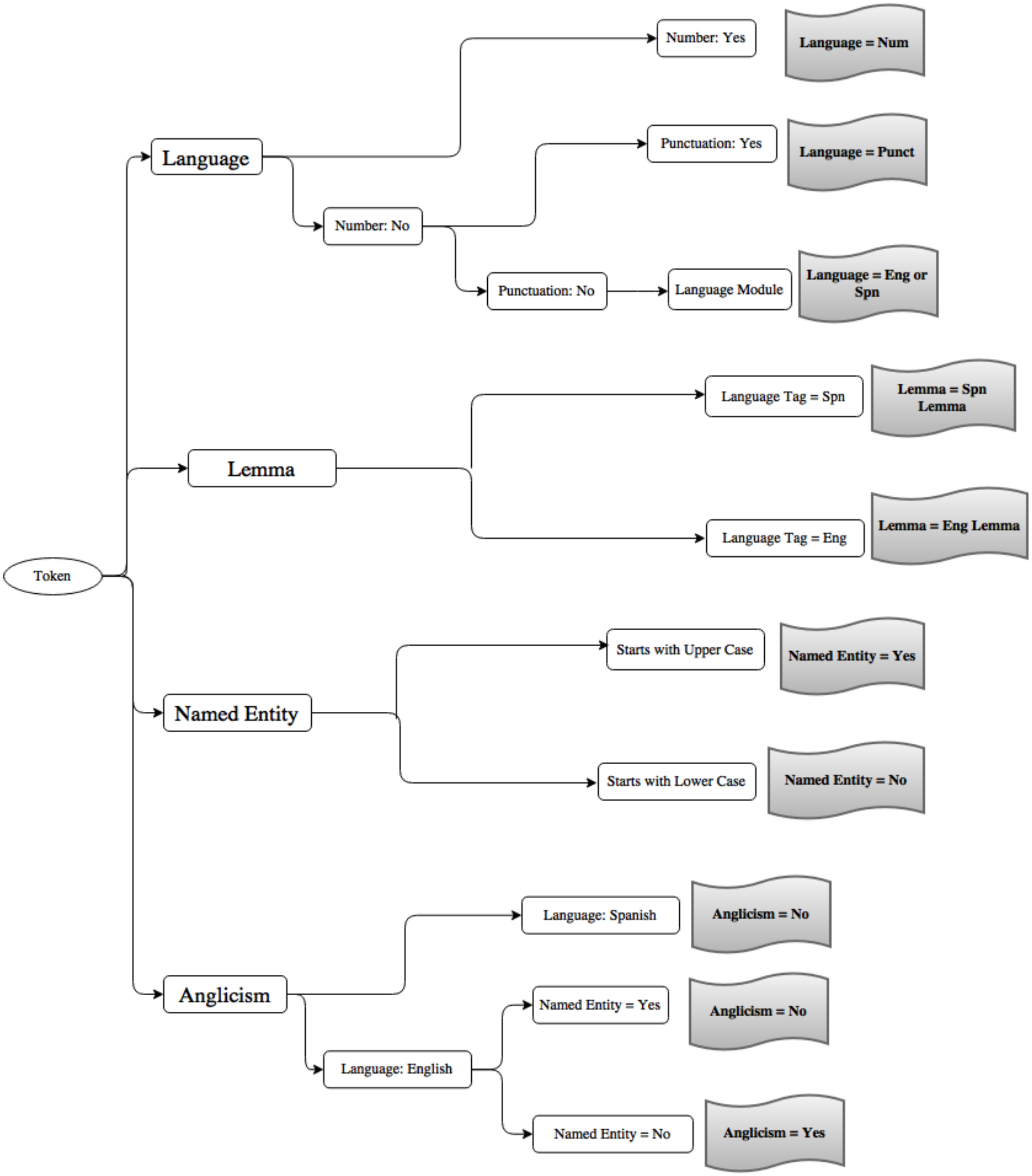

Figure 2: Architecture of the Anglicism Identifier.

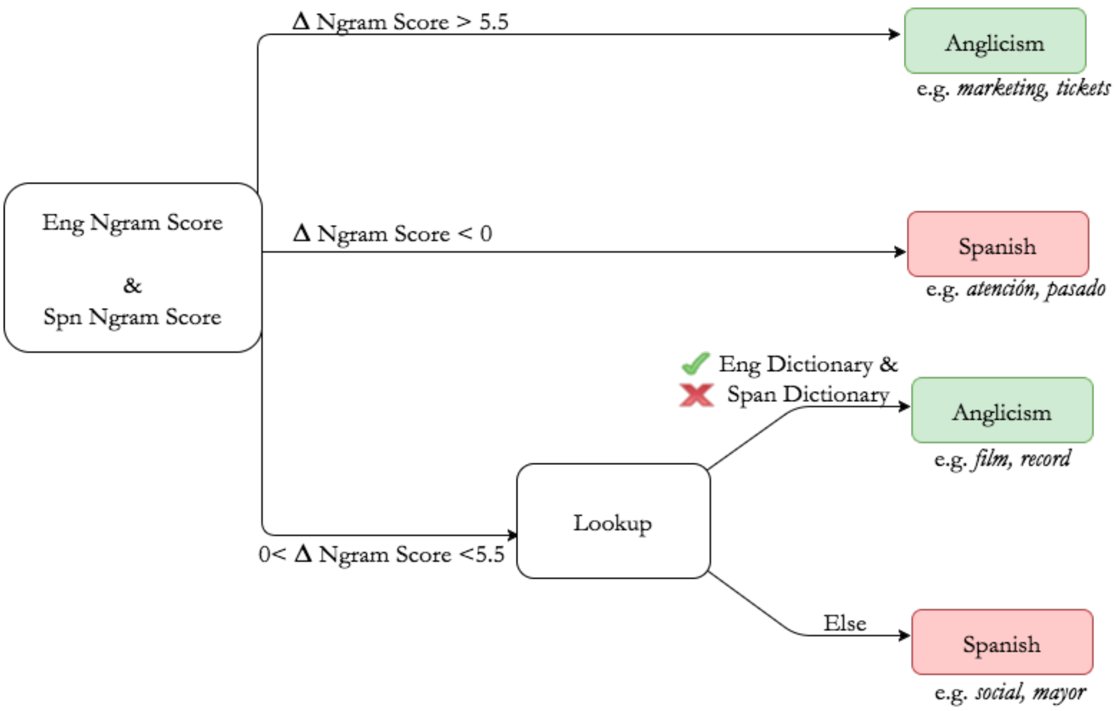

Figure 3: Architecture of the Language Classifier.

The character n-gram model calculates the probability that a token is English or Spanish and computes the difference between the two. Log probabilities are used due to the small scale of difference.

$$\Delta\, Ngram\ Prob\ =\ log(Eng\ Prob)\ -\ log(Spn\ Prob)$$

A threshold for the log-scale N-gram Prob is used to determine if there is a large difference between the two probabilities, in which case the language tag is determined by the character n-gram module, whichever language is more likely. If the difference is small, i.e. the likelihood of the token being either language is similar, the token is processed in the look-up module. Only if the token is both in the English dictionary and not in the Spanish Dictionary is it classified as English, otherwise it is Spanish. This bias towards Spanish is appropriate for this dataset, as native Spanish words overwhelmingly outnumber anglicisms.

For the NE tier, two options were tested and evaluated: the Stanford NE recognizer with the English and Spanish parameters and a simple capitalization test (if token starts with capital letter, then NE = Yes, else NE = No). The capitalization test is clearly very error-prone in that it over-identifies NEs, labeling all sentence initial tokens as NE. However, as the purpose of this system is to correctly identify anglicisms, the NE module functions only to avoid the over-identification of anglicisms posed by English-origin NEs. For example, US companies (e.g. *General Motors*) and Hollywood movies (e.g. *The Godfather*) are identified as English in the language module, but by also having the NE label, these tokens are not labeled as anglicisms.

A slightly more sophisticated capitalization test was considered, one in which only capitalized words that are not sentence-initial are considered NEs. For example sentence *Las peleas en Hollywood abundan* 'Fights in Hollywood abound', the capitalized *Hollywood* would be labeled as an NE, whereas the capitalized feminine article *La* would not, due to its sentence-initial position. While this approach mitigates some of the NE over-identification, it leads to a less successful identification of anglicisms because the model finds many more false positives, resulting from numerous instances of sentence-initial English-origin NEs, such as *Mark Zuckerberg reveló varios detalles* 'Mark Zuckerberg revealed several details'. Meanwhile, the benefits of this fine-tuned capitalization test were almost nil; only one previous missed anglicism was captured in the gold standard – the token *We* found in the code-switched phrase *We love* *celebrities*. The lack of anglicisms in sentence-initial position is to be expected, as Spanish generally does not allow bare nouns in sentence initial position.

The Anglicism tag results from the simple formula;

If language = English and NE = No, then anglicism = Yes, else anglicism = No.

Additionally, to capture loan phrases, such as *little black dress*, anglicisms occurring sequentially are marked as a phrase and count as one anglicism.

The final annotation layer, lemma, is determined using an English or Spanish lemmatizer from the python module Pattern (De Smedt & Daelemans, 2012), whichever corresponds to the language tag. The lemmatization of the token is what allows for a more complete look-up in the Spanish and English dictionaries. For example, the conjugated Spanish verb *habla* 'he/she speaks' or the English plural noun *dogs* do not appear in their respective dictionaries, yet their lemmatized forms *hablar* and *dog* do, so by looking up the lemma rather than the token, the language model achieves higher accuracy.

RESULTS

Anglicism Identifier

The Anglicism Identifier was trained on a 30,000-word subsection of the NACC and ACTIV-ES corpora to set the parameters and tested on another 30,000-word subsection of the corpora. To evaluate the model's performance, the F-score is utilized instead of accuracy. Accuracy, which measures the number of correct predictions out of all the predictions made, can be less effective for highly unbalanced datasets, such as anglicisms within a corpus of native words. As an illustrative example, a model that tags everything as a native word, in a corpus with 1% loanwords, would receive an accuracy of 99% even though it fails to capture a single loanword. A better metric for these types of unequally distributed datasets is the F-score. As mentioned in the literature review on existing loanword identifiers, the F-score is calculated from precision and recall, precision being defined as true positives divided by the sum of true positives and false positives and recall being defined as true positives divided by the sum of true positives and false negatives.

$$Fscore = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

While some studies evaluate loanword identifiers on word lists, compiled from newspaper corpora and existing neologism lists, both the training and test data for this study are subsets of the corpora, which is important for two reasons. First, context, which is lacking in word lists, can determine whether a token is a borrowing or not. For example, in a word list the English-Spanish homograph *animal* would be easily classified as a native Spanish word. However, given the context, *una remera con estilo animal print* 'a shirt with an animal print style', as appears in NACC, the bigram *animal print* should be classified as a loanword. Secondly, using a representative corpus instead of a word list leads to a more accurate calculation of precision. As anglicisms are often so few in number relative to native words (rates around 1% are in fact high for many datasets), precision may be artificially inflated when working with a word list containing a high percentage of anglicisms (e.g. 5.63% in Andersen (2005), and 15%, 4%, and 2% anglicisms in Leidig et al. (2014)). As highlighted in Leidig et al. (2014), the F-score dropped dramatically from 75% to 55% when their model was evaluated on a word list containing 15% anglicisms versus a word list containing 2% anglicisms.

In the development of this model, the following parameters were set using the training data: NE identifier (Stanford NE recognizer vs. capitalization), n-gram number, and threshold. The model with the highest F-score used the following parameter:

capitalization for NE recognition, 4-gram, and a threshold range of 0 < x 5.5, meaning that any token whose difference between the Spanish and English log n-gram probabilities falls between 0 and 5.5 is sent to the look-up model. This model achieves an F-score of 79.41% on the training data and 76.25% on unseen test data. To highlight the value of the combined approach over the n-gram and look-up modules alone, both modules were evaluated separately on the test data and, as expected, performed poorly, with F-scores of 19% and 47%, respectively. To show the contribution of the lemma and the NE modules, the model without the Lemma component and the model without the NE component were each evaluated on the test data, receiving F-scores of 72% and 42%. The NE component makes a greater contribution to the overall performance by lowering the number of false positives, such as *General Motors*, and therefore increasing precision.

Table 2: Performance of the Anglicism Identifier.

	N-gram	Lookup	Without Lemma or NE	Without Lemma	Without NE	Mixed (Training)	Mixed (Test)
Accuracy	96%	100%	98%	100%	99%	100%	100%
Precision	10%	47%	23%	67%	28%	75%	72%
Recall	93%	46%	85%	77%	85%	84%	77%
F-Score	19%	47%	36%	72%	42%	79%	75%

Reviewing the remaining errors in the highest performing model reveals some challenges for this model, primarily related to the over-identification of anglicisms: neologisms/NEs, borrowings from other languages (mostly French and Italian), and adapted loanwords. These issues are not merely technical glitches, but rather reflect the challenges of defining a loanword to begin with.

Many neologisms identified by the model as loanwords are themselves related to English vocabulary, adapted from NEs and often technology-related, like *googlear* 'to google' or *instagramear* 'to instagram'. Other adapted NEs also appeared, such as the adjectives *thatcherismo* and *shakesperiano*, based on famous English historical figures Margaret Thatcher and William Shakespeare, respectively. While they are not considered anglicisms under the definition used in this study, they are closely related and represent a connection between the Spanish and English speaking worlds. In the same vein, the English adjective *quixotic* is derived from the Spanish literary character Don Quixote, yet is not considered to be a borrowing from the Spanish language itself.

The second major challenge for this model is borrowing from other languages, primarily French and Italian. The model in its current state creates a binary classification system in which each token is evaluated to see if it is more like Spanish or more like English. This binary view clearly does not represent reality, thus an improvement to the model would take that fact into account. French poses an additional challenge in that there is a great deal of overlap with English, resulting from the large influx of French loanwords into the English language through its history (e.g. *amateur, garage, corset, romance*). Many of these French loanwords appear in Spanish as well. While some instances of French in the corpora are clearly directly borrowed from French, such as *oui oui* 'yes yes' and *mon amour* 'my love', some may be borrowed through English. For example *pantalones [cigarettes]* 'cigarette pants' or *hotel [boutique]* 'boutique hotel' are concepts that originated in English utilizing French borrowings (*cigarettes* and *boutique*) that are so well established in the lexicon that many native English speakers may not recognize them as Gallicisms. Here again we see that context is key to understand the nuance of these words. Without context, *cigarettes* and *boutique* may be classified as Gallicisms in Spanish, but given the correct context, perhaps they are borrowed via

English. For the sake of consistency, any word that appears in both English and French and is identified as being of French origin in the Diccionario del Real Academia Español is not considered an anglicism. Thus, words such as *amateur, chance, garage, corset,* and *romance*, along with *boutique* and *cigarette,* are not labeled as anglicisms.

Lastly some adapted loanwords are identified by the model, such as *samplear* 'to sample', *pitchear* 'to pitch' and *singlista* 'singles tennis player'. Under the definition utilized in this study, these adapted loans are not considered anglicisms; however, they prove of interest when examining the influence of English on Argentine Spanish, using a broader definition of lexical borrowings. As training and test datasets for this study are not annotated for adapted loans, it is currently unknown what percentage of adapted loanwords the model captured. However, re-annotating the test and training datasets will allow this model to be retested to see its success at capturing both adapted and unadapted borrowings. What to include and exclude is unfortunately not as clear-cut as it seems a priori, as many other studies attest (Andersen, 2005). Thus, while to a certain extent, the criteria may be arbitrary, the most important factor is consistent and reproducible methods. The guidelines set for the annotation process allow for the reproduction of this study.

Corpus

The analyses of anglicisms in both corpora are presented in this section. For this analysis, the anglicisms identified by the classifier were manually reviewed to remove all false positives; this manual inspection positively affects precision, raising it to approximately 100%, though recall (measured at 77% in the test set) is unaffected. The newspaper corpus, NACC, is treated as three separate datasets, resulting from the three national newspapers, in which each article represents a data point. The film corpus, ACTIV-ES, is treated as one dataset. The 106 scripts, which comprise the dataset, have an average word count of 9,070. To use each script as a data point, while conceptually sound in that each script represents a unique, independent work, proves problematic when comparing them to the newspaper articles, since the articles have an average length of only 570 words. In order to divide the film corpus into shorter segments, each script is divided into 20 equal length segments, resulting in segments with an average word count of 465, much closer in length to the newspaper articles. As this method of segmentation does not produce data points that are truly independent of one another, film segments will be compared to full film scripts throughout the analysis. The datasets were filtered to remove any data points with word counts of less than 100; this served to remove several articles that provided merely captions to pictures or advertisements. Additionally, two films, *Unen Canto con Humor* and *Grandes Hitos Antologia,* were removed from the analysis because, as musical performances of the group Les Luthiers, the vast majority of the scripts is composed of song lyrics and thus is not representative of speech.

The datasets are subjected to three statistical analyses; all analyses were performed in R, an open source programming language designed for statistical computing (R Core Team, 2017) and the visuals were produced using the R package ggplot2 (Wickham, 2009). The first analysis provides a general overview, examining the rates of loanword usage across all datasets. The second takes a binary approach, considering if data points (articles and film segments) contain loanwords or not. The third analysis examines the rates of loanword usage across all datasets for only data points containing loanwords. The first and third analyses are based on anglicism rates calculated with type counts, rather than token counts. For example, a 1,000-word text containing the anglicisms [*set*, *set, tie break*] is considered to have two types of anglicisms [*set, tie break*], resulting in an anglicism rate of 2/1,000*10,0000 = 200 anglicism types per 100,000 words.

First Data Analysis

This analysis considers all articles and film segments, including those with no anglicisms, with respect to their anglicism usage. The boxplot in Figure 4 shows the rate of anglicism types per 100,000 words across the four datasets. To ease the visual comparison of the datasets, the outliers, which account for less than 2% of the data, have been removed. As seen in Figure 4, the data heavily skews to the right, so robust measures of central tendency (i.e. median and interquartile range) are preferred over the mean, as they are less sensitive to outliers and thus, perform better for skewed data; these statistics, median and interquartile range, are presented in Table 3. As seen in Figure 4, there are two primary divisions across the datasets. First is that separating La Nación from the rest of the datasets, reflected in its median of 142, compared to the zero medians of Clarín, Crónica and Film Segments. The zero medians result from the high number of articles and film segments without anglicisms. Additionally, La Nación shows much greater variability in anglicism rates, reflected in its wide interquartile range, which is over double the size of other datasets. While La Nación clearly stands apart from the other datasets, all three newspapers (La Nación, Clarín, and Crónica) reflect much higher use of anglicisms than the Film Segments, of which almost 80% did not employ a single anglicism, hence their null representation in the boxplot (zero interquartile range and median).

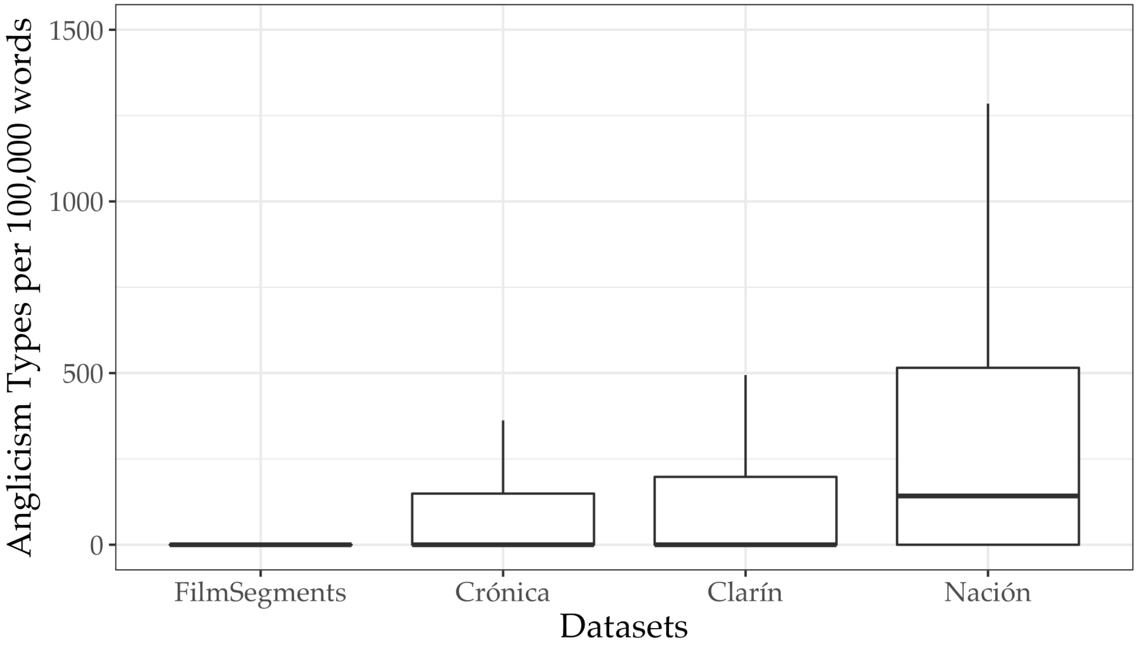

Figure 4: Anglicism Type Rates Across Datasets.

Table 3: Robust Statistics by Dataset

	FilmSegments	cronica	clarin	nacion
Number of Observations	4474	967	1119	970
Median	0	0	0	142
Interquartile Range	0	149	198	515

In order to test the significance of the anglicism rate distributions across the datasets, a linear regression was considered with the predictor variable, Dataset, and the response variable, Anglicism Types per 100,000 words. However, the skew of the data and the nonlinear pattern of the residuals, visualized in the Quantile Quantile Plot in

Figure 5, indicate that the assumptions for a linear regression are not met. Attempting to normalize the data with a log or logit link function still resulted in a nonlinear patterning of the residuals, also seen in Figure 5.

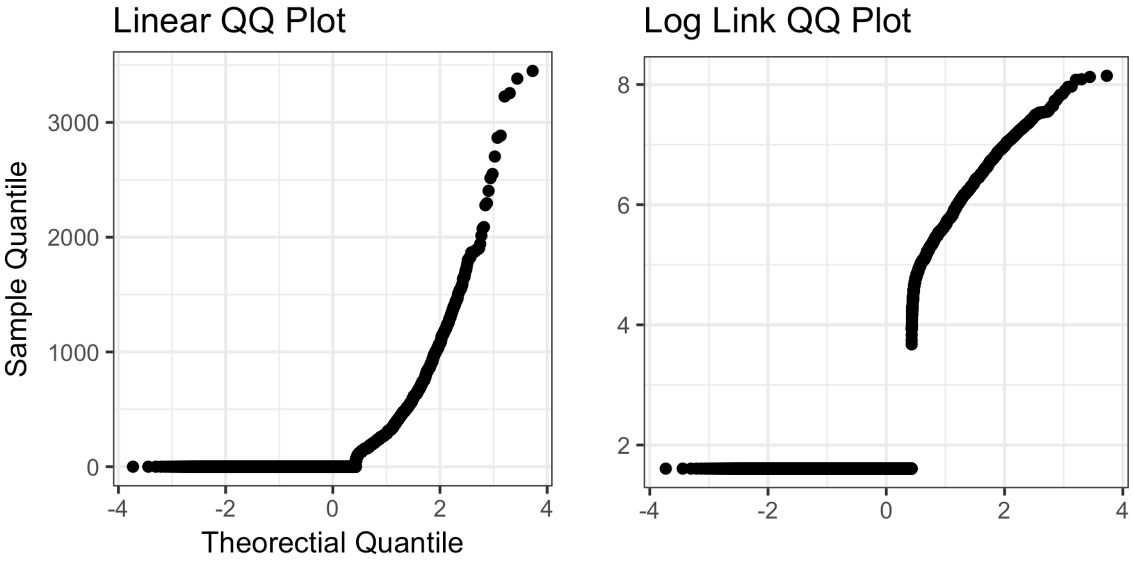

Figure 5: Quantile Quantile Plots.

Therefore, a non-parametric alternative is preferred. The data meets the assumptions for a non-parametric test (i.e., samples are from populations with similar distributions), as visualized in the histograms in Figure 6 below[5]. All have a similar form in that they skew to the right and are unimodal. Using the R package FSA: Fisheries Stock Analysis (Ogle, 2017), the Kruskal-Wallis rank sum test was performed with a post-hoc pairwise comparison, the Kruskal Dunn test, which is appropriate for groups with unequal numbers of observations (Zar, 2010). Table 4 shows unequal numbers of observations between datasets. The Kruskal-Wallis rank sum test reveals a significant difference

[5] It should be noted that the y-axis of the histograms was shorted, cutting off part of the zero counts bar, in order to improve the visualization of the overall shape of the distribution. The scale of zero counts relative to non-zero counts is represented in Figure 7.

between datasets and, as seen Table 4, the Kruskal Dunn test provides pairwise comparisons, all of which are significant. The vast number of zero counts, which strongly affect the median and interquartile range, relative to non-zero counts motivates the need for additional analyses to better divide the data. The following analysis treats the data in a binary fashion, in which article and film segments are classified as containing loanwords or not, to bring a sense of proportion to the zero counts.

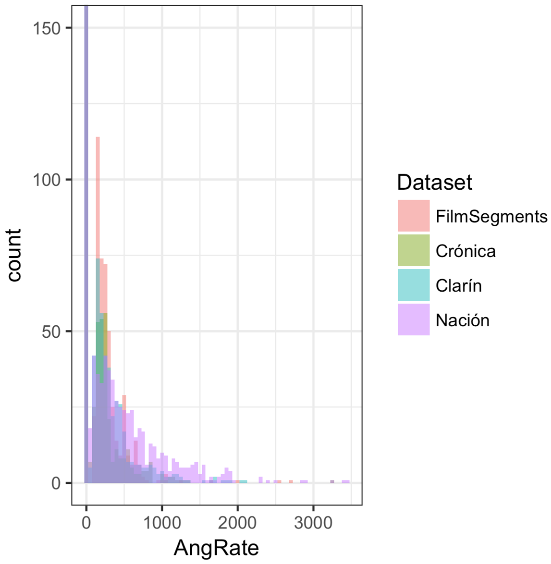

Figure 6: Histogram of Anglicism Rates across Datasets

Table 4: Post Hoc Pairwise Comparison.

Comparison	Z	P unadjusted	P adjusted
Clarín - Crónica	4.031177	5.55e-05	5.55e-05
Clarín – Film Segments	15.922992	0.00e+00	0.00e+00
Crónica – Film Segments	10.016473	0.00e+00	0.00e+00
Clarín – Nación	-11.880283	0.00e+00	0.00e+00
Crónica - Nación	-15.363990	0.00e+00	0.00e+00
Film Segments - Nación	-29.741861	0.00e+00	0.00e+00

Binary Data Analysis

From a binary perspective, the documents (articles and film segments) are analyzed using the response variable: (1) contains anglicisms or (0) not. Table 5 shows the number of documents containing anglicisms and not for each dataset and Figure 7 shows the equivalent percentages. In Figure 7, we see a pattern similar to that in the first analysis; La Nación stands apart from the other datasets in its likelihood to use anglicisms in its articles. In fact, more often than not, an article in La Nación will employ an anglicism, whereas Crónica and Clarín will use anglicisms in only about one in every three articles. The film segments employ anglicisms at an even lesser rate, one in every five. In contrast, when the film scripts are viewed as a whole, they employ anglicisms at the highest rate, 86%. However, it is not relevant to compare this rate to the other datasets; due to the much larger word count, there are simply many more chances of having a loanword. This implies that, while the films have the lowest rates of anglicism types, as seen in analysis 1, it is almost unavoidable to use at least one loanword at some point in a film. Of 106 films, only 15 did not contain anglicisms. Notably, ten of those 15 films are set in the past, thus while these films themselves may be current, they depict language from a different era, such as in *Fierro,* an animation film about Martín Fierro set in the 1800s, or in *Vientos de agua,* the story of a Spanish immigrant in Buenos Aires during the 1930s. This leaves only five films set in the present that do not implement loanwords.

Table 5: Counts of Documents Containing Anglicisms Across Datasets.

	Yes	No	Total
Films	92	15	107
Film Segments	934	3540	4474
Crónica	300	667	967
Clarín	426	693	1119
Nación	557	413	970

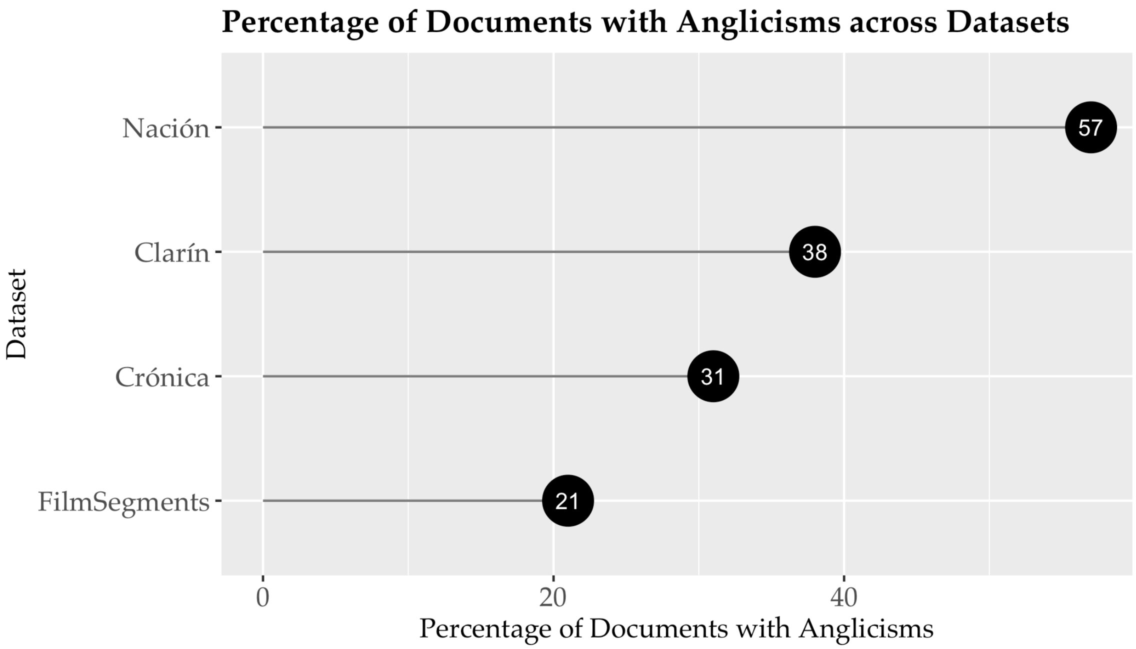

Figure 7: Percentage of Documents with Anglicism across Datasets.

Using the glm and pairwise.prop.test functions in R, a logistic regression with a Post-Hoc comparison was conducted, revealing significant differences between all four datasets with respect to the binary response variable: contains loanwords or not. To measure the effect size, the odds ratios of all dataset pairs are presented in Table 6. La

Nación stands apart from all the other datasets: a La Nación article is at least twice as likely to contain loanwords as any other dataset and jumps to 4.8 times as likely as a film segment to contain loanwords (i.e. 380% more likely). The difference between Clarín and Crónica, while significant, is small in comparison; a Clarín article is only 37% more likely to contain loanwords than one of Crónica. A slightly greater difference is found between Crónica and Film Segments; a Crónica article is 60% more likely to contain loanwords.

Table 6: Odds Ratios Between Datasets.

	Clarín	Crónica	Film Segments
La Nación	2.19***	3.00***	4.80***
Clarín		1.37***	2.19***
Crónica			1.60***

Anglicisms Only Analysis

The final analysis considers the anglicism rates only in the articles and film segments that contain loanwords. The boxplot in Figure 8 shows a similar pattern to Figure 4, except that in these boxplots the medians are no longer weighed down by zeros. La Nación again is distinct from the other three datasets; its median and interquartile range nearly double that of the other datasets. The distinction between oral data (Film Segments) and other newspapers (Clarín and Crónica) does not appear to be as distinct as in the first analysis, reflecting that much of the difference between the datasets is accounted for by zero count proportions. Once the zero counts are removed, the remaining articles and segments employ anglicisms at more similar rates, reflected in their similar medians.

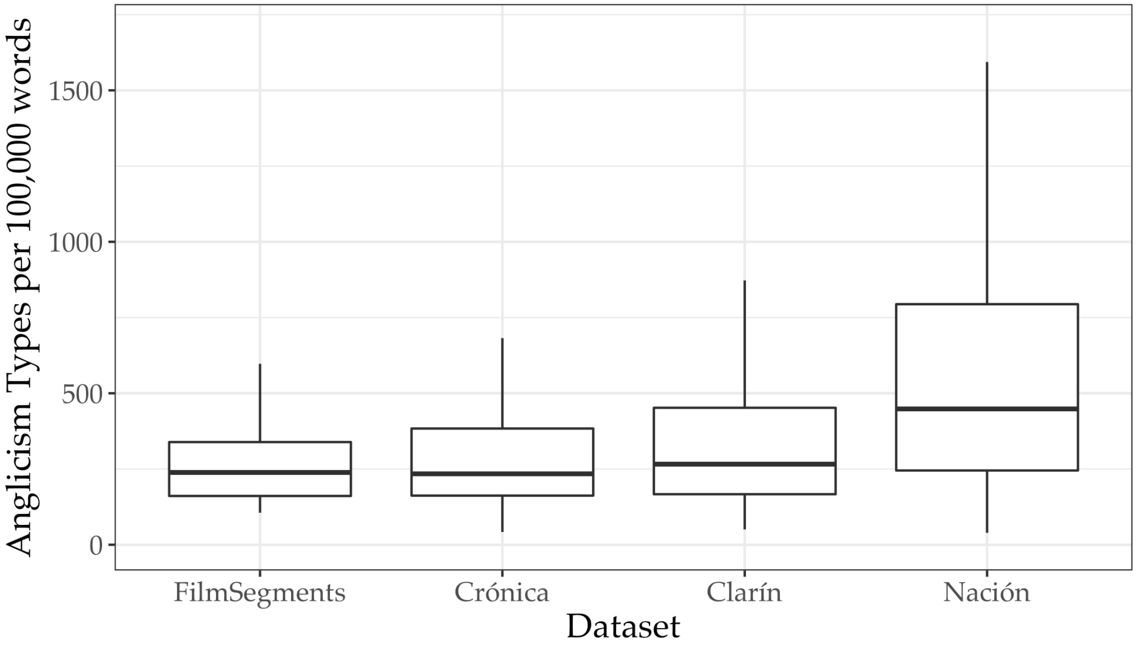

Figure 8: Anglicism Type Rates Across Datasets (Zero Counts Removed).

Table 7: Robust Statistics for Datasets (Zero Counts Removed)

	Film Segments	Crónica	Clarín	Nación
Number of Observations	457	300	426	557
Median	239	234	266	448
Interquartile Range	178	221	285	549

As in Analysis 1, a linear regression and a linear regression with a log or logit link function were considered to test the significance across the datasets. However, the nonlinear pattern of the residuals, visualized in the Quantile Quantile Plots in Figure 9, indicate that the assumptions for a linear regression (with or without a link function) are not met. Therefore a Kruskal-Wallis rank sum test was performed with a post-hoc pairwise comparison, the Kruskal Dunn test. The results of the post-hoc test are listed in Table 7 below. La Nación differs significantly from all other datasets. However, the pairings between Clarín-Crónica and Crónica-Film Segments did not differ significantly.

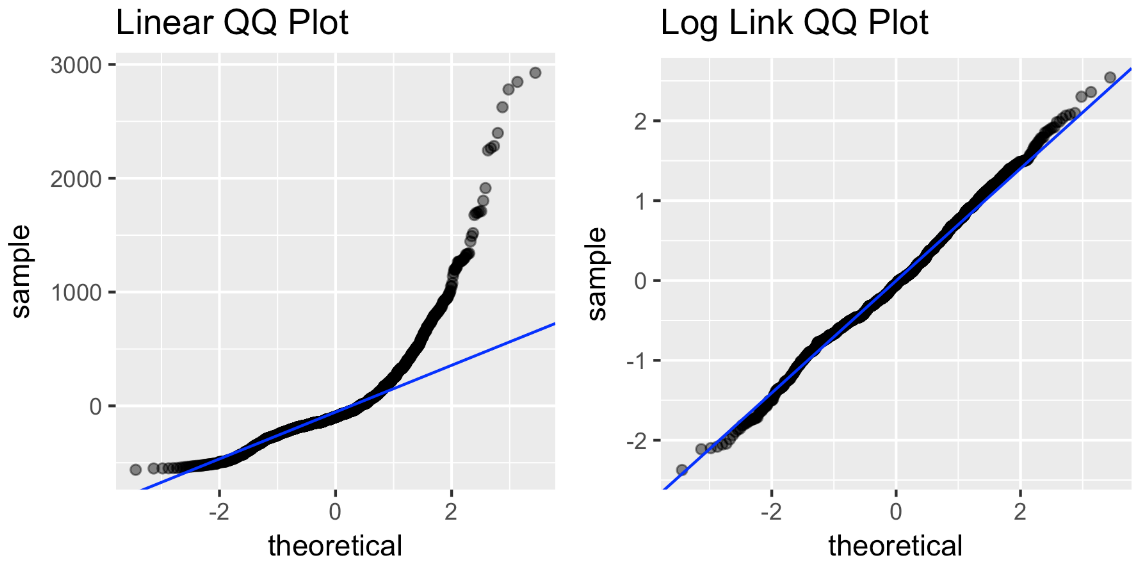

Figure 9: Quantile Quantile Plots.

Lastly, a qualitative analysis of the film corpus is provided to address issues arising from the segmentation of the corpus. Variability in anglicism rates within is most likely due to content changes (characters, settings, or registers). Analyzing the film segments with the highest usage of anglicisms revealed three motives for the use of English: specialized lexicon, stylistic code-switching, and word repetition. *Whisky Romeo Zulu*, a film based on an infamous Buenos Aires plane crash in 1999, had several dialogues that utilized technical vocabulary related to flying.

(1)Yo no te pedí opinión. <Before start check list> dale. Discúlpeme, señor comandante. <Emergency exit lights>

'I didn't ask for your opinion. Before start check list ok. Excuse me, sir commander. Emergency exit lights'

In contrast, in the psychological thriller *El Método* 'The Method', a Spanish-speaking character switches between Spanish, English, and French, to highlight his point about the importance of knowing multiple languages in this interconnected world, code-switching between languages not out of need for technical vocabulary but as a stylistic choice to better prove a point.

(2) <Great Britain is the mother of the greatest empire in the world.> A nosotros nos la suda el imperio ese porque tenemos todo el mercado latinoamericano. Nous aussi avons étroite relation avec les États Unís. Nous sommes tout ce qu'ils ne peuvent jamáis être.

'Great Britain is the mother of the greatest empire in the world. We don't give a damn about the empire because we have the whole Latin American market. We also have close relation with the United States. We are what they can never be.'

The indie film *El Custodio* 'The Minder' contains several segments with high anglicism rates, due solely to the repetition of one or two loanwords; as seen in the example below, the anglicism *tofu* appears five times within the 42 words. Note that while *tofu* comes from Japanese, it is noted in the Spanish Dictionary as also being attributed to English, thus is included in this analysis. More on the choice to include these types of borrowings will be discussed in Chapter 4.

(3) No , era <tofu> y no lo has querido. ¿Cómo era? Como un queso con soja. No, esto es ... Es <tofu>. No, esto no es tofu. Camarero, ¿Qué es el <tofu>? El <tofu> es un queso de soja, totalmente natural. No, entonces, no.

No, it was tofu and you didn't want it. What was it like? Like a soy cheese. No this is… It is tofu. No, this isn't tofu. Waiter, what is tofu? Tofu is a soy cheese, totally natural. No, then, no.

DISCUSSION

Anglicism Identifier

The Anglicism Identifier presented in this chapter achieves an F-score of 79.41% on training data and 74.50% on test data, outperforming several recent models in the literature (Andersen, 2005; Leidig et al., 2014). It is outperformed by Alex (2008), which achieves an F-score of 83.18%; however, the model presented here is less computationally taxing and doesn't suffer from internet search limits, which can present challenges when processing large datasets.

This model is novel in that it addresses the role of NEs, lemmas, and loan phrases. NEs can be particularly problematic for newspaper corpora, often used in anglicism studies, and can lead to an over-identification of loanwords. The NE recognizer used here was a simple capitalization test. While it led to an overall higher F-score for anglicism identification than the Stanford NE Recognizer, there are still errors resulting from NEs and thus future advances in NE recognizers could be applied to this model to improve its performance. Additionally the capitalization test, while being somewhat functional for

English and Spanish, would not work for other languages such as German, where all nouns are capitalized.

One unforeseen challenge for this loanword identifier is the presence of code-switching in the data. While the majority of English insertions are borrowed words or phrases within Spanish text, some films contain code-switched dialogue between English-Spanish or even complete dialogues in English, such as in the film *El Nido vacío* 'The Empty Nest'.

(1) *Hello, I am Leonardo ... Your father. Maybe you remember me?*

These instances of English prove problematic for the model, which defaults to Spanish, meaning the homograph *me* is incorrectly identified as Spanish. Hidden Markov Models (HMMs), utilized in the code-switching models of Solorio & Liu (2008) and Guzman, Serigos, Bullock & Toribio (2016), successfully address this issue by considering the language tag assigned to the previous word(s) when determining the language of the current token. While HMMs may be well suited to code-switched texts, for predominately monolingual texts with sporadic loanword insertions, these models are less helpful as the language tag of the previous word is irrelevant in the case of single word insertions. Testing the code-switching model of Guzman, Serigos, Bullock & Toribio (2016) on the test set developed from NACC and ACTIV-ES reveals how the HMM greatly overestimates the number of anglicisms, i.e. there is a high number of false positives, and thus overall performs poorly with a F-score of 54%. Future work is needed to consider how to combine these two models to handle text with both sporadic foreign word insertions and code-switching. Finally, future work on the classifer is needed to identify adapted loanwords, such as *bloggero* 'blogger' or *crashear* 'to crash'. Possibly using an automated stemmer, a natural language processing tool that reduces inflected or derived words to their word stem, could help identify English-origin tokens modified with Spanish morphology.

Corpus Study

The results of the corpus study clearly demonstrate that anglicisms are not equally distributed across the data. Among the four datasets, La Nación, the newspaper with highest prestige, employs anglicisms at a greater rate and in more articles than the other newspapers, Clarín and Crónica, and the film corpus. The films, representing oral everyday speech, show the lowest rate of anglicisms usage, but also employ instances of code-switching and dialogue in English. These findings support the claim that loanwords function as prestige markers in Argentina, which may be a logical consequence of the mode of contact: the upper socio-economic status group has greater access to outlets where loanwords seem to emerge, such as the media, internet, and second language education. Loanwords, particularly less common ones, may index a speaker as having studied English, traveled abroad or been in contact with media in English. This stands in stark contrast to English and Spanish contact within the US, where Spanish speakers often have a more direct means of accessing English given that the surrounding area is English-speaking, perhaps removing its allure and uniqueness.

The fact that all newspapers employ loanwords at a higher rate than movies suggests that transmission of loanwords may emerge from the top, through the media, and less through direct contact, supported by its lower rates in everyday speech. Additionally, La Nación's exceptionally usage of anglicisms relative to the other two newspapers shows that select media outlets, rather than the media in general, lead the way in terms of English influence. Also contributing to the contrast between the newspapers and film datasets could be topical difference; newspapers may more frequently discusses topics that require anglicisms, for example technology or business, while ignoring topics that do not lend themselves to loanwords. This hypothesis is supported in the fact that much of the difference in anglicism rates between the two less prestigious newspapers and the film corpus disappeared when only data points containing anglicisms were considered. To explore this hypothesis, additional analysis is needed that would involve tagging articles and movie segments for topics to see the patterns.

The present analysis could additionally be improved by finding a more sophisticated way to segment the film corpus. In this study, the film transcripts are divided into 20 equal length segments to approximate the length of articles, but ideally segmentation would be more conceptually sound, such as dividing by scene or grouping each character’s dialogue. This requires a more detailed annotation of the corpus, achieved either through manual means or, ideally, using a natural language processing method that could perform this task automatically.

Chapter 3: Distributional Semantics in Loanword Research[6]

As seen in the literature review presented in the Introduction and second chapter of this dissertation, loanwords have long been of interest in historical, socio, and contact linguistics. However, in spite of the growing body of research on this contact feature, the semantic properties of individual loanwords remain largely understudied. Attention to the semantics of loanwords has been hindered by the challenges associated with the empirical testing of aspects of word meaning, such as how to quantifiably compare them. This chapter attempts to offer a novel approach to this underserved topic. As the availability of 'big' data sourced from internet text (e.g., Twitter, blogs, online newspapers) has bolstered empirical research on loanwords, these large datasets render possible computational, quantitative techniques for studying the semantics of loanwords. The study of loanword semantics is important in that loanwords can alter recipient-language semantic systems (Bookless, 1982), and an understanding of the semantic properties of loanwords can shed light on motivations for their adoption. The semantics of another contact feature, bilingual code-switching, have been analyzed by Backus (2001), who proposes that "embedded language elements in code-switching have a high degree of semantic specificity" (128). While Backus's Specificity Hypothesis was put forward to account for dynamic bilingual alternations, the present work seeks to extend its scope to foreign insertions in primarily 'monolingual' texts.

[6] Serigos, J. (2016). Using distributional semantics in loanword research: A concept-based approach to quantifying semantic specificity of Anglicisms in Spanish. *International Journal of Bilingualism*.

In pursuing this aim, the present chapter sets forth innovative applications of corpus data and computational tools in testing whether anglicisms extracted from a large corpus of Argentine Spanish have a high degree of semantic specificity. This research implements a concept-based approach (Zenner et al., 2012), referencing the recipient language's semantic equivalent for each anglicism to create a consistent and appropriate unit of comparison. For example, the specificity of the anglicism *manager* is compared to the specificity of the native Spanish equivalent *gerente*. In addition, the work presents a definition and operationalization of specificity, informed by distributional semantics. Lastly, the study offers a novel approach to computing specificity, utilizing an entropy measure of the target word's environment, the assumption being that more specific nouns have less variety in their surrounding context. This approach in its current state is limited to single lexical insertions and requires a large dataset; however, further research within this area will hopefully expand its reach to also address issues central to code-switching. Extending Backus's Specificity Hypothesis to borrowing, we predict that loanwords are more specific than their native equivalents.

The chapter proceeds as follows. First, the extant research on the semantics of loanwords will be reviewed, followed by a brief discussion of distributional semantics and its relevance for the present investigation of loanword specificity. At the center of the chapter are the corpus description and methods; the 24 million word corpus of Argentine Spanish is described in detail along with the three stages of corpus processing: loanword identification, selection of semantic equivalents, and measurement of specificity. Next, the loanwords identified in the newspaper corpus are subjected to both quantitative and qualitative analyses. Finally, the results are discussed to draw conclusions on the role of semantic specificity in loanword adoption and success.

LITERATURE REVIEW

Semantic analysis of loanwords

Of the numerous studies on loanwords, few undertake semantic analyses. Though small in number, those studies that do attend to loanword semantics pursue a variety of approaches to the topic. One approach is to classify loanwords based on the semantic relationship between loanwords and the existing lexicon of the recipient language. This relationship, mentioned as early as Weinreich (1953), can be used to create two categories of loanwords: (i) loanwords with no existing semantic equivalent in the recipient language, e.g., the English loanword *blog* in Spanish, and (ii) loanwords that do have an existing semantic equivalent in the recipient language, e.g., the English loanword *casting* in Spanish, which is synonymous with *reparto*. These contrasting categories have been described using various terms: necessary and luxury (cf. Onysko & Winter-Froemel, 2011), unique and synonymic (Bookless, 1982), core and cultural (Myers-Scotton, 2002), catachrestic and non-catachrestic (Onysko & Winter-Froemel, 2011). Nonetheless, the basic premise remains the same. In one of the first studies to focus solely on this distinction, Bookless concludes that unique loanwords possess high referential value and cause minimum semantic rearranging in the receptor language because they do not displace other words. In contrast, loanwords with existing equivalents contain more stylistic than referential value and often demand a reshuffling of the recipient-language semantic system; this disruption, he argues, results in greater confusion on the part of speakers, as connotations and shifting meanings must be considered in choosing between the loan and the native word.

One limitation of Bookless' work is its reliance on isolated examples of English loanwords in Spanish selected by the author, i.e. the work presents a conceptual analysis of the unique vs. synonymic taxonomy, rather than an empirical study. More recently, Onysko & Winter-Froemel (2011) adopt an empirical approach in their study of the distinction between loanwords with semantic equivalents and those without, drawing on a 5 million-word corpus of the German newsmagazine Der Spiegel. The authors posit that this distinction can be used to explain the pragmatic feature of markedness: loanwords with an existing semantic equivalent are more marked than those without one. Their analysis revealed that about a third of the loanword tokens enter the German language as semantic innovations, while the remaining two thirds compete with existing German equivalents. Though Onysko & Winter-Froemel conclude that this classification is viable, they caution that the distinction is not always clear-cut, due to complications such as polysemy of the loanword or its change in meaning over time. Thus, this categorization may only be possible from a synchronic perspective and is not strictly an either/or decision.

Another corpus-based study of loanwords, published by Zenner et al. (2012), also makes use of this semantic distinction as a factor in its model to predict the success of English loanwords appearing in two Dutch newspaper corpora, together totaling over 1.6 billion words. For their purposes, the authors distinguish the two categories by the status of the loanwords at their time of adoption. Necessary loanwords, as they refer to them, are introduced to fill a lexical gap, and thus do not initially compete with an existing equivalent, although the authors note that later equivalents may be invented to fulfill this role (e.g. *voetbalvandaal* for *hooligan*). In contrast, luxury loanwords already have an existing equivalent when they are introduced into a recipient language. A comparison of the two categories indicated that necessary loanwords had slightly higher chances of success than loanwords with native equivalents.

Other studies focus on specific loanword and native equivalent pairs, comparing the two. Peterson and Vaattovaara (2014) have analyzed the use of the Finnish native politeness marker *kiitos* in comparison with the English loanword *pliis*, finding that the two share little overlap as they are used differently "grammatically, pragmatically and in terms of social distinctiveness" (247). Andersen (2014) explores the use of English borrowings in Norwegian, specifically their pragmatic function in the recipient language as compared to that in the source language, revealing two scenarios: functional stability or functional adaption, such as broadening, narrowing, or shift. Finally, Hornikx, Meurs, & Boer (2010) explore the use of loanwords relative to their equivalents to understand if Dutch participants' comprehension of loanwords affects their attitudes towards them, specifically their use in Dutch advertising.

Another approach to the semantics of loanwords is to consider semantic domains, also called lexical fields, to classify loanwords into categories, as it is often posited that loanwords arise more frequently when discussing topics relevant to the source language (Aaron, 2015; Zenner et al., 2012; Winter-Froemel & Onysko, 2012). The analysis of semantic domains is seldom the sole focus of a study, but rather a supplementary component to a larger question of loanword adaptation processes. In analyzing anglicisms in Dutch, Zenner et al. (2012) found sports and information & technology to be the two categories in which loanword incorporation was most successful, in comparison to the other three semantic categories they coded for: business, social, and deviance. In a similar study, Onysko and Winter-Froemel (2011) found the most common semantic categories for anglicisms in German to be information and computer technology, business, music and television, clothing, and sports. Different findings are offered by Aaron (2015), whose work on Spanish in New Mexico showed that of the 17 categories coded for, kinship terms made up the most frequently borrowed items. The divergence of Aaron's findings may be explained by differences in genres (newspaper text versus colloquial speech) and/or by differences in the metrics to assess the success of a semantic domain.

For each semantic domain, Aaron totaled all uses of nouns pertaining to that domain and determined what percentage of those nouns were loanwords. For instance, of the 238 kinship terms used in the corpus (e.g. *abuela, papá, sister, grandpa*), 136 were English-origin insertions, around 57%. In contrast, Onysko & Winter-Froemel (2011) and Zenner et al. (2012) gathered only anglicisms, determined the semantic category of each, and counted which categories held the highest number of types relative to the other categories. The differences in findings may also suggest that the semantic categories that are most frequently borrowed reflect the type of contact (intense vs. weak, see Zenner et al. (2014)) as well as the cultural value of the source language within the given community (see Bullock, Serigos, & Toribio, 2015).

Yet another approach to the study of the semantics of loanwords is to divide them into those with one word sense and those with multiple senses. For example, the anglicism *Chip*, as used in German, has multiple meanings, referring to a potato-based snack, a technological component, or the currency used in casinos, while the anglicism *T-shirt* has only one meaning. This categorization has been studied in relation to its potential impact on the success of a loanword. Chesley & Baayen (2010:1353) posit that, in possessing an "increased range of denotata," polysemous loanwords can be used with more frequency and therefore have an increased likelihood of successful entrenchment. They find this prediction to hold true for loanwords in culturally restricted contexts, i.e. contexts that refer to the culture associated with the source language of the loanword. However, the same pattern does not hold true for loanwords in culturally unrestricted contexts, i.e. those that don't necessarily refer back to the source language culture. Hlavac (2006) finds that the distinction between uni- and polyfunctional words impact their frequency of use when comparing English borrowings to their Croatian counterparts;

polyfunctional words, whether borrowed or native, increase in frequency compared to their alternative.

One of the most thorough discussions of the semantics of loanwords is provided by Winter-Froemel (2013). Following a usage-based approach, she tackles the issue of multiple forms of the same loanword (e.g. *people* & *pipole* in French) and semantic change of loanwords from the source to the recipient language (e.g. in Spanish *sombrero* = 'hat', yet in English *sombrero* = 'type of hat with a wide brim, associated with Mexico'). Approached from the level of speaker-hearer, her analysis highlights the hearer's role in introducing new interpretations of loanwords, specifically resulting in two types of semantic shifts: specialization and metonymy.

For the most part, previous approaches to the semantics of loanwords share two features: they employ manual and non-numerical means to categorize loanwords into discrete categories and most compare the differing semantic properties among loanwords, rather than compare the semantic properties of loanwords to those of receptor-language equivalents. This chapter offers a new approach, classifying loanwords using a numeric measurement that contrasts them with native words, while additionally addressing an understudied aspect of loanword semantics: specificity.

The notion of specificity within the literature of contact linguistics was examined by Backus (2001) with reference to Turkish-Dutch bilingual speech. He observed that first-generation bilinguals inserted Dutch words into Turkish clauses and that these lexical switches typically came from specific semantic domains. Drawing on Myers-Scotton's (1997) Matrix Language Frame model, Backus formalized this observation in the Specificity Hypothesis: "Embedded language elements in code-switching have a high degree of semantic specificity" (2001, p. 128). The hypothesis encapsulates the idea that insertion is facilitated for words that have highly specific meaning and whose cross-linguistic equivalents, where they exist, conjure up different connotations (see also Myers-Scotton & Jake, 1995). Support is offered in Anderson and Toribio's (2007) study of Spanish-English mixing, where insertion of lexical items denoting specialized concepts within the Little Red Riding Hood fairytale (e.g. *hunter*) were judged to be less marked/more felicitous than insertion of core nouns (e.g. *bed*) in the same story.

Left unspecified in these studies is the unit of comparison when testing the specificity of embedded items—a loanword is more highly specific than what? The present study adopts a concept-based approach, using the semantic equivalents of loanwords as a unit of comparison. Also left unresolved in previous studies is an empirically testable definition of specificity. Backus offers the replacement test, i.e. highly specific words are hard to replace with a synonym, while more general words are easy to replace with something more specific. For example, the general term *tree* is easily replaced by a more specific term such as *oak* or *dogwood*; while those two terms are not readily replaceable with a singular specific term. However, this definition of specificity presents issues that must be surmounted. For instance, due to its subjectivity, the results may not always prove reliable or replicable. In addition, the definition only divides words into two categories—specific, i.e. cannot be replaced, or non-specific, i.e. can be replaced— leaving many word pairings uncomparable. Thus, while this definition may address one aspect of specificity, it leaves information uncaptured. In order to find a more robust definition of specificity and a means by which to quantify it, this study turns to the field of distributional semantics, explored further in the following section.

Distributional semantics

Distributional semantics has roots in both cognitive science and computational linguistics (Lenci, 2008). The cognitive perspective establishes a framework for understanding word meaning, which this study will extend to define word specificity. The computational perspective provides methods by which to capture and quantify word meaning on a large scale, which this study will adapt to measure word specificity.

Before defining word specificity, it is necessary to understand how distributional semantics approaches word meaning in general. Within this field, a word’s meaning can be understood through its surrounding context. This relationship between word meaning and its context is often summed up by the quote from Firth (1957:11), "You shall know a word by the company it keeps", i.e. a word's meaning can be derived from its surrounding context. To exemplify this point, consider sentence (1) below, presented in Erk (2015).

1. *On our last evening, the boatman killed an* ***alligator*** *as it crawled past our campfire to go hunting in the reeds beyond.*

The context surrounding the word *alligator* provides readers unfamiliar with this word with several clues to its meaning; learning that an alligator crawls and can be killed, readers may deduce it is an animal. Learning that an alligator goes hunting, readers may deduce it is carnivorous. Thus, from the context alone, a characterization of the meaning begins to form. That definition may be further refined as readers come across additional contexts in which the word alligator occurs and begin to build a profile of the word’s meaning from its distribution.

In addition to meaning, word contexts can be used to provide information as to how specific and/or general a word is. Specificity has been measured using several techniques. One technique considers the distribution of adjectives modifying the word in question; the idea is that specific nouns are rarely modified in text, while general nouns are frequently modified (Caraballo & Charniak, 1999; Ryu & Choi, 2006). In a similar vein, specificity has been represented by the distribution of verbs that accompany the word in question; this approach makes the assumption that specific words co-occur with a narrow set of verbs while general terms are associated with multiple verbs (Cimiano, Hotho, & Staab, 2005). Both of these approaches require a syntactic parsing of the corpora to identify adjectives and verbs. The approach chosen for this study does not require any POS parsing and is based on the lexical diversity of the surrounding context. The idea here is that the more specific a word is, the less variety it will have in its surrounding context, *i.e.,* it will co-occur with a narrower set of words.

For example, consider the words *writer* and *novelist*, both from the same semantic domain with different levels of specificity. To compare the contexts of these words, I collected data from the British National Corpus (BNC); a quick search of the database revealed that all nouns within a 3-word window of *novelist* are found within a 3-word window of *writer* as well. However, over 50% of the nouns within a 3-word window of *writer* do not appear with *novelist*. A few examples of these contexts from the BNC are presented in Table 5, where we see that the two words share some contexts (*poet, wife,* and *career*), as to be expected given that they pertain to the same semantic domain. However, *novelist* does not co-occur with *television*, *business*, and *staff* as the word *writer* does. Thus, the variability of a word's surrounding context can provide insight as to how specific the word is.

Table 8: Examples of *Writer* and *Poet* in the British National Corpus.

Writer	***Poet***
*the first Caribbean **writer** and **poet** to receive the honour*	*this man of diverse talent -- **poet**, **novelist**, song-writer, performer*
*decided on a **career** as a **writer***	*a successful **career** as a **novelist***
*An elderly **lady**, a **writer** of crime-stories*	*a **lady** **novelist** like Jane Austen*
*a well-known **television** **writer***	-
*a specialist **writer** on **business** affairs*	-
*the most entertaining **writer** on your **staff***	-

Closely related to specificity is the semantic relationship hypernymy, which reflects the status between a superordinate term and a subordinate term within a taxonomy. In addition to the subordinate term having greater specificity than its superordinate term, the two terms share a semantic class. As in the *writer/novelist* example, *writer* (the hypernym) ranks above *novelist* (the hyponym) within the taxonomical hierarchy. Hypernymy, though closely related to specificity, captures a relationship between two words, rather than a property of a word itself, and implies an additional layer of information, that of the shared semantic category within a taxonomical hierarchy.

In identifying hypernymy distributionally, much of the work is based on a distributional similarity hypothesis outlined by Weeds, Weir, & McCarthy (2004) and further refined by Geffet and Dagan (2005) as the Distributional Inclusion Hypothesis, which states that the contexts in which a hypernym appears should be a superset over the contexts of its hyponym. Weeds et al. (2004) find a strong link between distributional generality, relative frequency and semantic generality. However, Roller, Erk, & Boleda (2014) find the Distributional Inclusion Hypothesis only holds if inclusion is applied to the set of relevant dimensions, those that were deemed necessary by the classifier.

For the study presented in this chapter, specificity is defined distributionally, yet does not follow the distributional inclusion approach because, while loanwords and native equivalents are clearly semantically related, we do not want to assume the hierarchical relationship of hypernymy, but rather more broadly explore loanword's relative levels of specificity. Thus, taking a non-inclusive distributional perspective, this study defines word specificity relatively: Words used within a narrow set of contexts are more specific than words used across varied contexts.

This distributional understanding of word meaning and word specificity may be quantified, using methods from computational linguistics. These methods for capturing and quantifying word meaning are based on their distribution in large bodies of text (Erk, 2012; Turney & Pantel, 2010). They rely on big data to construct profiles of words that reflect their meaning and relationship to other words. In practical terms, distributional models utilize large corpora to create high-dimensional context vectors. The context vector for a given word contains the words with which it co-occurs, along with the number of co-occurrences. For example, the context vector for writer, collected from the BNC, includes the following words and counts (among many others): actor:1963, fiction:1927, pen:627. Context vectors have been utilized for information retrieval (Manning, Raghavan, & Schütze, 2008), word sense disambiguation (McCarthy, 2009), and word specificity (Caraballo & Charniak, 1999), to name only a few of their current applications.

It is the application of vectors to quantifying word specificity that is most relevant to this study. Caraballo & Charniak (1999) create a numerical score for word specificity, which allows them to rank words hierarchically, using entropy calculated from context vectors. The details of the entropy calculation and context vectors will be discussed in detail in the following section. Within the field of psycholinguistics, McDonald & Shillcock (2001) use the same method of entropy calculated from context vectors to quantify the concept they introduce: Contextual Distinctiveness. Their definition of Contextual Distinctiveness, which is based on the informativeness a word provides about its contexts of use, very much relates to the notion of specificity used here and in Caraballo & Charniak (1999). They show that Contextual Distinctiveness is a better predictor of lexical decision latencies than word frequency.

THE PRESENT STUDY: QUESTIONS AND METHODS

The present study pursues a semantic analysis of loanwords, examining their specificity relative to that of their native-language equivalents. More specifically, the work extends Backus's Specificity Hypothesis, developed for lexical insertions in oral bilingual code-switching, to the analysis of donor-language loans in 'monolingual' recipient-language contexts. Thus, the study aims to address theoretical questions and methodological concerns: Are loanwords more semantically specific than their receptor language alternatives? What does it mean for a word to be semantically specific? How can specificity be empirically measured?

In responding to these questions, the study utilizes a concept-based approach from cognitive linguistics, a context-based understanding of word meaning from distributional semantics, and distributional models from computational linguistics. This section describes the corpus used for the analysis, defines the scope of loanwords and the processes applied in extracting them from the corpus, and presents the definition of specificity used in this study along with the model and parameters used to measure it.

Corpus

As mentioned in Chapter 2 of this dissertation, Argentine Spanish is particularly amenable to loanword research because it demonstrates a large influx of anglicisms that often do not undergo orthographic alteration, making them easier to identify in written texts. The NACC corpus created for the study in Chapter 2 was expanded from the original 1 million words to 24 million words by the downloading of articles from the daily archives for the one-year period from June 2013 through May 2014. In gathering additional data, only the newspaper La Nación was used because it was the only newspaper of the three that made its archives easily available through its website. As in the NACC corpus, articles from all sections, such as Politics, Economy, International, Opinion, Sports, Technology, were included to ensure that a broad range of topics were covered. Issues with downloading and text encoding caused a few days to be eliminated, thus not all days from this time period are represented. The corpus was subsequently parsed for part-of-speech (POS) and lemma using the open-source probabilistic tagger, TreeTagger (Schmid, 1994).

Loanwords

Definition

The same definition used for the study in Chapter 2 was used for the current study: loanwords are words whose phonemic shape and meaning have been imported into a recipient language with no morphemic substitution (Haugen, 1950).

Unlike the study in Chapter 2, which included loanwords from all parts of speech, this study includes only loanwords that function as nouns because – as this study is concept-based, meaning each loanword will be assigned a semantic equivalent – nouns better lend themselves to having equivalents and are universally the most widely borrowed part of speech (Haspelmath, 2008). However, this specificity measure and other distributional models may be applied to other open-class parts of speech, including verbs, adjectives, and adverbs.

Identification

To find all tokens of English origin in the Argentine corpus, an automated English identifier was developed. The system designed for this study[7] makes use of a lookup method, similar to Alex (2008). The two stages that comprise this system —(1) collecting all tokens that are not recognized as Spanish and (2) checking the non-Spanish tokens (which include borrowings from various languages, proper names, onomatopoeia, etc.) to see if they are English — are further explained below.

To identify words that are not Spanish, the algorithm makes use of two special tags generated by TreeTagger. The first tag *palabra extranjera* (PE) 'foreign word' is a POS tag that TreeTagger outputs when it recognizes the token as foreign, along with the corresponding lemma tag. However, most foreign items, including many anglicisms, are not recognized by TreeTagger. When TreeTagger fails to recognize a word, it outputs the lemma tag <unknown>. Table 6 below illustrates a sample TreeTagger output for the sentence *pidió input de sus managers* 'he/she asked for input from his/her managers'. The Spanish token *pidieron* 'they asked for' receives the POS tag verbo lexical finito (VLfin) 'finite lexical verb' and the lemma *pedir* 'to ask for', while the two foreign items *input* and *managers* receive the POS tag PE and the lemma tag <unknown> respectively.

[7] The model used in this current study served as the base upon which the model presented in Chapter 2 of this dissertation was built and further refined.

Table 9: Sample Annotation from TreeTagger.

Token	POS Tag	Lemma
pidieron	VLfin	pedir
input	PE	input
de	PREP	de
sus	PPO	suyo
managers	NC	<unknown>

With this first stage of the algorithm, all tokens with the POS tag PE or the lemma tag <unknown> are collected. The one exception is that tokens beginning with a capital letter are not included in order to avoid proper nouns, such as *General Motors*. Though this process also excludes sentence-initial nouns, the exclusion is not considered consequential because Spanish typically does not permit sentences beginning with bare nouns.

Additionally, in order to preserve the integrity of loan phrases, such as *think tank*, anglicisms are collected and evaluated in chunks, using the IOB (Inside, Outside Beginning) tagging technique commonly used in named Entity Recognition (Jurafsky & Martin, 2008). The B- tag is given to anglicisms at the beginning of an anglicism phrase, i.e. an English token with a Spanish word preceding it. The I- tag is given to any anglicisms on the inside of a loan phrase, i.e. an anglicism following a B-tag or I-tag. The O- tag indicates that the token is outside of the loan phrase, i.e. it is Spanish. Consider the example, *El/O manager/I del/O think/B tank/I es/O alemán/O* 'The manager of the think tank is German'. The loanword *manager* and the loan phrase *think tank* are identified as two distinct chunks, the former a one-word chunk and the latter a two-word chunk. These loan phrases are included in the analysis and treated just as any other single-word token. For example, the phrase *think tank* appears 69 times in the corpus, so it is recorded as one unit with a count of 69, rather than as two separate single-word entries: *think* with a count of 69 and *tank* with a count of 69.

The tokens returned by this automated process include the target items, i.e. anglicisms (e.g. *input*, *managers*), but also comprise borrowings from other languages (e.g., *spaghetti* from Italian), misspellings (e.g. *pidio* instead of *pidió* 'he/she asked for'), and Spanish words that are not recognized by TreeTagger due to sparse training data (e.g., *canonización* 'canonization', *ancestral* 'ancestral'). Thus, in the second stage, to remove non-English tokens from this list, the tokens are lemmatized using an English lemmatizer from the Natural Language Toolkit (Bird, Klein & Loper, 2009). The lemmas are checked for membership in the English dictionary from the UNIX operating system. Additionally, it is checked that they are not present in the Spanish dictionary, Diccionario de la Real Academia Española, so homographs between the two languages that are not recognized by TreeTagger, such as *ancestral*, are not marked as English.

The final stage involves manually inspecting all remaining tokens for any words erroneously identified as anglicisms, such as loanwords from other languages that were included in the English dictionary (e.g., *cadenza*, *leitmotiv*, *burka*) and homographs between the two languages that are not present in the Spanish dictionary (e.g., *postdoctoral*). The major challenge for the automatic identification is the fact that UNIX's English dictionary is extremely robust in comparison to the Spanish lexicon used by TreeTagger and Diccionario de la Real Academia Española; it encompasses numerous foreignisms and homographs with Spanish that are absent from Spanish sources, hence the need for the subsequent manual inspection.

After anglicisms in the corpus are identified, those that comply with two criteria are selected for analysis: first, loanwords must function as nouns; second, they must appear 50 or more times throughout the corpus, as the ensuing computational analysis uses word vectors that require higher counts for improved accuracy.

Selection of semantic equivalents

The working definition for semantic equivalents is drawn from the discussion of near-synonyms versus true synonyms as presented in Zenner et al. (2012), in which the researchers sought to distinguish "those near-synonyms which are maximally equivalent with a given English ...noun" (760) from true synonyms. True synonyms – "two words [that] can replace each other in any given context without changing the propositional content of the sentence they are used in" (see Edmonds & Hirst, 2002: 107) – are actually quite rare and not a feasible concept for natural languages.

This current study is interested in the relationship formed between loanwords and the existing semantic system, specifically loanwords relative to their closest native alternative, be they true synonyms or near-synonyms. The aim is to capture the closest alternative available to a speaker, even if that alternative functions more as a hypernym, such as *apariencia* 'appearance' for *look*, rather than a true synonym, such as *celebridad* ‘celebrity’ and *celebrity*. The same reasoning applies to those loanwords whose semantic equivalents are polysemous (e.g. the loanword *team* is equivalent to the Spanish *equipo,* which, in addition to meaning 'team', also means 'equipment'). Limiting the equivalents to true synonyms would exclude the majority of tokens from this study and thus unnecessarily narrow the scope of this study.

With this goal of selecting near-synonyms, a list of potential equivalents was gathered using the online translator WordReference and the Spanish dictionary Diccionario de la lengua española. Together the researcher and a native Argentine Spanish speaker selected the most viable semantic equivalent from the list of potentials.

Only one semantic equivalent was chosen for each loanword, as the goal of the study was to identify the one closest equivalent rather than to obtain a list of potentially related words. This choice had the benefit of maintaining a binary setup for the data, which allowed for a straightforward comparison in the statistical analysis.

In order to ensure appropriate and sufficient data for the quantitative analysis, loanwords require a semantic equivalent that also appears 50 or more times. Thus some loanwords are excluded from the quantitative analysis because the semantic equivalent does not appear in the corpus or appears with low counts, as is the case for *freezer* and its equivalent *congelador*. Other loanwords, such as *blog*, are excluded from the quantitative analysis because they lack a semantic equivalent. However, all of the gathered loanwords are discussed in the qualitative analysis. Several native equivalents function as multiple parts of speech, such as *puesto* noun ‘stand’ or past participle ‘placed’. To avoid non-relevant uses, only equivalents tagged with the POS tag *noun* are included in the analysis.

Computational Model: Measuring Specificity

The final step of data processing is measuring the semantic specificity of the loanwords and their native equivalents. Loanwords and equivalents will be referred to here as target words. Recall that for this study, specificity is defined distributionally; words used within a narrow set of contexts are more specific than words used across varied contexts, as exemplified by the BNC corpus results for *writer* and *novelist* discussed in the literature review. While the terms 'narrow' and 'varied' may appear subjective, they become quantifiable, replicable and thus objective by calculating the entropy of a word's distribution. Entropy, as used within information theory, measures the complexity or disorder of information within a system. Given a random variable X ranging over the set of χ and with a probability function of p(χ), the entropy of the variable X is calculated as follows:

$$H(X) = -\sum p(x) \log_2 p(x)$$

Conceptually, entropy can be understood as "a lower bound on the number of bits it would take to encode a certain decision or piece of information in the optimal coding scheme" (Jurafsky & Martin, 2016). Applied to word specificity, the entropy of a word increases as the distribution of a word becomes more complex, *i.e.,* varied.

The entropy measure proposed in this study is a modified version of the one presented in Caraballo & Charniak (1999). Caraballo & Charniak (1999) seek to determine the most accurate computational method for ranking words according to specificity. Drawing on a corpus of 15 million words, they evaluated nine measures on their ability to correctly reproduce three hierarchies of noun hypernyms[8] (see Figure 10 below as an example) from an unordered set of words, utilizing the corpus alone as a data set.

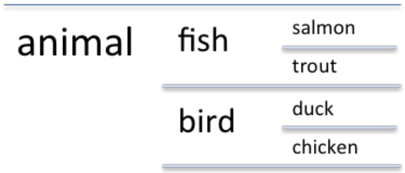

Figure 10: Sample Hierarchy.

The first four measures are based on the probability that the target word is modified (by a prenominal adjective, by a verb, by another noun, and by any of the three). The next four measures calculate entropy, that of the rightmost prenominal modifier and that of all the words occurring within a 2-word, 10-word and 50-word window of the target. The final measure is the frequency of the target word. The most

[8] Selected from WordNet.

accurate of the measures tested were: (i) entropy of the rightmost modifier, (ii) frequency of the word, and (iii) entropy of the 50-word window of the target, each of which performed with over 80% accuracy in reconstructing the hypernym hierarchy.

As they achieved equally high accuracy, the three measures were each considered to be used as the specificity measure in this study. The entropy of the rightmost modifier was discarded because it requires syntactic parsing. The frequency measurement was discarded because, by ignoring the surrounding context, it does not measure specificity in accordance with the distributional definition proposed in this study. From a linguist's perspective, though a word's frequency may strongly correlate with the specificity of its meaning – for example in the newspaper corpus used for this study, the hypernym *tela* 'fabric' occurs over three times as often and has a higher entropy score than its hyponym *lana* 'wool' – the two measures are conceptually distinct. As this chapter aims to quantify and analyze specificity, defined distributionally, frequency is not an appropriate measure. Thus, the entropy of the 50-word window, which considers the surrounding context and does not require syntactic parsing, was selected as the method to calculate specificity. The entropy measure for a given target word is calculated as:

$$H_n\ (target)\ =\ -\sum_{Words}[\mathrm{P}(context\ word|target)\ *\log_2 \mathrm{P}(context\ word|target)]$$

where Hn(*target*) is the entropy score, P(*context word|target*) is the probability that *context word* will appear within a n-word window of *target* and Words is the set of context words that appear in a target's context window. Probabilities are calculated for each target word using its context vector. The hypothesis is that nouns with greater specificity have less variety in their surrounding context, resulting in lower entropy scores.

This entropy calculation comprises several parameters, including context window size, context type, and dimensionality reduction. The parameter settings in this model were as follows: only nouns, verbs, adjectives and adverbs with a count of 30 or more, appearing within a 50-word window of the target, were included in the context vectors. Function words and light verbs – verbs that contribute little to no semantic content on their own – were excluded from the context vectors. These types of words, commonly referred to as stop words, are often ignored in distributional models because they offer little to no semantic information due to their ubiquity and lack of referential meaning. The function words were identified via their part of speech tags provided by TreeTagger, *i.e.*, tokens not tagged as nouns, verbs, adjectives or adverbs were excluded. The light verbs were identified by checking their lemmatized form, also provided by TreeTagger, with a list constructed by the researcher. The light verb list comprises the following verbs: *haber* 'to have', *hacer* 'to do/make', *estar* 'to be', *ser* 'to be', *tener* 'to have'.

One problem with the entropy measure as presented in Caraballo & Charniak (1999) is its sensitivity to frequency. For example, if we wish to compare word A, which occurs 100 times, to word B, which occurs 50 times, this entropy measure provides A double the opportunity to show variety in its surrounding context; thus frequency affects the resulting score. This fact did not pose a problem for Caraballo & Charniak (1999)'s task-oriented study. However, since the current study aims to measure the linguistic concept of specificity as separate from frequency, the overlap proves problematic.

To remove the effect of frequency from the entropy measure, a technique to artificially keep the frequency the same for all target words was added: bootstrapping. Bootstrapping is a statistical technique in which a parameter of a population is estimated by drawing random samples with replacement from the data set available (see Gries (2006) for another application of bootstrapping in a corpus study). For each target word, a sample of 50 occurrences, along with their surrounding contexts, was drawn with replacement from the corpus to calculate an entropy score, thus holding the frequency constant at 50 for all target words. This process was repeated 1000 times for each target word, resulting in a set of entropy scores. The two parameters – sample size: 50 and number of samples: 1000 per target word – were chosen because 50 was the minimum requirement for target words and 1000 samples ensured convergence across the estimate scores. Convergence was checked graphically, revealing that the value of the estimate became stable after approximately 700 iterations, thus 1000 is well beyond the point of convergence.

To check the accuracy of this resulting entropy model with bootstrapping on the 24 million-word corpus compiled for this study (small by computational linguistic standards, where 1 billion-word corpora are common), a test set of 10 word pairs, similar to that used in Caraballo & Charniak, was selected. The word pairs comprise one general term and one hyponym, following the definition: a Y is a hyponym of X if a native speaker accepts the sentence "Y is a kind of X". These word pairs appear in Table 7 below, along with the frequency of each word in the corpus and the entropy score. A paired T test was applied to the entropy scores of the general terms compared to their hyponyms. The results show that the general terms receive significantly higher entropy scores than their hyponym counterpart ($T9= 4.25$, $p < 0.01$). Thus, the entropy model was considered sufficiently accurate to measure specificity.

Table 10: General Term and Hyponym Entropy Scores.

General Term	Frequency	Entropy Score	Hyponym	Frequency	Entropy Score
animal 'animal'	1334	178	gato 'cat'	430	163
edificio 'building'	3444	186	casa 'house'	14489	184
tela 'fabric'	370	164	lana 'wool'	102	157
órgano 'organ'	628	184	pulmón 'lung'	162	170
bebida 'drink'	629	176	té 'tea'	257	162
escritor 'writer'	1880	178	novelista 'novelist'	214	146
comida 'food'	1068	180	ensalada 'salad'	102	156
planta 'plant'	2347	183	árbol 'tree'	768	179
médico 'doctor'	4306	184	cirujano 'surgeon'	124	163
mueble 'furniture'	328	176	silla 'chair'	735	174

This entropy model, trained on the 24 million-word corpus of Argentine newspaper articles, is used to measure the 30 loanword and equivalent pairs to empirically test the Specificity Hypothesis.

Results and Discussion

The English identifier model found 70 anglicisms that function as nouns and appear 50 times or more. This section will present the quantitative analysis of 30 anglicisms, those that have valid semantic equivalents, using entropy to measure their specificity, as well as a qualitative analysis of all 70 anglicisms identified in the corpus.

Quantitative

The 30 loanwords and their semantic equivalents are analyzed using an entropy measure. The entropy measure for each word is calculated by considering the probabilities of the surrounding context appearing with each of the target words. Thus, a word that appears in a greater variety of contexts receives a higher entropy score, indicating low specificity. The average entropy score for each loanword and its native equivalent[9] appears in Figure 1 below; the dotted lines represent the loanword-equivalent pairs in which the entropy score of a loanword is lower than its equivalent, *i.e.*, pairs that support the proposed Specificity Hypothesis. The dashed lines represent those pairings that show the opposite trend, *i.e.*, pairs in which the loanword had a higher entropy score than its equivalent. The solid bold line represents the general trend, calculated as the mean score of loanwords and equivalents.

[9] The table of these scores is located in the Appendix.

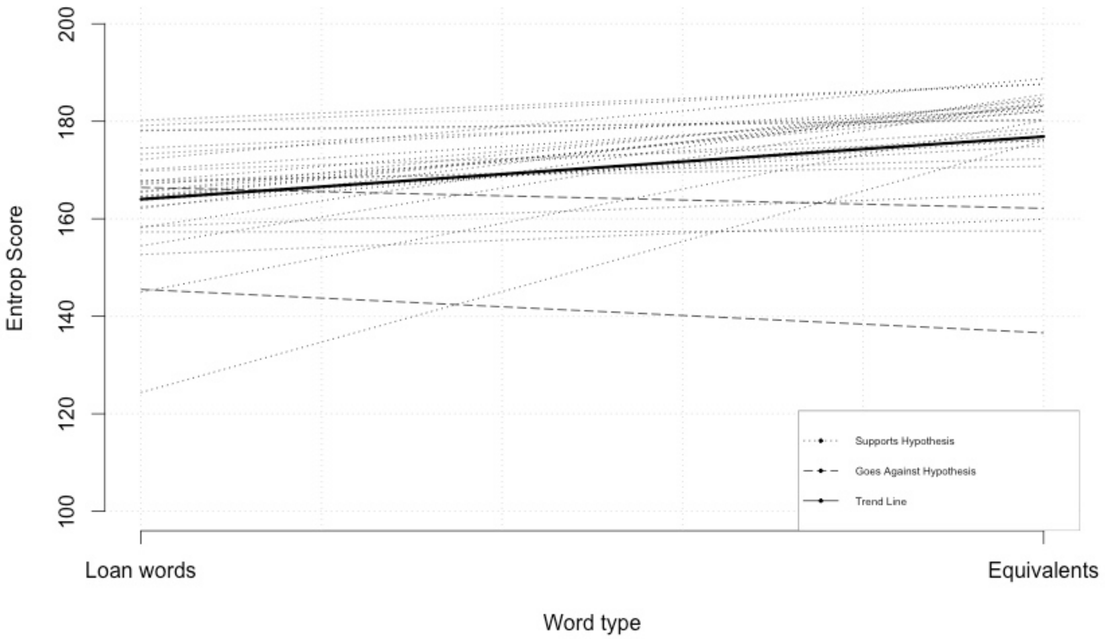

Figure 11: Entropy Scores of Loanwords As Compared To Their Equivalents.

Visually, it is clear that the dotted lines far outnumber the dashed lines and that the bolded trend line shows an increase in entropy score when moving from loanwords to native equivalents. To check the statistical significance of the relationship between Word Type (Loanword or Equivalent) and Entropy Score, the lme4 package (Bates, Mächler, Bolker & Walker, 2014) in R was used to conduct a linear mixed model regression. Word Type was entered as a fixed effect. As random effects, there were intercepts and random slopes for the variable Concept. The variable Concept has 30 levels, one for each loanword/equivalent pair. Each loanword and equivalent has 1000 observations, resulting from the bootstrapping technique applied to mitigate the effects of frequency on entropy. Results from a likelihood ratio test ($\chi 2(1)=23.44$, $p=1.29e-06$) of the full model with the fixed effect Word Type against the model without this effect reveal that Word Type (native vs loanword) significantly affects entropy. Going from a native equivalent to a loanword lowers entropy by about -13.05 ± 2.23 (standard errors), reflecting a less varied or narrower range of surrounding contexts.

The principal research question – Are loanwords semantically specific? – is answered in the affirmative, according to the general trend provided by the quantitative analysis and following the distributional definition of specificity. These anglicisms in the Argentine Spanish corpus are used within a significantly narrower set of contexts, meaning that they are more specific than their counterparts, supporting the extension of the Specificity Hypothesis from code-switching to loanwords.

Qualitative

Of the 70 high-frequency English-origin nouns found in the newspaper corpus, 40 were disqualified from the computational analysis due to two reasons. In some cases, the Spanish semantic equivalents had low counts or did not appear in the corpus. For example *congelador* 'freezer' did not appear in the corpus, though the loanword *freezer* was prevalent. The other reason for disqualification was that some loanwords have no single word or phase that serves as semantic equivalent. For example, the borrowed noun *thriller* may be replaced in Spanish by the compound adjective, *de suspenso*, added to the noun *libro* 'book' or *película* 'movie', but there is no single word or phrase that can reliably serve as an equivalent. In other cases, the polysemy of the loanword poses a problem because each sense of the word has a distinct semantic equivalent, such as the borrowing *stock,* which can be replaced by the Spanish *inventario* to refer to the supply of a store or by *acciones* to refer to a financial security.

These loanwords, along with the 30 from the quantitative analysis discussed above, will be analyzed qualitatively by examining the contexts in which they appear and considering their location along the loanword trajectories described in Weinreich's seminal book Languages in Contact. Weinreich (1953) proposes that loanwords initially cause confusion for a speech community and then follow one of two trajectories. One trajectory is that of replacement, in which the loanword overtakes and occupies the complete space of the original native word. The other is specialization, in which the native and borrowed terms refine their meanings so that both become more specialized, resulting in their sharing the semantic space that was once occupied by just the native term. Another trajectory that is present in this data set is one in which the loanword introduces a new concept into the existing recipient lexicon and therefore lacks a semantic equivalent.

Replacement

The semantic equivalents whose frequencies prove too low for computational analysis reflect the process of being replaced by the loanword. There are 17 such loanwords in the corpus, including *default, country, shorts, jeans, broker, ferry,* and *freezer,* which offer shorter or one-word alternatives to their native counterparts: *incumplimiento* 'default/incompletion', *barrio privado* 'private neighborhood', *pantalones cortos* 'shorts', *vaqueros* 'jeans', *corredor de bolsa* 'broker', *transbordador* 'ferry', and *congelador* 'freezer'. The complete list may be found in the appendix. These loanwords and equivalents fall into semantic domains commonly cited in other anglicism studies: technology, fashion and lifestyle. Rather than occupying a more specific semantic domain than their equivalents, these loanwords have completely or nearly completely occupied the entire semantic domain of the equivalent, thus explaining the shrinking to non-existent presence of the Spanish equivalent in the corpus.

In the quantitative analysis, three pairs exhibited evidence that also seemed to suggest that a process of replacement was underway. These three loanwords have a higher or an almost equal entropy score (i.e., lower and equal specificity) relative to their equivalents: *tablet/tableta, stud/padrillo* and *jockey/jinete.* The loanwords occupy similar, if not greater, semantic space than their native equivalents. Thus these pairs may represent an earlier stage of replacement, whereby the equivalent is still prevalent, yet distributionally, they are comparable, both competing for the same semantic space. For example, stud 'a male horse' is clearly gaining ground over the native *padrillo* in both distribution – with an entropy score of 146 compared to 137 – and in frequency – 146 tokens versus 54[10].

The pair *tablet:tableta* highlights a more complex case, where the semantic equivalent *tableta* shows much greater frequency (677 cases compared to 222), though a slightly more narrow distribution (162 compared to 166)[11]. The native equivalent *tableta* appears predominantly in reference to specific products (e.g. *La tableta T810 posee una pantalla HD* 'the tablet T810 has an HD screen'). The loanword *tablet* also appears in specific product contexts (e.g. *nueva tablet, llamada iPad Air* 'a new tablet called iPad Air'), but is also used in more general discussions of technology (e.g. *podemos asistir compu en mano, o tablet* 'we can attend with computer in hand or tablet'). Though more prevalent, the semantic equivalent *tableta* may have a slightly smaller scope than *tablet*; this could suggest it will be replaced at some point further in the future. Additionally, the cognate status between the two may complicate the issue; as the pair differs by only one letter, speakers may not make as large a distinction between them as they do between

[10] A review of the contexts shows that stud, while polysemous in English (i.e., 'male horse', 'a post within a wall', or 'an attractive male'), retains only the 'male horse' meaning in this corpus.

[11] Though tableta is technically polysemous, referring to both the technological device and a square of chocolate, only three of the 677 uses refer to chocolate and the rest refer to the technological device.

other non-cognate pairs, such as *stud* and *potrillo*. These three loanword and equivalent pairs (*tablet/tableta, stud/potrillo,* and *jockey/jinete*) do not seem to directly support the Specificity Hypothesis, but reflect another trajectory that loanwords may take in their integration into the recipient language.

Specificity

The process of specificity is well documented in the discussion of the quantitative analysis above, as 27 of the 30 pairs reflected this process. Two of the loanword-equivalent pairs with the largest entropy difference, *doodle/dibujo* and *bullying/abuso,* serve as clear examples. The loanword *doodle* is used exclusively to refer to the Google doodles (e.g. *El doodle de Google celebra los 50 años de Doctor Who*. 'Google's doodle celebrates 50 years of Doctor Who') and never refers to a generic sketch or drawing as it can in English. In those cases, the more generic term *dibujo* 'drawing' is utilized. The anglicism *doodle* has entered the lexicon in a very specialized manner to refer to one specific type of drawing and, in this corpus, shows no signs of expanding to broader contexts that the English usage would suggest are possible (e.g. *Her notebook is filled with doodles*). *Bullying*, too, has carved out a semantic space within the broader category of *abuso*, though its use more closely matches its English scope. Both words appear in contexts about children and families (*e.g*., *El bullying es un problema de chicos que lo resuelven los adultos*. 'Bullying is a problem of children that is resolved by adults.' *Las jóvenes y las familias que habrían sido afectadas y condenamos absolutamente cualquier acto de maltrato, abuso y agresión...* 'The girls and families that have been affected and we condemn absolutely any act of mistreat, abuse or aggression...'). However, *abuso* has a much greater scope, commonly appearing in collocations, such as *abuso de poder* 'abuse of power' and *abuso de autoridad* 'abuse of authority', in which *bullying* does not fit. Thus, although *abuso* has the ability to cover the same semantic domain as and more than that of *bullying*, the loanword references a more specific type of abuse that did not have its own name before.

While these loanwords create a new division within a semantic space, other loanwords offer an unambiguous alternative to polysemous native words. This is the case for the loanword and equivalent pairs *team*/*equipo* 'team' or 'equipment', *stand*/*puesto* 'stand' or 'position', and *casting*/*reparto* 'casting' or 'distribution'. Loanwords with polysemous equivalents clearly lend support to the Specificity Hypothesis in that the native equivalent may be used in broader contexts due to the equivalents' multiple meanings. The corresponding loanwords are more specific than the semantic equivalents by the simple measure of number of word senses. The polysemy of the semantic equivalent may contribute to the success of the loanword, as the loanword offers an unambiguous alternative. Polysemy is also present in two loanwords from the corpus: *stock* and *crack*—*stock* referring to either the supply of a store or a financial security and *crack* referring to the drug or a talented athlete. These pairs were not subject to the quantitative analysis, as the current design does not handle two semantic equivalents for one loanword. Future work hopes to remedy this limitation by using automated word sense disambiguation. Word sense disambiguation, identifying which of multiple senses is used for a particular instance of a word, would allow all instances of polysemous borrowings to be classified into sense categories, and each category could be compared to the correct semantic equivalent. The polysemy of borrowings has been shown to contribute to their success in a study on anglicisms in French conducted by Chesley (2010), who found that polysemous borrowings are more likely to become well entrenched into the recipient language than are borrowings with one word sense.

New concept

The last trajectory of loanwords concerns those that introduce a new concept, such as *hockey* or *blog*, which essentially create a new semantic space within the existing recipient-language lexicon. They are easily identifiable in that they lack a clear semantic equivalent. In this data set, these loanwords included numerous sports terms (*scrum, hockey, welter, wing, chukker, handicap, rally*), several technology terms (*blog, chat, drone, streaming, hacker, notebook*), a few business terms (*lobby, holdout*), and several that could loosely be grouped as entertainment terms (*thriller, rock, punk, pub, best seller*). These loanwords have no viable alternative or require a description to convey the same idea; some are challenging to classify, such as *lobby*, which expresses a similar idea as the phrases *presión política* 'political pressure' or *influencia política* 'political influence', but these are not perfect substitutes and may carry negative connotations not shared by the loanword. As these loanwords have no equivalents, it is impossible to quantify their specificity via the concept-based approach applied in this study. Interestingly, they may be categorized as highly specific via the replacement test offered by Backus (2001) — highly specific words are hard to replace with a synonym — thus still supporting the Specificity Hypothesis, even while lacking an equivalent.

When approached from a diachronic perspective, one complication to this seemingly clearly definable category arises: native equivalents may be subsequently introduced as equivalents to these terms, often multi-word collocations (e.g. the phrase *pirata informático* 'hacker' was introduced after to the loanword *hacker* entered the Spanish language, according to Spanish data on Google N-gram Viewer). This phenomenon of creating native equivalents after the adoption of a loanword has been documented in other loanword studies (Onysko & Winter-Froemel, 2011) and often results from notions of language purism or from the need to avoid loanwords in formal documents. A tongue and cheek example of this is the game *Speakons français!,* invented by the French public radio service, *Radio France Internationale,* in which participants suggest colorful and inventive native alternatives for existing anglicisms. While native synonyms may emerge as alternatives, they often struggle to gain ground after a lagging start. *Pirata informático* shows relatively low use as compared to its loanword counterpart in the present Argentine corpus and in Google N-gram Viewer. Future work may explore this diachronic perspective to understand the role of time depth in loanword semantics, though this current study limits itself to a synchronic perspective.

CONCLUSION

The research presented here has provided a semantic analysis of loanwords, pulling from various fields to create a unique perspective on this understudied aspect of lexical borrowings: from distributional semantics, specificity is defined distributionally; from cognitive linguistics, loanwords are analyzed using a concept-based approach; and from computational linguistics, distributional models are adapted to quantifiably measure specificity. Utilizing these methods, 70 high frequency noun borrowings were extracted and analyzed from a 24 million-word corpus of Argentine newspaper Spanish.

The quantitative measure for specificity presented here allows for processing large data sets in a replicable, unbiased manner. One limitation of this approach is that it constrains the number of tokens that may be analyzed. Another is that this model is restricted to large datasets and is most suitable for singletons or frequent multiple-word expressions, thus is not practical for more extensively bilingual texts involving longer code-switched spans. Further exploration of computational linguistic techniques, such as handling data sparsity via web sampling (see Geffet & Dagan, 2005), may remedy some of these limitations. Automated language annotation algorithms could aid in expanding the model proposed here to code-switched texts (see Solorio & Liu, 2008 and Guzman et al., 2016). Future work could also test specificity via other means, as the concept itself is somewhat vague and could be defined numerous ways. This study adopts a definition based on the surrounding lexical variety and applied a methodology based on that definition. However, future work can test alternative definitions, such as verb variety or adjective modification, or even explore other semantic relationships between loanwords and native lexicons. Computational techniques in identifying co-hypernyms, meronyms, and lexical entailment (Roller & Erk, 2016) could all prove useful to loanword-semantics research.

The limitations of the quantitative method are addressed through the qualitative analysis, which has allowed for the computational results to be interpreted within a larger discussion of loanword trajectories as presented by Weinreich, revealing two separate patterns that may define the semantic development of lexical borrowings. Thus, although specificity is not the only semantic option, it is a strong trend among new borrowings and may affect the long-term success of the borrowing. As both qualitative and quantitative approaches examine loanwords that appear over 50 times, the effect of specificity is probably still underestimated, given that low frequency loanwords are likely to be specific due to their limited contextual appearances.

The anglicisms in the corpus that are suitable for quantitative analysis demonstrate a clear pattern of higher specificity than their native equivalents. As specificity is operationalized as a measurement of variability of the surrounding context, these results reflect the fact that loanwords are utilized in more narrow contexts, implying a specific or nuanced meaning as compared to their counterparts. In a similar vein, the majority of loanwords that are not suitable for computational analysis also follow a pattern of high specificity. Thus, both the qualitative and quantitative findings support Backus's Specificity Hypothesis.

These findings may suggest that loanwords' specificity is itself a motive for borrowing, though it is not possible to empirically prove cause and effect with this type of dataset. Speakers of the recipient language may find the precision of a borrowed word motive enough to temporarily (or permanently) abandon the semantic equivalent that previously acted as the default and unmarked choice (see Mackey (1970) for a discussion on quantifying the integration of borrowings and the replacement native equivalents). Another motive for borrowing cited in the literature is loanwords’ ability to serve as prestige markers (Bullock et al., 2015; Ngom, 2000). In contexts where the source language is highly regarded, the prestige of a borrowed word may, in fact, contribute to its specificity. Since meaning is not just denotation, a more prestigious connotation would likely lead to a more specific contextual usage. However, methodologically separating the concepts of prestige and specificity is not yet possible given the analysis proposed here.

This study offers a novel perspective on loanwords with existing semantic equivalents, often viewed as 'unnecessary' when compared to loanwords that introduce new concepts into the recipient language. With the notion of specificity, we may understand these loanwords as disruptors of the semantic system of the recipient language, dividing up the semantic space formerly occupied solely by the native equivalent, thus increasing the level of nuance expressed in the original concept. This conclusion offers a different perspective from Bookless' observation that loanwords with existing equivalents contain more stylistic than referential value: the specificity value of loanwords may, in fact, imply a referential value as well.

In addition to the support found for the Specificity Hypothesis, two other patterns of loanword semantics emerged: replacement and introduction of a new concept. Three pairs from the quantitative analysis and loanwords with low equivalent counts from the qualitative analysis reflect a process of replacement rather than specificity, in which the loanword eventually takes over the whole semantic space previously occupied by the native equivalent. Other loanwords that had no equivalent introduce a new concept. These loanwords create a new semantic space within the existing lexicon, though they too may be viewed as specific, according to the replacement test presented by Backus (2001). A diachronic perspective may shed light on the trajectory of these three semantic patterns as change over time or the introduction of native equivalents may alter the final semantic role of lexical borrowings in the recipient lexicon.

In utilizing a corpus of Argentine Spanish, this study extended the empirical scope of Backus's Specificity Hypothesis to new domains, *i.e.*, to data often viewed as 'monolingual' or representing a situation of weak language contact. What remains to be tested is if this hypothesis holds true for loanwords appearing in situations of more direct bilingualism, such as the Spanish spoken across the United States. As language attitudes towards contact features vary greatly across communities, they may also impact the semantic reshuffling that occurs during loanword adoption.

Chapter 4: Conclusions

SUMMARY OF AIMS AND RESULTS

The overarching goal of this dissertation is to highlight the potential applications of computational methods for the field of contact linguistics. In doing so, this dissertation has presented two case studies, each of which uses computational methods to explore the semantic and social roles of loanwords in Argentine Spanish. The first case study presents the Anglicism Identifier, an algorithm for English loanword identification, which additionally provides the lemmatized form of the token, identifies Named Entities and preserves loan phrases. This algorithm was applied to two corpora, one oral corpus comprising subtitles of Argentine films and one written corpus comprising articles from three major Argentine newspapers, to explore the distribution of loanwords across mediums and social groups. The results reveal that anglicisms are not equally distributed across the data, but rather are most common in the prestigious newspaper La Nación and least common in the film subtitle corpus. This distribution suggests that loanwords may function as prestige markers in the press and that loanwords enter this speech community from the top, through the media, and trickle down into everyday speech.

The second case study draws on an expanded version of the newspaper corpus to measure the semantic specificity of loanwords in comparison to their native equivalents. This study aims to test whether loanwords are more specific than their native equivalents – a hypothesis extended from Backus's Specificity Hypothesis, which affirms that “embedded language elements in code-switching have a high degree of semantic specificity” (2001:128). To test this extended hypothesis, the study borrows from the field of distributional semantics to define semantic specificity and to measure it quantitatively. Both the qualitative and quantitative findings support the hypothesis overall and suggest that loanword specificity may be a motive for borrowing. To a lesser extent, there is also evidence of another semantic trajectory between native equivalents and loanwords: replacement, in which loanwords come to occupy the entire semantic space of the existing equivalents. Lastly, some loanwords introduce a new concept and thus enter the recipient language with no existing equivalent. These loanwords create a new semantic space within the existing lexicon.

LIMITATIONS

This dissertation has methodological and theoretical limitations that merit attention; future work will address some of these limitations.

In the first case study, the Anglicism Identifier presented is designed in a binary fashion wherein each word is classified as Spanish or English. Given the extent of lexical borrowings across languages, this is clearly not a viable premise, even for varieties considered predominantly monolingual. In both corpora utilized in this case study, there are numerous borrowings from French and Italian, along with borrowings from other languages, though to a lesser extent. Forced into the binary system, these borrowings were incorrectly classified as English or Spanish, depending on their particular character sequence. Those labeled as English were particularly problematic, as they increased the number of false positives and in turn negatively affected precision. Future work can remedy this issue by incorporating additional training corpora to augment the language tag set to include French, Italian, and other relevant languages; an augmented tag set offers the possibility of exploring borrowings from other languages in the corpora. Alternatively, character n-gram perplexity could be used, as in Mansikkaniemi & Kurimo (2012), where the researchers set a threshold value in which the highest perplexity values are labeled as foreign. This method could be adapted to create the tag set: Spanish, English or other.

While adding language tags would offer great improvements to the current system, issues in tagging still remain due to lexical overlap between languages, which raises both practical and theoretical considerations when assigning language tags. In both corpora used in this study, several identified borrowings were found to exist in both French and English. Due to extensive borrowing from French into English during numerous points in history, many core items in the English lexicon, such as *portrait, casserole, camp, debit, denim, empire,* and *finance,* are in fact borrowed from French. As a result of their long-standing history in the English language, many of these items are no longer perceived as foreignisms by native speakers. If words with long-standing history in multiple languages are to appear in a corpus of Spanish, or any other language, it can be challenging to determine from what language they are borrowed and therefore what language tag they should receive. An example seen in Chapter 2 is the borrowing *tofu;* the Diccionario del Real Academia Español recognizes the word origin to be from both English and Japanese. The challenge of correctly identifying the source language of a borrowing leads to the question of whether the true origin actually matters, or if the perceived origin is what matters. If Spanish speakers associate the collocation *hotel boutique* 'boutique hotel' with French, even though the term *boutique hotel* originated in the English language – coined by an American in reference to his New York hotel – then what type of borrowing should this term be considered: an anglicism or a gallicisim?

If perception is in fact relevant to a word's status as a borrowing, there is another limitation of the Anglicism Identifier, namely, that there is no distinction between long-standing and new borrowings, which are often perceived quite differently by native speakers. Examples of French in English highlight this point. The borrowings from French, such as *cassette, parachute, café au lait, à la mode, bon appétit, au revoir, raison d'être,* and *je ne sais quoi*, conjure up different levels of association with the French language and culture, possibly due to differences in their length of time in English, their general frequency in English, cultural ties, and their phonetic realization. For sociolinguistic study of language contact, these borrowings would ideally be treated differently. Future work may consider how both time depth and perception of borrowings may be addressed, both in the practical issue of identifying loanwords and in the theoretical issue of defining loanwords. A few loanword studies already add the stipulation that a word has to be recognizable as foreign to native speakers to be considered a borrowing (Zenner et al., 2012). The disjunction between the origin and perception of a word is highlighted in Diab & Kamboj (2011); they evaluate language tagging of code-switched text via the crowd-sourcing platform, Amazon Mechanical Turk, and find that approximately 17% co-turkers confused English words as Hindi.

Another limitation of the Anglicism Identifier is that it does not reliably identify code-switching or adapted loans. In weak contact settings, this limitation is less of a concern, as contact outcomes are mostly limited to lexical borrowings (Thomason & Kaufman, 1988). However, in situations of intense contact, any combination of overt mixing – code-switching or lexical borrowings – and covert mixing – semantic extensions or calques – may be realized. Thus to accurately handle the wide variety of contact outcomes, future work must explore how to combine existing language identifiers, designed for code-switched texts (Solorio & Lui, 2008; Guzman et al., 2016), with anglicism identifiers. Corpora designed to represent code-switched speech, such as the New Mexico Spanish-English Bilingual Corpus and the corpora from BangorTalk, likely include long stretches of monolingual speech with occasion foreign lexical insertions, in addition to code-stitching, as the quantity and quality of language mixing is known to vary based on speaker, situation and register. Loanword Identifiers may prove more accurate in language tagging these segments. Additional computational techniques are also needed to identify covert mixing, possibly using techniques from native language identification, which are also likely to abound in code-switched corpora.

As for the distributional model designed to quantify semantic specificity presented in the second case study, the main limitation of this tool is its dependence on large data. This dependence is common in work on computational semantics, where the corpora used often have word counts in billions. In its current state, this tool is best suited to high-resource language varieties, which most likely limits its scope to standard varieties representing weak contact situations, where data is abundant. To make this tool amenable to smaller datasets, future work may look to methods used for low resource languages (see Littell et al., 2016). For work in contact linguistics, the ability to handle small datasets is especially important as many sites of contact are not well represented in existing mega corpora and do not produce large amounts of easily accessible data, such as newspapers, magazines, or literature, when compared to more standard varieties. An additional limitation of the semantic specificity study is its dependence on semantic equivalents for quantitative analysis. Creating other points of comparison or means of classifying the lexicon could greatly expand what lexical items may be analyzed.

Both studies in their current states are focused on loanwords appearing in predominantly monolingual text; additional work may consider how to adapt these methods to other contact phenomena.

RESEARCH QUESTIONS AND CONTRIBUTIONS

This dissertation took shape around the question: What is the role of corpora in loanword research and, more precisely, what methods could enhance the way we process and analyze corpora to make loanword research more efficient and accountable? In response, two computational methods were presented and evaluated. While both have their limitations, addressed in the previous section, they allow for the efficient processing of large corpora. Therefore they afford greater feasibility and replicability when working with word counts in the millions. The codes for both studies are posted in the public Github repositories jacquelinelars/Semantic-Specificity-Model and jacquelinelars/ Anglicism-Identifier, making them readily available to any researcher wishing to reproduce or extend these methods to other datasets. Implementing the same algorithm to identify loanwords or measure semantic specificity ensures consistency when comparing across corpora, studies, and populations and, in this way, offers a strong advantage over manual annotation. In the case of manual annotation, even if two researchers follow the same annotation guidelines, they will inevitably have inconsistencies between them. This variability has been demonstrated in the creation of gold standards, where testing inter-annotator agreement for various annotation tasks has shown considerable disagreement among the human judges (see Bermingham & Smeaton (2009); Nowak & Rüger (2010); Plank, Hovy, & Søgaard (2014); Véronis (1998)).

In addition to the need for readily available tools, data must be open and shared between researchers. While these corpora for major standardized varieties do exist, this is much less the case for corpora representing contact varieties. Increased availability of these resources, both the methods and the data, will allow for the analysis of language mixing using objective, accountable, and replicable means. This is increasingly important as demand for replication studies is present across disciplines, though often lacking (Burman, Reed & Alm, 2010; Lamal, 1990; Neuliep & Crandall, 1993; Polio & Gass 1997).

In addition to its methodological contributions, this work contributes to theoretical understanding of loanwords and specificity. Issues of loanword classification were addressed by discussing the complications that stem from multiple word origins and the role of perception and time depth in determining the status of a word as foreign or native. The adoption of distributional semantics to define specificity provided a new perspective for contact linguistics, which has placed less attention on semantic properties of contact features and offered few operationalizable frameworks. While there is still much work to be explored, in terms of both the theoretical and methodological contributions that may be gleaned from computational tools in the field of contact linguistics, this dissertation has attempted to make a small step in that direction.

Appendix

Loan	Loan Count	Equivalent	Equivalent Count
casting	61	reparto	407
court	75	cancha	2355
team	55	equipo	12789
doodle	69	dibujo	491
hit	141	éxito	3098
jockey	182	jinete	175
motorman	261	maquinista	456
shopping	384	mercado	10745
delivery	57	entrega	1722
mix	97	mezcla	575
staff	122	personal	3490
test	84	prueba	3481
bullying	94	abuso	1342
stand	98	puesto	1680
stud	141	padrillo	54
celebrity	50	celebridad	211
think tank	55	centro de estudios	285
running	58	correr	162
amenity	62	comodidad	284
pack	74	paquete	875
management	78	gerencia	111
ticket	112	entrada	2577
coach	125	entrenador	2192
look	150	apariencia	289
manager	162	gerente	1003
commodity	182	producto	6086
tablet	183	tableta	677
show	243	espectáculo	1554
performance	363	actuación	1583
mail	590	correo eletrónico	273

References

Aaron, J. E. (2015). Lone English-origin nouns in Spanish: The precedence of community norms. International Journal of Bilingualism, 19(4), 459–480.

Albistur, E. R. (2006). Sistema nacional de consumos culturales argentina. Ciudad de Buenos Aires: Secretaría de Medios de Comunicación de la Jefatura de Gabinete de Ministros de la Presidencia de la Nación. Retrieved from www.consumosculturales.gov.ar

Alex, B. (2006). Integrating language knowledge resources to extend the English inclusion classifier to a new language. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006) (pp. 2431–2436).

Alex, B. (2008). Comparing Corpus-based to Web-based Lookup Techniques for Automatic English Inclusion Detection. In Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08) (pp. 2693–2697).

Andersen, G. (2005). Assessing algorithms for automatic extraction of Anglicisms in Norwegian texts. Presented at the Corpus Linguistics Conference, Birmingham.

Andersen, G. (2014). Pragmatic borrowing. Journal of Pragmatics, 67, 17–33.

Anderson, T. K., & Toribio, A. J. (2007). Attitudes towards lexical borrowing and intra-sentential code-switching among Spanish-English bilinguals. Spanish in Context, 4(2), 217–240.

Aslam, J. A., & Frost, M. (2003). An information-theoretic measure for document similarity. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 449–450).

Backus, A. (2001). The role of semantic specificity in insertional codeswitching:-Evidence from Dutch-Turkish. Jacobson, Rodolfo (Hg): Codeswitching Worldwide. Bd, 2, 125–154.

Balteiro, I. (2011). A reassessment of traditional lexicographical tools in the light of new corpora: sports Anglicisms in Spanish. International Journal of English Studies, 11(2), 23–52.

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender Identity and Lexical Variation in Social Media. Journal of Sociolinguistics, 18(2), 135–160.

Barrs, K. (2014). Lexical Semantics of English Loanwords in Japanese. In Proceedings of the 47th BAAL Annual Meeting (pp. 29–35). University of Warwick, Coventry.

Bates, D., Maechler, M., Bolker, B., Walker, S., & others. (2014). lme4: Linear mixed-effects models using Eigen and S4. R Package Version, 1(7).

Bein, R. (n.d.). La situación de las lenguas extranjeras en la Argentina 1. Introducción. Retrieved from http://linguasur.org.ar/panel/archivos/f9227ef50db3de732e1d3f897de85aa8Bein%20lenguas%20extranjeras.pdf

Bermingham, A., & Smeaton, A. F. (2009). A Study of Inter-annotator Agreement for Opinion Retrieval. In Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 784–785). New York, NY, USA: ACM.

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. Sebastopol, California: O'reilly.

Boczkowski, P. J., & de Santos, M. (2007). When More Media Equals Less News: Patterns of Content Homogenization in Argentina's Leading Print and Online Newspapers. Political Communication, 24(2), 167–180.

Bookless, T. C. (1982). Towards a semantic description of English loan-words in Spanish. Quinquereme, 5, 170–85.

Bordelois, I. (2011). El país que nos habla. Buenos Aires, Argentina: Penguin Random House Grupo Editorial Argentina.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41(4), 977–990.

Bullock, B., Serigos, J., & Toribio, A. J. (accepted). Exploring a loan translation and its consequences in an oral bilingual corpus. Journal of Language Contact.

Bullock, B., Serigos, J., & Toribio, A. J. (2015). The status of Anglicisms in Puerto Rican Spanish-language Press. In Code-switching in the Spanish-speaking Caribbean and its diaspora (C. Mazak, M. Parafita Cuoto, R. Guzzardo). New York: John Benjamins.

Bullock, B., & Toribio, A. J. (2013). Spanish in Texas Corpus. http://www.spanishintexas.org

Burman, L. E., Reed, W. R., & Alm, J. (2010). A Call for Replication Studies. Public Finance Review, 38(6), 787–793.

Caraballo, S. A., & Charniak, E. (1999). Determining the specificity of nouns from text. In Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (pp. 63–70). Citeseer.

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval (pp. 161–175). Las Vegas.

Chan, T. W., & Goldthorpe, J. H. (2007). Social Status and Newspaper Readership. American Journal of Sociology, 112(4), 1095–1134.

Chesley, P. (2010). Lexical borrowings in French: Anglicisms as a separate phenomenon. Journal of French Language Studies, 20(3), 231–251.

Chesley, P., & Baayen, R. H. (2010). Predicting new words from newer words: Lexical borrowings in French. Linguistics, 48, 1343–1374.

Cho, E., Myers, S. A., & Leskovec, J. (2011). Friendship and Mobility: User Movement in Location-based Social Networks. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1082–1090). New York, NY, USA: ACM.

Cimiano, P., Hotho, A., & Staab, S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. J. Artif. Intell. Res.(JAIR), 24(1), 305–339.

Crystal, D. (1999). World English: past, present, future,. Presented at the ASKO Europa-Stiftung symposium on Weltgesellschaft, Weltverkehrssprache, Weltkultur, “Globalisierung vs. Fragmentierung.”

Crystal, D. (2012). English as a global language. Cambridge University Press.

Daulton, F. E. (2004). The Creation and Comprehension of English Loanwords in the Japanese Media. Journal of Multilingual and Multicultural Development, 25(4), 285–296.

De Longueville, B., Smith, R. S., & Luraschi, G. (2009). “OMG, from Here, I Can See the Flames!”: A Use Case of Mining Location Based Social Networks to Acquire Spatio-temporal Data on Forest Fires. In Proceedings of the 2009 International Workshop on Location Based Social Networks (pp. 73–80). New York, NY, USA: ACM.

De Smedt, T., & Daelemans, W. (2012). Pattern for python. Journal of Machine Learning Research, 13(Jun), 2063–2067.

Deuchar, M., Davies, P., Herring, J., Couto, M. P., & Carter, D. (2014). Building bilingual corpora. Advances in the Study of Bilingualism, 93–111.

Diab, M., & Kamboj, A. (2011). Feasibility of Leveraging Crowd Sourcing for the Creation of a Large Scale Annotated Resource for Hindi English Code Switched Data: A Pilot Annotation (pp. 36–40). Presented at the 9th Workshop on Asian Language Resources, Chiang Mai, Thailand.

Dunning, T. (1994). Statistical identification of language (Technical Memo). Las Cruces, NM: Computing Research Laboratory, New Mexico State University. Retrieved from https://pdfs.semanticscholar.org/bff2/b05f369187775640593dd152f8af723b76cd.pdf

Edmonds, P., & Hirst, G. (2002). Near-synonymy and lexical choice. Computational Linguistics, 28(2), 105–144.

Eisenstein, J. (2017). Written dialect variation in online social media. In C. Boberg, J. Nerbonne, & D. Watt (Eds.), Handbook of Dialectology. Wiley.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. Language and Linguistics Compass, 6(10), 635–653.

Erk, K. (2015). What do you know about an alligator when you know the company it keeps? Semantics and Pragmatics, 9, 1–63.

Firth, J. (1957). A synopsis of linguistic theory 1930-1955. Studies in Linguistic Analysis.

Fisherman, H. (1990). Attitudes toward foreign words in contemporary Hebrew. International Journal of the Sociology of Language, 86(1), 5-40.

Francom, J., Hulden, M., & Ussishkin, A. (2014). ACTIV-ES: a comparable, cross-dialect corpus of “everyday”Spanish from Argentina, Mexico, and Spain. In The Ninth International Conference on Language Resources and Evaluation (pp. 1733–1737).

Friedrich, P. (2003). English in Argentina: attitudes of MBA students. World Englishes, 22(2), 173–184.

Furiassi, C., Pulcini, V., & González, F. R. (2012). The Anglicization of European Lexis. Amsterdam, The Netherlands/Philadelphia, PA: John Benjamins Publishing.

Gall, T. L., & Hobby, J. M. (Eds.). (2007). Argentina. In Worldmark Encyclopedia of the Nations (12th ed., Vol. 3, pp. 11–34). Detroit: Gale.

Geffet, M., & Dagan, I. (2005). The distributional inclusion hypotheses and lexical entailment. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 107–114). Association for Computational Linguistics.

Graedler, A.-L. (2004). Modern loanwords in the Nordic countries. Presentation of a project. Nordic Journal of English Studies, 3(2), 5–22.

Graedler, A.-L. (2014). Attitudes towards English in Norway: A corpus-based study of attitudinal expressions in newspaper discourse. Multilingua, 33(3–4).

Grefenstette, G. (1995). Comparing Two Language Identification Schemes. In JADT (1995) 3rd International conference on Statistical Analysis of Textual Data. Rome.

Gries, S. T. (2006). Exploring variability within and between corpora: some methodological considerations. Corpora, 1(2), 109–151.

Guzman, G. A., Serigos, J., Bullock, B. E., & Toribio, A. J. (2016). Simple Tools for Exploring Variation in Code-switching for Linguists. In Proceedings of the Second Workshop on Computational Approaches to Code Switching (pp. 12–20). Austin, Texas: Association for Computational Linguistics.

Haspelmath, M. (2008). Loanword typology: Steps toward a systematic cross-linguistic study of lexical borrowability. Empirical Approaches to Language Typology, 35, 43-62.

Hassall, T., Murtisari, E. T., Donnelly, C., & Wood, J. (2008). Attitudes to western loanwords in Indonesian. International Journal of the Sociology of Language, 2008(189), 55-84.

Haugen, E. (1950). The Analysis of Linguistic Borrowing. Language, 26(2), 210–231.

Hlavac, J. (2006). Bilingual discourse markers: Evidence from Croatian–English code-switching. Journal of Pragmatics, 38(11), 1870–1900. https://doi.org/10.1016/j.pragma.2006.05.005

Hornikx, J., Meurs, F. van, & Boer, A. de. (2010). English or a Local Language in Advertising? The Appreciation of Easy and Difficult English Slogans in the Netherlands. Journal of Business Communication, 47(2), 169–188.

Hovy, D., Johannsen, A., & Søgaard, A. (2015). User review sites as a resource for large-scale sociolinguistic studies. In Proceedings of The 24th International World Wide Web Conference (WWW).

Hughes, B., Baldwin, T., Bird, S. G., Nicholson, J., & MacKinlay, A. (2006). Reconsidering language identification for written language resources. In Proceedings, 5th International Conference on Language Resources and Evaluation (LREC2006). Genoa, Italy.

Johnson, A. (2009). The rise of English: The language of globalization in China and the European Union. Macalester International, 22(1), 131-168.

Jurafsky, D., & Martin, J. H. (2008). Speech and Language Processing, 2nd Edition (2nd edition). Upper Saddle River, N.J: Prentice Hall.

Jurafsky, D., & Martin, J. H. (2016). Speech and Language Processing, 3rd Edition. Upper Saddle River, N.J.

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In English in the world: Teaching and learning the language and literatures (pp. 11–36). Cambridge University Press.

Kikui, G. (1996). Identifying, the coding system and language, of on-line documents on the Internet. In Proceedings of the 16th conference on Computational linguistics-Volume 2 (pp. 652–657). Association for Computational Linguistics.

Koo, H. (2015). An unsupervised method for identifying loanwords in Korean. Language Resources and Evaluation, 49, 1–19.

Lamal, P. A. (1990). On the Importance of Replication. Journal of Social Behavior and Personality; Corte Madera, CA, 5(4), 31–35.

Leidig, S., Schlippe, T., & Schultz, T. (2014). Automatic Detection of Anglicisms for the Pronunciation Dictionary Generation: A Case Study on Our German IT Corpus. In Spoken Language Technologies for Under-Resourced Languages.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. Italian Journal of Linguistics, 20(1), 1–31.

Lins, R. D., & Gonçalves, P. (2004). Automatic language identification of written texts. In Proceedings of the 2004 ACM symposium on Applied computing (pp. 1128–1133). ACM.

Littell, P., Goyal, K., Mortensen, D., Little, A., Dyer, C., & Levin, L. (2016). Named Entity Recognition for Linguistic Rapid Response in Low-Resource Languages: Sorani Kurdish and Tajik. In the 26th International Conference on Computational Linguistics.

Loveday, L. (1996). Language contact in Japan: a socio-linguistic history. Oxford□: New York: Clarendon Press□; Oxford University Press.

MacKenzie, I. (2012). Chapter 1. Fair play to them: Proficiency in English and types of borrowing. In C. Furiassi, V. Pulcini, & F. Rodríguez González (Eds.), The Anglicization of European Lexis (pp. 27–42). Amsterdam: John Benjamins Publishing Company.

Mackey, W. (1970). Interference, Integration and the Synchronic Fallacy. In George Washington University round table on languages and linguistics 23 (pp. 195–227).

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. New York: Cambridge University Press.

Manovich, L. (2011). Trending: the promises and the challenges of big social data. In Debates in the Digital Humanities. U of Minnesota Press.

Mansikkaniemi, A., & Kurimo, M. (2012). Unsupervised vocabulary adaptation for morph-based language models. In Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT (pp. 37–40). Association for Computational Linguistics.

Matus-Mendoza, M. (2002). The English Lexical Loan: A Class Marker. Journal of Hispanic Higher Education, 1(4), 329–337.

McCarthy, D. (2009). Word sense disambiguation: An overview. Language and Linguistics Compass, 3(2), 537–558.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the Word Frequency Effect: The Neglected Role of Distributional Information in Lexical Processing. Language and Speech, 44(3), 295–322.

McEnery, T., & Wilson, A. (2003). Corpus linguistics. The Oxford Handbook of Computational Linguistics, S, 448–463.

McNamee, P., & Mayfield, J. (2004). Character n-gram tokenization for European language text retrieval. Information Retrieval, 7(1–2), 73–97.

Melgarejo, G. (2011, April 11). Préstamos legítimos y préstamos innecesarios. La Nación. Retrieved from http://www.lanacion.com.ar/1364531-prestamos-legitimos-y-prestamos-innecesarios

Messineo, C., & Cúneo, P. (2006). Las lenguas indígenas de la Argentina: situación actual e investigaciones. In Third International Workshop on (Semi) Numerical Techniques in Polynomial Equation Solving, in Honor of Joos Heintz's 60th. Buenos Aires.

Myers-Scotton, C. (1997). Duelling languages: Grammatical structure in codeswitching. Oxford University Press.

Myers-Scotton, C. (2002). Contact Linguistics: Bilingual Encounters and Grammatical Outcomes (1St Edition). Oxford: Oxford University Press, USA.

Myers-Scotton, C., & Jake, J. L. (1995). Matching lemmas in a bilingual language competence and production model: Evidence from intrasentential code switching. Linguistics, 33, 981–1024.

Neuliep, J. W., & Crandall, R. (1993). Reviewer Bias Against Replication Research. Journal of Social Behavior and Personality; Corte Madera, CA, 8(6), 21–29.

Ngom, F. (2000). Sociolinguistic motivations of lexical borrowings in Senegal. Studies in the Linguistic Sciences, 30(2).

Ngom, F. (2003). The social status of Arabic, French, and English in the Senegalese speech community. Language Variation and Change, 15(3), 351–368.

Nielsen, P. M. (2003). English in Argentina: A Sociolinguistic Profile. World Englishes, 22(2), 199–209.

Nowak, S., & Rüger, S. (2010). How Reliable Are Annotations via Crowdsourcing: A Study About Inter-annotator Agreement for Multi-label Image Annotation. In Proceedings of the International Conference on Multimedia Information Retrieval (pp. 557–566). New York, NY, USA: ACM.

Ogle, D. H. (2017). FSA: Fisheries Stock Analysis.

Onysko, A., & Winter-Froemel, E. (2011). Necessary loans – luxury loans? Exploring the pragmatic dimension of borrowing. Journal of Pragmatics, 43(6), 1550–1567.

Otheguy, R., & Stern, N. (2011). On so-called Spanglish. International Journal of Bilingualism, 15(1), 85–100.

Pagano, M. (2013, January 19). Las frases en inglés, un guiño de informalidad. La Nación. Retrieved from http://www.lanacion.com.ar/1547183-las-frases-en-ingles-un-guino-de-informalidad

Peterson, E., & Vaattovaara, J. (2014). Kiitos and pliis: The relationship of native and borrowed politeness markers in Finnish. Journal of Politeness Research: Language, Behavior, Culture, 10(2), 247–269.

Plank, B., Hovy, D., & Søgaard, A. (2014). Learning part-of-speech taggers with inter-annotator agreement loss. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (pp. 742–751). Gothenburg, Sweden: Association for Computational Linguistics.

Polio, C., & Gass, S. (1997). Replication and reporting. Studies in Second Language Acquisition, 19(4), 499–508.

Poplack, S., & Sankoff, D. (1984). Borrowing: the synchrony of integration. Linguistics, 22(1), 99-135.

Poplack, S., Sankoff, D., & Miller, C. (1988). The social correlates and linguistic processes of lexical borrowing and assimilation. Linguistics, 26(1), 47–104.

R Core Team. (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Roffo, J. (2016, June 5). Afirman que los anglicismos tecno empobrecen el castellano. Clarín. Retrieved from https://www.clarin.com/sociedad/Afirman-anglicismos-tecno-empobrecen-castellano_0_Ekfi7naQZ.html

Roller, S., & Erk, K. (2016). Relations such as hypernymy: Identifying and exploiting hearst patterns in distributional vectors for lexical entailment. In In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas, USA: Association for Computational Linguistics.

Roller, S., Erk, K., & Boleda, G. (2014). Inclusive yet Selective: Supervised Distributional Hypernymy Detection. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics (pp. 1025–1036). Dublin, Ireland.

Rosner, M., & Farrugia, P.-J. (2007). A tagging algorithm for mixed language identification in a noisy domain. In Proceedings of the Eighth Annual Conference of the International Speech Communication Association (pp. 190–193). Antwerp, Belguim: ISCA.

Ruellot, V. (2011). English in French print advertising from 1999 to 2007. World Englishes, 30(1), 5–20.

Ryu, P.-M., & Choi, K.-S. (2006). Taxonomy learning using term specificity and similarity. In Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge (pp. 41–48).

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In Proceedings of international conference on new methods in language processing (Vol. 12, pp. 44–49).

Serigos, J. (2016). Using distributional semantics in loanword research: A concept-based approach to quantifying semantic specificity of Anglicisms in Spanish. *International Journal of Bilingualism*.

Solorio, T., & Liu, Y. (2008). Learning to predict code-switching points. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 973–981). Honolulu, Hawaii.

Thomason, S. G., & Kaufman, T. (1988). Language contact, creolization, and genetic linguistics. Univ of California Press.

Torres Cacoullos, R., & Travis, C. E. (2015a). New Mexico Spanish-English Bilingual (NMSEB) corpus, National Science Foundation 1019112/1019122. Retrieved from http://nmcode-switching.la.psu.edu/

Treffers-Daller, J. (1999). Borrowing and shift-induced interference: Contrasting patterns in French–Germanic contact in Brussels and Strasbourg. Bilingualism: Language and Cognition, 2(1), 1–22.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37(1), 141–188.

Varga, D., Orešković Dvorski, L., & Bjelobaba, S. (2012). English Loanwords in French and Italian Daily Newspapers. Studia Romanica et Anglica Zagrabiensia, 56, 71–84.

Varra, R. M. (2013). The social correlates of lexical borrowing in Spanish in New York City. Unpublished PhD thesis, City University of New York.

Véronis, J. (1998). A study of polysemy judgements and inter-annotator agreement. In Programme and advanced papers of the Senseval workshop (pp. 2–4).

Weeds, J., Weir, D., & McCarthy, D. (2004). Characterising measures of lexical distributional similarity. In Proceedings of the 20th international conference on Computational Linguistics (p. 1015). Association for Computational Linguistics.

Weinreich, U. (1953). Languages in contact, findings and problems. New York: Linguistic Circle of New York.

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Retrieved from http://ggplot2.org

Winford, D. (2003). An Introduction to Contact Linguistics (1st ed.). Wiley-Blackwell.

Winter-Froemel, E. (2013). Formal variance and semantic changes in borrowing: Integrating semasiology and onomasiology. New Perspectives on Lexical Borrowing: Onomasiological, Methodological and Phraseological Innovations, 7, 65-100.

Zar, J. (2010). Biostatistical analysis. 5th Prentice Hall. Inc Upper Saddle River, NJ, USA.

Zenner, E., Speelman, D., & Geeraerts, D. (2012). Cognitive Sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch. Cognitive Linguistics, 23(4), 749–792.

Zenner, E., Speelman, D., & Geeraerts, D. (2014). A sociolinguistic analysis of borrowing in weak contact situations: English loanwords and phrases in expressive utterances in a Dutch reality TV show. International Journal of Bilingualism, 19(3), 333–346.