Cross-Dialectal Usage of Augmentative and Diminutive Suffixes in Spanish Online:

AN EXERCISE IN CORPUS LINGUISTICS

Carlos Hervey Gómez, Jr.

An Honors thesis presented to the Department of Linguistics at the University of Colorado – Boulder and before the following Committee:

Jeremy Calder, Ph.D. (Assist. Prof., Linguistics), Chase Wesley Raymond, Ph.D. (Assoc. Prof., Linguistics and Family Medicine), and José Javier Rivas-Rodríguez, Ph.D. (Assoc. Prof., Spanish and Portuguese)

under the direction of Susan Windisch Brown, Ph.D. (Assoc. Director., CLASIC and Senior Research Assoc., Computer Science and Linguistics)

University of Colorado – Boulder

March 21, 2023

# Table of Contents

[Table of Contents 3](#_Toc132752582)

[Table of Figures 5](#_Toc132752583)

[Abstract 6](#_Toc132752584)

[1 – Introduction 7](#_Toc132752585)

[1.1 – Guiding Statements 7](#_Toc132752586)

[1.2 – Background Information 7](#_Toc132752587)

[1.2.1 – Shingles of Linguistics 7](#_Toc132752588)

[1.2.2 – Corpus Linguistics 9](#_Toc132752589)

[1.3 – Literature Review 10](#_Toc132752590)

[1.3.1 – Los llamados diminutivos y aumentativos en el español de México (Gaarder, 1966) 10](#_Toc132752591)

[1.3.2 – La función del diminutivo en español (Zuluaga Ospina, 1970) 13](#_Toc132752592)

[1.3.3 – A contrastive study of English and Spanish synthetic diminutives (Hagg, 2016) 14](#_Toc132752593)

[1.3.4 – Variación y variedad del diminutivo en español a través de dos corpus originales (Criado and Andión, 2016) 15](#_Toc132752594)

[1.3.5 – “Sólo un poquito” El uso y funciones del diminutivo en español peninsular en dos grupos de Facebook (Maíz-Arévalo, 2018) 16](#_Toc132752595)

[1.3.6 – Gaps in the literature and the need for this paper 16](#_Toc132752596)

[1.4 – Thesis Project Structure 17](#_Toc132752597)

[2 – Part I: Corpus Creation 18](#_Toc132752598)

[2.1 – Background Information 18](#_Toc132752599)

[2.1.1 – Regions of Study 18](#_Toc132752600)

[2.1.2 – Other Corpora 19](#_Toc132752601)

[2.1.3 – SpanRed 20](#_Toc132752602)

[2.1.4 – Sketch Engine Compilation 21](#_Toc132752603)

[2.2 – Materials and Methods 21](#_Toc132752604)

[2.2.1 – Survey Method 21](#_Toc132752605)

[2.2.2 – Optimization Algorithm 22](#_Toc132752606)

[2.2.3 – Compilation 23](#_Toc132752607)

[2.3 – Results 23](#_Toc132752608)

[2.4 – Discussion and Conclusions 24](#_Toc132752609)

[3 – Part II: Replication Experiment with Gaarder’s System 26](#_Toc132752610)

[3.1 – Background Information 26](#_Toc132752611)

[3.1.1 – Endings 26](#_Toc132752612)

[3.1.2 – Semantic Categorization 27](#_Toc132752613)

[3.1.3 – Three Primary Layers of Meaning 28](#_Toc132752614)

[3.2 – Materials and Methods 29](#_Toc132752615)

[3.2.1 – Data Sets 29](#_Toc132752616)

[3.2.2 – Annotation 29](#_Toc132752617)

[3.3 – Results and Analyses 30](#_Toc132752618)

[3.3.1 – Interrater Agreement Quotient 30](#_Toc132752619)

[3.3.2 – Confusion Matrix 30](#_Toc132752620)

[3.4 – Discussion and Conclusions 31](#_Toc132752621)

[4 – Part III: Experiment with Revised Categories 33](#_Toc132752622)

[4.1 – Creating Final Categories 33](#_Toc132752623)

[4.2 – Materials and Methods 33](#_Toc132752624)

[4.2.1 – Final Category List 33](#_Toc132752625)

[4.2.2 – Corpus, Exclusion/Inclusion Lists, and Annotators 34](#_Toc132752626)

[4.2.3 – Category Calibration Experiment 34](#_Toc132752627)

[4.3 – Results and Discussions 34](#_Toc132752628)

[4.3.1 – Recap of Research Questions and Hypotheses 34](#_Toc132752629)

[4.3.2 – Research Question 1 35](#_Toc132752630)

[4.3.3 – Research Question 2 40](#_Toc132752631)

[4.3.4 – Research Question 3 45](#_Toc132752632)

[4.3 –Final Discussion 49](#_Toc132752633)

[5 – Conclusions 51](#_Toc132752634)

[5.1 – Final Conclusions 51](#_Toc132752635)

[5.2 – Areas of Future Study 51](#_Toc132752636)

[Appendices 52](#_Toc132752637)

[Appendix B – Part I 52](#_Toc132752638)

[Appendix C – Part II 52](#_Toc132752639)

[Appendix D 56](#_Toc132752640)

[Works Cited 66](#_Toc132752641)

# Table of Figures

[Equation 1 – Interrater Agreement Quotient 27](#_Toc129884873)

[Figure 1 – The Shingles of Linguistics 8](#_Toc132739710)

[Figure 2 – Example of Concordance 30](#_Toc132739711)

[Figure 3 – Skew in Referential Suffix Usage, by Primary 45](#_Toc132739712)

[Figure 4 – Skew in Expressive Suffix Usage, by Primary 45](#_Toc132739713)

[Figure 5 – Skew in Expressive Suffix Use, by Primary 47](#_Toc132739714)

[Table 1 – Characteristics and Features of Independent Corpora 18](#_Toc129884879)

[Table 2 – Characteristics of SketchEngine Corpora 18](#_Toc129884880)

[Table 3 – Cutoff Thresholds of Dialectal Sub-corpora 21](#_Toc129884881)

[Table 4 – Confusion Matrix (Replication Experiment) 28](#_Toc129884882)

[Table 5 – Augmentative frequency per million nouns and adjectives, by geographic region 32](#_Toc129884883)

[Table 6 – Diminutive frequency per million nouns and adjectives, by geographic region 33](#_Toc129884884)

[Table 7 – Share of Augmentation 34](#_Toc129884885)

[Table 8 – Ratio of Augmentation to Diminution 35](#_Toc129884886)

[Table 9 – Share of Augmentative Semantic Function 37](#_Toc129884887)

[Table 10 – Share of Diminutive Semantic Function 38](#_Toc129884888)

[Table 11 – Share of Augmentation by Primary Category 39](#_Toc129884889)

[Table 12 – Share of Diminution by Primary Category 39](#_Toc129884890)

[Table 13 – Share of Positivity and Negativity in Augmentation (by Primary Category) 43](#_Toc129884891)

[Table 14 – Share of Positivity and Negativity in Diminution (by Primary Category) 43](#_Toc129884892)

[Table 15 – Share of Positivity and Negativity in Augmentation (any Inclusion) 43](#_Toc129884893)

[Table 16 – Share of Positivity and Negativity in Diminution (any Inclusion) 43](#_Toc129884894)

[Table 17 – Ratio of Positivity to Negativity in Augmentation and Diminution (by Primary Category) 44](#_Toc129884895)

[Table 18 – Ratio of Positivity to Negativity in Augmentation and Diminution (Any Inclusion) 44](#_Toc129884896)

[Table 19 – Share of Referentiality and Expressivity in Augmentation (by Primary) 45](#_Toc129884897)

[Table 20 – Share of Referentiality and Expressivity in Diminution (by Primary) 45](#_Toc129884898)

[Table 21 – Pure Referentiality vs. any Expressivity in Augmentation 46](#_Toc129884899)

[Table 22 – Pure Referentiality vs. any Expressivity in Diminution 46](#_Toc129884900)

[Table 23 – Ratio of Referentiality to Expressivity in Augmentation and Diminution (by Primary) 46](#_Toc129884901)

[Table 24 – Ratio of Pure Referentiality vs any Expressivity 46](#_Toc129884902)

# Abstract

Using SketchEngine, 10 million tokens were compiled from Reddit in 2021 and annotated from seven nations representing seven major dialectal regions of the Spanish language (Mexico, the Dominican Republic, Guatemala, Colombia, Chile, Argentina, and Spain) and sorted into regional sub-corpora. A categorization scheme was devised for the semantic functions of augmentatives and diminutives in Spanish, which, once verified, was then applied to a corpus sample, yielding results that confirmed that different dialectal regions of Spanish speakers do exhibit statistically significant regional variations in the augmentative and diminutive suffixes they prefer, the rate at which they augment or diminish, and the meanings they most and least often convey with such suffixes.

Notable among the findings is the distinct Colombian trend of using augmentative and diminutive suffixes sparingly but with focused purpose, the near-universal propensity of Spanish speakers to express positive attitudes through Affectionate diminutive use while expressing negative ones through Contemptuous augmentation, the monopoly of Magnitude and Accentuation in both diminutive and especially augmentative usage, and the cross-dialectal conflation of Magnitudinal and Accentual augmentation, perhaps due to metaphorical mapping. An important corollary finding is that Peninsular Spanish maintains distinct currents from Magnitude and Accentuation to traditionally dedicated augmentatives (-*ón*/*ote/azo* and -*ísimo*, respectively) whereas in all studied American dialects, to varying degrees, the distinction has eroded and the semantic flows comingle among the augmentative suffixes.

# 1 – Introduction

## 1.1 – Guiding Statements

The aim of this study is to observe the usage patterns in augmentative and diminutive suffixation within the Hispanosphere, with the secondary focus of surveying the existence and usage patterns of multiple augmentation and diminution. This investigation was guided by the following Research Questions:

1. Does regional variation exist in augmentative or diminutive use in Spanish?
	1. Concerning regional preference for various suffixes
	2. Concerning regional propensity to augment or diminish
2. What semantic functions do Spanish augmentatives and diminutives serve?
3. Do augmentative or diminutive semantic functions vary across geographic regions?

These questions, as well as a survey of background literature, have yielded the following Hypotheses, rooted in the Spanish-speaking world only:

1. Regional variation in augmentative and diminutive use is expected
	1. Regional preferences in suffix use are expected
	2. Regional variation in overall augmentation and diminution is expected
2. Augmentatives and diminutives are expected to perform various distinct semantic functions
3. Augmentative and diminutive semantic functions are expected to vary by region

## 1.2 – Background Information

### 1.2.1 – Shingles of Linguistics

To understand these aims, research questions, and hypotheses, a grounding in basic linguistic terminology as well as a familiarization with key background literature is required. The concept that will guide the exploration of this study’s foundational knowledge is what the Primary Investigator has termed the Shingles of Linguistics, which visualizes the broadening of linguistic fields of meaning as a linear progression of layers along two axes, scope and context-dependence, starting with phonetics and progressing through to pragmatics and beyond, each stratum layered atop the previous but with a significant overhang towards more and more broader context-dependence, illustrated belower:

Figure – The Shingles of Linguistics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Scope** |  |  |  |  |  | Pragmatics |
| Semantics |  |
| Syntax |  |
| Morphology |  |
| Phonology |  |
| Phonetics |  |
|  | **Context-dependence** |

The Author further prefers to group this progression into thirds: Phonetics and Phonology, Morphology and Syntax, and finally Semantics and Pragmatics. This visualization is not meant to be exact or authoritative, it is merely presented as a visual aid and adaptation of other, pre-established, conceptual models.

This investigation deals with the semantic characteristics of Spanish suffixes; therefore, phonetics and phonology are not fields of meaning pertinent to this study. Had this been a diachronic (over time) study, they still would have only been pertinent as possible explanations for the origin and distribution of certain suffixes of study, but being that this is a mostly synchronic study, the exploration of background material shall begin in morphology.

#### 1.2.1 §1 – Morphology and Syntax

Morphology is the study of how the shape of lexemes, i.e., how and from what they are constructed, affects their meaning. A lexeme is a unit of meaning, a word or phrase that means something within the structured list of words and meanings in the mind of the speaker, or the lexicon. Words are, depending on the nature, or typology, of the language, often composed of smaller, indivisible subunits, known as morphemes. These morphemes are either inflectional or derivational; that is, they are either required to participate in the grammatical structure of the language or are used to create novel words or impart new meanings, respectively. The distinction is also made between free and bound morphemes, or morphemes that can exist independently, often as separate words, and those that cannot, respectively.

The binding of morphemes to primitive bases, words or roots that are considered unmodified, is known as affixation, the morphemes being bound known as affixes. Affixation specifically to the ends of bases is known as suffixation, with the morphemes being bound in this manner known as suffixes. The multiple application of affixes of any class, including augmentatives and diminutives, is known as multiple iteration. The multiple application of the same affix is known as affix reiteration. The Spanish language has a particularly wide selection of bound derivational suffixes known as augmentatives and diminutives, suffixes that, nominally, enlarge or shrink the referents of their roots, respectively. While multiple iteration is a widely documented phenomenon, reiteration is relatively poorly attested and -understood.

Productivity describes the level of activity or usage a linguistic process or element possesses as an independent entity in the minds and creative processes of a speaker or group of speakers. There is, theoretically, a question of degrees of productivity, however, as, in a language that utilizes suffixes often, suffixation may be characterized as productive, but if the same language is in the habit of deriving new words almost exclusively with prefixes, prefixation may, according to some models, including the one used in this study, be characterized as the more productive derivational strategy in this hypothetical language. Elements which used to be productive but no longer are, i.e., processes which have fallen into disuse, are referred to as non-productive. Thus, a corollary definition of productivity, used by Bauer, is the ability to produce novel forms. This study distinguishes between productive and non-productive augmentative and diminutive suffixes because the two sets affect speech patterns in markedly diverse ways, with productive suffixes being the larger influence in language usage and the future of the language.

In the space between Morphology and Syntax and Semantics and Pragmatics, there exists the concept of expressivity, a term coined by Roman Jakobson to describe language that “functions to express the speaker’s attitude towards what he is speaking about.” This draws a distinction between expressive and referential language, which more directly describes the speaker’s subject matter. Though sometimes presented as a binary status, there is some literature to suggest that there exist degrees of expressivity. This concept affects the study on a morphological and semantic level, lining up nicely with the following, more semantics-specific, concepts.

#### 1.2.1 §2 – Semantics and Pragmatics

Semantics and pragmatics are both fields of linguistics that deal primarily with meaning. The primary distinction between these fields, as illustrated above, is the dependence of this meaning on context. This distinction has a few important implications, the most relevant of which is the spectrum between entirely semantic and entirely pragmatic meanings.

At the concrete end of this spectrum are denotations, semantic meanings that function as the delineations of the concepts being manipulated during discourse. Rooted in morphology and syntax, these meanings are lexical, that is, they are prefabricated and imported wholesale into each linguistic encounter, and rigid, that is, largely impervious to context save to select which denotative meaning is being used, should multiple be assigned to a given word. They comprise the text of discourse and are far more likely than not to be understood by the listener.

Being that pragmatic meanings are almost entirely drawn from contextual clues in individual speech interactions, they are characterized as spontaneous and amorphous. They are spawned *ad hoc* within individual speech interactions and are exceedingly malleable and subjective. Like negative space, they rely on the juxtaposition of more concrete forces as well as the overlay of more ephemeral and non-verbal channels of communication and context, such as tone, body language, and timing. These meanings, which comprise a sizable portion of discourse subtext, are among the most likely to be missed by listeners and readers without an audiovisual medium, owing to their amorphousness, spontaneity, and often non-verbal medium. Because of this, though they exist and influence the study, pragmatic meanings will not be the subject of concentration here. However, their presence is useful to anchor the end of the spectrum opposite to denotation, allowing for the existence of an interstitial category of meaning: the connotation.

If denotations delineate discursive concepts, connotations instead color them with expressive value and attitude judgements from the speaker towards their referent object. Like denotations, connotations a largely lexical set of semantic meanings, however, unlike denotations, these meanings are not always rigid. Their fabrication is dependent on the culture and personal values of the speaker or listener, sometimes resulting in a disconnect in the transfer of subtextual information. This possibility of disconnection or misinterpretation are an example of the susceptibility of connotations to the contextual pressures of speech interactions, making them more pliable, and, to some extent, spontaneous, than their denotative counterparts.

As noted above, connotations are usually quite expressive, often comprised of appreciative or depreciative value judgments and attitudes the speaker holds towards the primitive base. Expressivity can manifest itself even more purely and pragmatically, encoding attitudes and value judgments of the speaker towards the subject matter in general and even the audience itself, but, as stated before, such pragmatic meanings are outside the purview of this investigation.

#### 1.2.1 §3 – Augmentative and Diminutive Suffixes

The key point of focus for this study, it would be useful to discuss *augmentative* and *diminutive* suffixes, both cross-linguistically and in Spanish in particular, here after discussing both Morphology and Syntax and Semantics and Pragmatics, as the juncture between these linguistic regions is their realm.

Cross-linguistically, augmentatives and diminutives are morphological forms, most often affixes, which modify the base word to primarily convey its enlargement or shrinkage, respectively, with the concomitant approbatory and pejorative connotations that the culture in question associates with largeness or smallness. English has its own class of augmentatives and diminutives, with examples like *super-* (as in *superstore*, *superpower*, *supermodel*, and *superstar*) and *-ie/y* (as in *Charlie*, *dearie*, and *doggy*), respectively.

Spanish is an especially rich language in this regard; it boasts a wide degree of often interchangeable augmentatives and diminutives, which are almost exclusively suffixes. The degree of productivity and associated approbatory and pejorative connotations vary widely from entirely neutral to almost always negative or positive, as seen in the analysis of Part III. The list of augmentative and diminutive suffixes that the study will be focused on will be explained in more detail in the introductory passage to Part II below.

### 1.2.2 – Corpus Linguistics

The nature of the questions that guided this study necessitated the availability of copious quantities of linguistic data as well as techniques and tools to help sift through these data, analyze the selections, and draw conclusions. The field of linguistics that was most equipped to meet these requirements was, of course, corpus linguistics, and we relied on Stefanowitsch’s *Handbook for Corpus Linguistics* for the practical and analytical framework for this investigation. The *Handbook* treads over much ground, but the concepts of interest to this study are those of key corpus characteristics.

The first and most important key characteristic of every corpus is its size. As with most other matters in statistics, the larger the corpus, the less prone to statistical aberrations and anomalies the analyses drawn from it are (McEnery & Hardie, 2012, p. 7). The second most important characteristic, which depends on the purpose of the corpus itself, is its composition. The composition of the corpus is comprised of the origins of the corpus data as well as the nature of the data themselves. A corpus of movie scripts should be treated differently than a corpus of transcribed emergency service calls. The composition of the corpus also determines its type, of which there are several, differentiated along numerous axes.

The first distinction is between general and specialized corpora, the difference being the scope of the composition. General corpora aim for the largest array of provenance when compiling their data, aiming to be as representative as possible, whereas specialized corpora may vary in representativeness within a given scope, such as the Michigan Corpus of Academic Spoken English. Though this corpus attempts to be as representative as possible, its limitation to Reddit makes it a specialized corpus. It is also specialized because it sources only from written material.

Another axis of comparison is between native and learner language. Though it has no official controls or filtration against learner input, the corpus is intended, and assumed, to be rooted in native linguistic information, making it a native corpus.

A key structural distinction is between monolingual and comparable or parallel corpora. Monolingual corpora, as the name suggests, contain entries in only one language, whereas parallel and comparable corpora contain information in more than one language, differing mainly in formatting. A comparable corpus, according to McEnery (McEnery & Hardie, 2012, p. 20), is a battery of corpora or sub-corpora that have similar levels of representativeness and genre balance over roughly the same period. The degree of comparability varies along a spectrum according to these characteristics, culminating in the most comparable of all corpora: the parallel corpus. This type is a corpus comprised of two sub-corpora that are essentially direct translations of each other, utilizing the same formatting and, if possible, encoding similar or analogous figurative language and idiomatic expressions. Though, as above, there is no specific control mechanism against the intrusion of other languages, the corpus is taken to be monolingual.

Corpora can also be diachronic, collecting data from across time to document changes, or synchronic, sampling only within a limited time window. This corpus, pulling social media data from 2021 only, is considered synchronic.

A balance of origins and nature of data termed representativeness should be strived for, the degree to which depends on the intended purpose of the corpus. The more diverse the origins and the nature of the data, the closer to representing the range of natural speech the corpus is, and, consequently, the more likely the data the scholar is looking for in natural speech is recorded and available for study. This concept works closely in tandem with corpus size. If a corpus is both large and diverse in nature and origin, it is more likely to contain valuable linguistic information.

## 1.3 – Literature Review

### 1.3.1 – Los llamados diminutivos y aumentativos en el español de México (Gaarder, 1966)

#### 1.3.1 §1 ­– Introduction, Materials, and Methods

Gaarder (1966) is the root of this investigation, its inspiration and template. For this reason, this study requires especially close attention, as it directly informed the shape of these experiments and thesis project. Focusing on the Spanish of Mexico, Gaarder “intends for [the study] to be a meticulous examination of the numerous [augmentative and diminutive] affixes…” (Gaarder, 1966, p. 1). Even in the 1960s, Gaarder recognizes that “[t]hese affixes…provide the Mexican…a means to express himself, a receptacle where one can discharge their subjectivism,” indicating that linguists at that time understood that the use of these affixes was not solely restricted to simply enlarging or shrinking their referents, as was the extent of their explanation in many grammatical texts from the preceding period.

He states that he would compile a “corpus,” as it would be now termed, with three principal sources of data: a specific anthology of Mexican literature over a two-century period, a selection of Mexican periodicals with a reputation of prizing direct quotation, and the anecdotal recollections and probing of the author. Next, he would locate and annotate *by hand* each instance of suffixation as either augmentative or diminutive, collect personal data about the speaker in question, and assign “nuances of meaning” according to a categorization system that was developed by the author for this investigation based on the writings of one Amado Alonso (Noción, emoción, acción y fantasía en los diminutivos, 1951). He would then compile the annotation according to speaker and then analyze the results.

#### 1.3.1 §2 – Gaarder’s Categories

Because they would directly inform the categorical structure of this thesis project’s own classification system, it is useful to examine the original categories here. They were, with his own examples:

1. Material modifications (including metaphorical cases)
	1. Reduction of size, quantity, intensity, extent, or duration
		1. Example: “He brought Juan a *little rifle*…” (Talking about a toy for the child)
	2. Increase in size, quantity, etc.
		1. Example: “…a *huge* gray *beard* was hiding his lips.”
2. Affective modifications (emotional domain)
	1. Appreciation (affection, anticipatory joy, etc.)
		1. Example: “He gave a small slap to Julian and said affectionately, ‘Now, *little friend*, to work.’”
	2. Depreciation (resentment, hatred, disgust, contempt, etc.)
		1. Example: “Clemente Ramírez, breaking down into tears and wails, reiterat[ed] his decision to immediately stab Ana María Bermejo, who that very evening had presented herself at the Salón México at the arm of a *little serviceman*, with whom she danced a ton of dances…”
	3. Condescending superiority (Not a single pure instance, that is, without any other nuances of meaning, was found of this class, proposed by A. Alonso (p. 203))
		1. Example: “Come on, *young man*…”
	4. Impertinent familiarity (another case proposed by Alonso)
		1. Example: “Hit it, because my general Natera will giving him his *little eagle*.”[[1]](#footnote-2) (this is an “impure” case.)
	5. Timidity through humility
		1. Example: “And I thought: “Well, I’ll go give the owner a *little greeting*.”
	6. Strategic timidity (false modesty, deceit, etc.)
		1. Example: “*Manuelito*, by God, come to your senses!” (Directed at a “murderer” hated by all, to calm him down.)
	7. Euphemistic softening
		1. Example: “We gave (the dog) all kinds of things to eat, until his *little stomach* was about to burst, and nothing happened.”
	8. Emphatic or attritional (real or imagined)[[2]](#footnote-3) intensification
		1. Example: “And the *very, very, very first* thing I saw…”
	9. Playful, joyful, or ironic extravagance
		1. Example: “…I was very distracted talking with a *big drunk*.”
	10. Tenderness towards the surrounding reality
		1. Example: “He would smoke his *little herb*, he fell asleep there, and some vagrants…shot him…”
	11. From habit (neutral cases)[[3]](#footnote-4)
		1. Example: “So that you can feel what the *little maize field* felt!” (There was no pure case of this.) This is considered a neutral case because the peasant invariably says *milpita (little maize field)*, no matter the circumstances.
	12. Baby talk
		1. Example: “*Daddy*, I’m hungry…”
	13. Slang[[4]](#footnote-5)
3. Approximate modifications (intellectual domain)
	1. Depreciation
		1. Example: “…he gives her the *little remedy*.”
	2. Appreciation
		1. Example: “What *great friends* he had…” (it is of the intellectual domain because those friends were not even personally known by the speaker)
	3. New voices [*sic*] following a logical basis[[5]](#footnote-6)

#### 1.3.1 §3 – Results, Discussion, and Conclusions

He found through this process 557 instances, or “tokens,” of relevant suffixation, “omitting all instances that exist as their own lexical item (p. 586),” from which he annotated 837 “nuances of meaning,” or about 1.5 “nuances” per token.

Through discussion of numbers, Gaarder deduces that (a) *-ito* and *-ón* are the most common diminutive and augmentative, respectively, and that (b) only the following “affective” categories had any real significant usage:

* Appreciation
* Depreciation
* Strategic timidity
* Euphemistic softening
* Emphatic intensification
* Playful or ironic extravagance
* Tenderness due to surrounding reality

Further, he surmises that the categories of appreciation, strategic timidity, courtesy/respect, and humility could have been folded under tenderness due to surrounding reality because they all share “the intent to not offend the listener, a certain trepidation…an attitude of vacillation, irresolution before the hardness…of the silhouette of reality, and the consequent desire to…soften this silhouette.” Thus, because most of the suffix usage falls under this new super-category, Gaarder concludes that Mexican people have a cultural need to encase themselves in a “protective sheath” against reality and use affective suffix usage as a way to soften the harshness of this reality.

Gaarder touches on the issue of whether a diminutive can be “augmentative.” A more detailed discussion of this rather outdated issue is included in 1.3.2 below.

### 1.3.2 – La función del diminutivo en español (Zuluaga Ospina, 1970)

Alberto Zuluaga Ospina, working at around the same time as Gaarder, also manually creates his own corpus, but instead chooses to focus in-depth on the diminutive suffix *-ito,* widely recognized as the most common and important diminutive suffix of the Spanish language.

He first establishes a narrowing scope of investigation, starting from all diminutives down to instances of *-ito,* then indicates that he filtered out “words that, although modified [with *-ito*]…have been lexicalized, that is to say, they have lost their diminutive meaning or have specialized or somehow restricted [it] (Zuluaga Ospina, 1970, p. 27).” He elaborates, saying that “[i]n spite of their…structure, these forms have lost their diminutive meaning in relation to their base lexeme and function as independent lexical [items]…For us, semanticity is the defining and characteristic trait of matters of language (p. 29).”

Included under Zuluaga’s “lexicalizations” are instances where the diminutive is used as its own base for further derivation, i.e., that the diminutive is “buried” within the word, such as in the progression *chico*, *chiquillo*, *chiquillada*. Zuluaga calls this process “integration (p. 30)” and explores the issue of whether integration indicates lexicalization. In any case, these instances are also discarded.

Working with the 500 instances of *-ito* this filtration process has left, Zuluaga proceeds with analysis through, as with Gaarder, the lens of Alonso (Noción, emoción, acción y fantasía en los diminutivos, 1951). He groups Alonso’s “functions of the diminutive,” the same as used by Gaarder above, by “intentional directionality,” either *towards the object* or *towards the interlocutor* (Zuluaga Ospina, 1970, p. 30)*.* However, he argues that Alonso’s characterizations of the functions of the diminutive are too shallow; Zuluaga argues that Alonso’s functions are epi-semantic. That is, Zuluaga recognizes that there is a difference between the semantic, denotative and pragmatic, connotative discursive layers, that the semantic is more “object-oriented” and the pragmatic more “interlocutor-oriented,” and that the pragmatic layer is where there is room for the interlocutor to project their “stylistic values,” as he calls Alonso’s categories (p. 30).

But the interesting wrinkle is that Zuluaga then champions the stance that the diminutive semantic meaning is always present, elaborating that:

“[the] insistence on the fundamental function of diminution [must] not be interpreted as a rejection or ignorance of connotational or stylistic values…they can make themselves present, and are explainable by normal psychological attitudes towards the small…[H]ere [it must be] insist[ed] that these stylistic effects cannot be confused with the basic, grammatical, function of the diminutive suffix… (p. 33)”

This implies that Zuluaga recognizes that an instance of diminutive use can have co-occurring semantic and pragmatic meanings. However, he insists that the semantic element is always present and is always diminution.

Zuluaga, as Gaarder independently before him, explores the idea of “augmentative diminutives,” that is, diminutives that add to the already diminutive meaning of the base lexeme. Therefore, he argues, the semantic and pragmatic functions of the diminutive can change and interact either morphologically with the meaning of the base lexeme and/or syntactically or pragmatically with the meanings of surrounding words. In this way, the ever-present semantic meaning of diminution, he posits, can be “hidden” by these interactions.

It is useful here to note that, in the years intervening between 1970 and the present, the related concepts of intensification and emphasis (as well as their negative corollaries, attenuation and understatement) have, in modern linguistics, been formally broken off from the concept of “augmentation” found in these older texts. This disambiguation is what allows for these concepts’ constellation into the categories of magnitude (collapsed from augmentation and diminution), accentuation, and minimization, respectively, in Experiments II and III below.

With the theoretical framework out of the way, Zuluaga moves on to note the allomorphy of -ito before enumerating this very detailed list of functions of the diminutive (p. 42), supposedly derived from his own observation from his undisclosed data (numbering and lettering is the work of the author):

1. Affective
	1. Derogatory
	2. “Of a weakly positive affective temper”
	3. Of such an affective value that overpowers its diminutive meaning
	4. “Of a predominant affective sense, but the diminutive sense is covered by context[[6]](#footnote-7)”
	5. “[d]iminutive function replaced by affection in nouns with predicate”
	6. “Of a definite affective sense, in vocatives”
	7. “Of a definite affective sense, in proper nouns[/personal names]”
2. Condescending superiority
3. Impertinent familiarity
4. “Son of…(name)-ito”
5. Euphemism
6. Playfulness
	1. Narrator’s smirk
	2. “Infantile grace and tenderness”
	3. Happy-go-lucky

The detail of the subcategories and the short length when compared to Gaarder’s are both noteworthy. The inclusion of identical language indicates that the same Alonso material is used as reference, and its shortened length hints that it was perhaps cut down (or that Gaarder added to his own reading of Alonso). The detail of the subcategories includes the interactions between the lexemic and suffixal meanings and their interactions; strangely, however, positive affection is broken down into categories 1b-g whereas negative affection, “derogatory,” is limited to only one category.

He concludes that, in most cases, as was expected, “the diminishing function of the suffix is evident, that is…[one can] abstract a diminutive meaning in relation with the base lexeme. (Zuluaga Ospina, 1970, p. 48)” However, with “nouns indicating person plus predicate, vocatives, and proper names, diminution is hidden and replaced with an affective meaning, which can be explained, fundamentally, by the meaning of the base lexeme (p. 48).”

### 1.3.3 – A contrastive study of English and Spanish synthetic diminutives (Hagg, 2016)

Submitted in 2016 by Anna Theresa Hagg, this master’s thesis endeavored to compare the productivity and semantic distribution of synthetic diminutives in English and Spanish from a quantitative approach that harnessed the power of corpora.

After laying out the background material, such as an explanation of contrastive linguistics and synthetic diminutive suffixes, Hagg enumerates the English and Spanish diminutive suffixes featured in the study, of which the Spanish diminutives are listed below:

* -ito
* -illo
* -ico
* -ín
* -ete

She then lays out the implications of her contrastive approach on her experimental framework and reviews previous contrastive studies between English and Spanish, indicating that none focusing on diminutives had yet been done. Hagg explained how to search the 100-million-word *Corpus del Español* and the 400-milion *Corpus of Historical American English* and did so, yielding a huge set of data that needed to be sifted through. The author then discussed the “process of sorting relevant data from the irrelevant (Hägg, 2016, p. 52).” Because Spanish is the author’s third language, dictionaries were heavily relied upon. “[L]exemes that explicitly expressed a diminutive meaning, and were formed on an authentic diminutive suffix, were swiftly ruled diminutive. Instances…which simply terminate in the same letter combinations as a diminutive suffix have of course been ruled non-diminutive (p. 52).”

In the discussion and results section, the author notes the English and Spanish (Peninsular and all South American) varieties included in the study before transitioning to the results and analyses. She splits her analyses into frequency and semantic distribution. She looked at the most common individual synthetically diminished lexemes in both languages, unimportant to this study, but moved on to describe -*illo* as the most productive of the studied Spanish diminutives, noting its strong affinity to “specialize” (i.e., lexicalize), followed closely by -ito.

Moving onto the semantic distribution analysis, Hagg lays out “the five semantic categories:”

* Smallness
* Hypocorism
* Pejoration
* Specialization
* Ambiguous

This list of categories is much, much less extensive than that of the Gaarder study, and includes one, Specialization, that expressly holds non-compositionally diminutive base lexemes, and another, Ambiguous, that, by its very nature, will be inscrutable from an analytical perspective. This list implies that Hagg, like Zuluaga, distills the functions of the diminutive into a small handful of broader categorical functions. It is important to note that Gaarder, too, concluded a collapse of the many specific Alonso categories into fewer general ones, and that the small size of the list may be to make crosslinguistic comparison between English and Spanish easier.

Utilizing this small list of semantic categories, Hagg presents the result that lexicalized instances (“specializations”) comprise almost 60% of Spanish diminutive use. Of the semantic categories with meaning, Smallness and Pejoration have similar frequencies, with Smallness only more popular by a single percentage point (16% and 15%, respectively). Hypocorism comes a distant last, at about 10%.

Through comparison and contrast of the data for English and Spanish, Hagg concludes that English diminutive usage is more vibrant than many give it credit for, but that Spanish remains far and away the richer language in this aspect.

### 1.3.4 – Variación y variedad del diminutivo en español a través de dos corpus originales (Criado and Andión, 2016)

This study mines two original custom-made corpora created from 970 written newspaper articles and six and a half hours of cooking program to characterize the diminutive suffix for the betterment of pedagogy in Spanish as a Foreign Language.

In reviewing prior literature, the authors point out that “[r]egrettably, we could not find a pan-Hispanic study that delved into the behavior of the appreciative diminutive suffixes in the different geographic regions,” and that, because of differing methodologies, direct comparison between studies that characterize diminutive usage in particular areas yields little results. The authors detail the geographic distribution and extent of the various Spanish diminutive suffixes (as done in Section 3.1.1 below), as well as some of their distinctive regional behaviors, including a small note about “reduplicated variant[s].”

When discussing the creation of their proper corpora, the authors discuss assuring representativeness by including a “diversity of genres…and registers,” and by ensuring proportionality of contribution. With this in mind, the authors compiled the materials attentively, indicating that they also referred to the much larger preexisting corpora like the *Corpus de Referencia del Español Actual* (CREA) and the *Corpus del Español del Siglo XXI* (CORPES), selected for their size and diversity. They drew materials from Spain, the US, Cuba, Puerto Rico, Venezuela, Colombia, Chile, and Argentina, to name a few locales.

They reveal that 95% of the written data is from 2012-14, of which only 3% did not contain a single diminutive of any kind. Through frequency analysis, the authors determined that diminutives were used much more commonly outside of Spain than within, and that, regardless of location, the most common ending is -ito. All other endings saw far higher rates of incidence within Spain than without. “The functional variety of the diminutives is wide and difficult to delineate,” they write, concluding that their use falls into roughly five categories: diminishing, affective, ironic, emphatic, and respectful. Of note in this analysis is the conclusion that ironic use of the diminutive was most common in Spain (being its most common function there) and uncommon without, almost unheard of in Mexico or the Caribbean. In all other regions, the affective use is predominant, most especially in the Caribbean, where it comprises a whopping 83.5% of diminutive use. The oral corpus yielded nearly the same results, with the added peculiarity that the ending -*illo* did not appear in any linguistic area outside of Spain at all.

The study concludes by presenting a graduated pedagogical list for the different proficiency levels of Spanish learners to better instruct them on their use.

### 1.3.5 – “Sólo un poquito” El uso y funciones del diminutivo en español peninsular en dos grupos de Facebook (Maíz-Arévalo, 2018)

Similarly to the previous studies, the author of this study constructs a corpus from digital data, but, crucially, this data comes from social media. In this case, from Facebook. This is to see how the usage of these suffixes changes between different media, as half of the corpus is from an entirely online Facebook group whereas the other is from another group where the members have relationships both on- and offline in Spain. The author collected posts and first comments from the two Facebook groups over a period of two months. One group is a public page dedicated to a Spanish singer, the other a private one comprised of the author’s friends and family, centered in Madrid. She describes the ethical dilemma and the process of anonymizing data. The resulting corpus consisted of 8,088 words in “67 conversational exchanges (Maíz-Arévalo, 2018, p. 37).”

Four categories were set up to annotate the diminutives’ functions: affection, pejoration, attenuation, and familiarity. The following suffixes were mined for:

* -ito
* -ete
* -illo
* -uelo

Among the anonymous online-only Facebook users, diminutives were very rare and solely comprised of -ito instances, of which almost 80% consisted of attenuation. The mixed-medium corpus, on the other hand, utilized diminutives at five times the rate, using all suffix types (of which -ito was far and away the most used, -uelo least) and functions (of which affection was most common and pejoration least).

The author then concludes that personal close relationship is directly related to the productivity and variety of form and function of diminutive endings in at least Peninsular Spanish.

### 1.3.6 – Gaps in the literature and the need for this paper

As seen in the studies covered above, the vast majority of the literature focuses either primarily or entirely on the functions and roles of the diminutive, probably because of, as Jurafsky describes, “the extraordinary, often contradictory[,] range of its senses… (1996, p. 1)” Even the work the seminal study was based on dealt primarily with the diminutive. In keeping with the seminal study’s spirit, the author intends to extend the methodology and focus on the diminutive to its logical opposite, the augmentative, along with augmentative- and diminutive-adjacent suffixal categories.

Beyond this, the older literature lacked substantively in volume of data, which often led to compromises, such as extending the window of data not only to years, but decades and, in the case of Gaarder, two centuries. The creation of the subfield of corpus linguistics has allowed for the exploration of thinner and thinner slices or snapshots of linguistic usage, such as the one I wish to explore, the use of Spanish across the globe on Reddit.

## 1.4 – Thesis Project Structure

The thesis project will be broken into three separate but interconnected parts. The first will be to establish a corpus from which all the data will be taken and analyzed for the subsequent experiments. The second will be the testing of Gaarder’s classification system as he began his experiment (i.e., without the modifications he recommends in concluding his study) and analyzing its performance for improvement. The third and final portion will be the utilization of this categorization system to annotate portions of the corpus and to analyze the usage patterns of the designated endings.

# 2 – Part I: Corpus Creation

## 2.1 – Background Information

### 2.1.1 – Regions of Study

The Spanish-speaking world is vast, being the second most natively spoken language in the world by some estimates, after only Mandarin Chinese and beating out English by one hundred million speakers according to those same estimates. Though the geographic distribution of its use is not as vast as that of English, the area of native Spanish speech is much bigger, leading to a higher number of broad accents and regional dialects, which have already been divided by other scholars into eleven distinct dialect groups, namely (in geographic order from northwest to southeast) (Lipski, 2012, p. 3):

* Mexican
* Caribbean
* Central American
* Highlander
* Chilean
* Rioplatese
* North Peninsular
* South Peninsular
* Islander
* Guinean
* Philippine

Of these, the Guinean and Philippine dialect groups exist in countries where the Spanish language either no longer occupies a central governmental or cultural role or the speaker group is too small or poor to have a significant social media presence.

Because of resource constraints, a seminal assumption was necessary to make the data manageable. This assumption had wide-ranging implications throughout the design and execution of the study, detailed in further detail in the sections below. This central assumption was that **speech is uniform between speakers within and throughout the geographic and virtual spaces occupied by speakers of said dialect group**. This core premise is, of course, accepted for convenience and is not considered axiomatic. It would not have been made or operated upon had the resources available to the study been infinite, but the adoption of this foundational supposition facilitated the manipulation of several key properties, the first of which is the corollary that *the speech of each dialect group can be accurately represented by the citizens of a single country*. Furthermore, the nature of the data, explained in more detail below, necessitated an additional, related presumption: that *speech between citizens of a country is also uniform*.

These assumptions combined enabled the selection of one country as the representative of the speakers of the entire dialect group but also forced the normalization of speech within each nation’s borders by the selection of one dialect group above any others, should they exist in a single country. A dialect was only chosen in countries in which a dialect from one of the comprising dialect groups enjoys an either *de jure* or *de facto* standard status, otherwise the country was rejected outright. Thus, while the institutions of the Royal Spanish Academy ensure that the data from Spain is regarded as representative of North Peninsular, those from the United States were excluded from the study, given that no dialect in the nation enjoys any standard status, neither *de facto* nor *de jure*.

The final list of regions, and their respective countries, studied was therefore:

* Mexican - Mexico
* Caribbean - Dominican Republic[[7]](#footnote-8)
* Central American - Guatemala
* Highlander - Colombia
* Chilean - Chile
* Rioplatese - Argentina
* North Peninsular – Spain

### 2.1.2 – Other Corpora

The fundamentals of corpus linguistics having already been established in Section 1.2.2 above, a decision needed to be made as to which corpus should be the base of the study. Gaarder’s own corpus, limited by the technology and knowledge of the time, was assembled by hand, and was based upon newspaper articles and magazines, but mostly consisted of just three literary anthologies, comprised of works spanning two or three centuries. Though it can be argued that this was representative, the small sample size was susceptible to statistical anomalies and outliers, particularly because the data seems to have been collected and annotated by hand. Zuluaga’s own primitive corpus has not, to the author’s knowledge, been released, precluding its use.

In the years intervening between the work of Gaarder and Zuluaga and this thesis project, a wide array of readily available Spanish-language corpora has made itself available to the linguistic scholar, both on their own websites and hosted on SketchEngine, the corpus compilation and analysis tool chosen for this project (rationale detailed further below). It would be prudent at this juncture to examine them in detail and discuss why they were not appropriate for use in this thesis project.

#### 2.1.2 §1 – Independent Corpora

##### Corpus del Español – dialectal (CdEd)

This corpus is quite large, clocking in at slightly over 2 billion words from web pages representing 21 Hispanophone countries. This corpus comes equipped with a wide variety of analytical and presentational functions, however the search interface is less than user-friendly for anyone without a computational background, requiring some adjustment. Furthermore, while the data itself was collected in a relatively narrow band of time (2013-14) and is from media informal enough to likely yield novel forms (further articulated in Section 2.2.2 below), the data’s age (9-10 years at the time of writing) was less than ideal. In addition, there was no way to easily collect lexicalizations and coincidental forms into exclusion lists that could be easily filtered out during the initial search (more on this in Section 3.2.1 §1 below).

##### Corpus del Español – NOW (CdEN)

Part of the same system as the CdEd above, the CdEN is a standalone corpus along the same vein. Comprised of an impressive 7.3 billion words drawn from web pages from the same 21 Hispanophone countries as above, the interface and analytical tools remain the same above, as do their limitations. Although much larger than the CdEd above, the data originates from a wider band of time (2012-19), though this is mitigated by the granular search filters that can home in on data from individual months and days. Regions can still be filtered for, but easy exclusions remain impossible, and the data remain of suboptimal age.

##### Corpus del Español del Siglo XXI (CORPES XXI)

Published by the Real Spanish Academy (RAE), this corpus is smaller than the CdE corpora, clocking in at 381 million words. However, the major selling point of this corpus is the increased granularity in annotation, though this feature is irrelevant to this particular study. Similarly to the CdEN above, the data comprising this corpus as of this writing was collected from a wider band of dates (2001-2022), this was not the case in the version available at the beginning of this project, which had only collected data until 2020. In addition, the vast majority of the corpus data remains from between 2006 and 2011.

Though the inclusion of *oral* transcripts would have been a major advantage, there was no effective or efficient way to sort by or filter out certain geographic regions, and the capability to easily exclude words remained elusive.

##### Corpus de Referencia del Español Actual (CREA)

The smallest of the reviewed corpora, the CREA, containing 155 million words from each of the Hispanophone countries, included oral transcripts, the ability to filter regions, and the ability to filter by date. However, the date range for this corpus is the largest of those examined (1975-2004), and data towards the upper limit is nearly a decade old.

##### Summary

A table is included below to summarize and visualize the characteristics and features of each corpus:

Table – Characteristics and Features of Independent Corpora

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Corpus*** | *Size* | *Age* | *Oral/Web?* | *Sort Dates?* | *Sort Region?* | *Exclusions?* |
| **CdEd** | 2 B | 9 y | Y | N/A | Y | N |
| **CdEN** | 7.3 B | 4 y | Y | Y | Y | N |
| **CORPES** | 381 M | 3 y\* | Y | Y | N | N |
| **CREA** | 155 M | 9 y | Y | Y | Y | N |
| **Sketch.** | (9.3M) | (1 y) | Y | Y | Y | Y |

\* - as of 2023, using version available Jan 2020

#### 2.1.2 §2 – SketchEngine Corpora

As seen in the summary table above, SketchEngine maintained all these capabilities, but it itself hosted an array of Spanish-language corpora, which will be tabulated in a similar manner below.

Table – Characteristics of SketchEngine Corpora

|  |  |  |  |
| --- | --- | --- | --- |
| ***Corpus*** | *Size* | *Age* | *Sort by Regions?* |
| **Span. Web ‘11** | 9.5 B | 12 y | Y |
| **Span. Web ‘18** | 16.9 B | 5 y | Y |
| **Elexis Span Web ‘20** | 1 B | 3 y | N |

The table above illustrates that, although each of these corpora was close to ideal, Spanish Web 2011 was far too old for our purposes and Elexis Spanish Web 2020 could not be sorted through by geographic region or dialect. Spanish Web 2018, which did meet most of the criteria and whose data was passably recent, was, however, largely comprised of non-conversational web data, that is, online publications of varying degrees of professionalism. Thus, the need to create SpanRed was recognized.

### 2.1.3 – SpanRed

Though many of the SketchEngine corpora above came close to meeting criteria, none was sufficiently recent while also being able to sort by dialectal region. This called for the creation of an original corpus, which was prepared according to the following scheme: the medium was decided upon, the programming language and ancillary programming packages were chosen, targets were pre-surveyed and selected, and the data were collected and compiled.

#### 2.1.3 §1 – Medium and Programming Language

Social media were decided to be the optimal medium for harvesting for this project. Easily machine-readable and widely available to the public, it was known that there also existed a large amount of data collection operations occurring outside of linguistics and that this infrastructure could be utilized to speed up the corpus creation process. Social media draw from users of widely varying backgrounds, thus making representativeness easier to address. We explored numerous popular social media for harvesting and settled on Reddit for both practical and theoretical reasons. Reddit is an online forum/social medium , subdivided into subforums that made it easy to identify specific forums in different geographical regions. Its user interface made it relatively straightforward to scrape data from it. Finally, Reddit is almost entirely comprised of posts and comments to those posts. This format often results in lengthy back-and-forth exchanges, exactly the type of conversational situation in which we hypothesized augmentatives and diminutives would be frequent.

The programming language R and Reddit were ultimately chosen due to two factors: A) the work by Levshina, *How to do Linguistics with R*, was instrumental but dealt solely with the R language, and B) the R library packages for Reddit were much simpler than for other social media platforms, such as Twitter, which are used to more complex algorithmic analytics, and thus were much harder to learn. These platforms also required an application process to collect their data, something that was not required by Reddit, as Reddit does not seem to collect as much personally identifiable metadata (or at least does not make these harvestable with comments) as any of the other platforms, being a forum.

#### 2.1.3 §2 – Targeting and Survey

The portions of the vast Reddit forum included in the study were surveyed and cleared for harvesting by taking advantage of the formatting of the site to find where the largest representative populations from each country were likely to be active according to a set of assumptions, some informed by the decisions abovementioned and others by online experience.

The base unit of Reddit is the post. It consists of a title, usually accompanied by a photo, video, or a block of body text. Beneath this “original post” is a branching tree of comments and sub-comments. When the comments enter a loop of replies and responses, essentially a conversation, this is known as a (comment) thread; these are the main content of Reddit and comprise the harvested material for this corpus. These posts are collated into “subreddits,” or collections of posts, almost always related or adjacent to a central theme, with which the subreddit often shares its name. Reddit users, known as Redditors, can subscribe to as many subreddits as they are interested in for updates. This format has a couple of implications that can constitute the bedrock for some convenient assumptions and be harnessed for directed data harvesting.

First, the fact that subreddits are subscribed to almost always leads to most of the threads of posts in the subreddit being participated in by subscribers. The assumption can be made, then, that the participation level of non-subscribers will be low, and that most of the linguistic data will be produced by the subscribers, essentially all of it. An exploitable consequence of this, and the fact that subscription is an opt-in process, which implies interest in the theme upon subscription, is that most linguistic data in posts of a certain theme will belong to subscribers. If these themes are strongly connected to the identity of a certain people group, it would then make sense that the majority of those interested would be those who identify themselves with this theme, i.e., belonging to this people group.

Second, and contingent on the previous assumption, is that the subreddits concerning strong but general identifiers of a country, most of all the subreddits claiming to center upon and/or represent said country, will largely contain comment threads authored by subscribers, who are likely to be citizens of the country represented. Thus, the lion’s share of linguistic data in these subreddits is assumed to have been authored by citizens of the country and thus the threads of this subreddit are assumed to be of the dialect of interest.

It may be, as it is inevitable on such a format, that there exists cross-pollination between subreddits, and thus trace amounts of linguistic data from dialects other than those studied (or of other studied dialect groups) may be found in a given sample, but under these assumptions it unlikely that these cross-pollinations are of real statistical significance, especially with increasing sample size.

Though the above assumptions were followed, they were, as with all assumptions in this study, taken as tools of convenience, not as axioms. Knowing reality to not be as simple as the assumptions establish, in the interest of representativeness, the designation of subreddits for harvest was kept as general as possible, restricted almost entirely to country subreddits in the interest not only of sample size, as these subreddits tended to be quite active and informationally large, but also of representativeness, again, to limit susceptibility to local statistical anomalies and the over- or underrepresentation of certain speech groups within the featured countries.

### 2.1.4 – Sketch Engine Compilation

In addition to being the compilation tool with which the advisor to this project is most familiar, Sketch Engine was selected because it offers excellent, user-friendly search and analytical functionality and an extensive library of corpora, large or small, corporately funded or individually sourced, and in many languages, to which the corpus used for this project could be added on its conclusion. The corpus involved in this thesis project could also be, and is planned to be, hosted on SketchEngine for future independent linguistic study.

## 2.2 – Materials and Methods

### 2.2.1 – Survey Method

The URLs for the scraped subreddits will be provided in the Appendices but consisted entirely of those dedicated to the appropriate nations (i.e., r/Colombia, r/Mexico, etc.). The subreddits were pre-surveyed qualitatively for the existence of “tourist fora,” or sections for answering tourists’ questions, and also for visibly significant intrusion of foreigners and tourists. Luckily, this was not seen to be the case in any of the main subreddits for the studied countries, so there was no need to deviate to another, smaller, subreddit.

### 2.2.2 – Optimization Algorithm

Due to hardware limitations, an optimization algorithm was developed to maximize the efficiency and effectiveness of the harvesting process, synthesized from both the assumptions highlighted above and lessons learned from the better part of two decades in online environments like Reddit.

The first axiom is that not all speech is created equal; that is, the quality and/or density of linguistic information derivable from every utterance is neither uniform nor is it assured to be valuable. Each utterance may, in fact, contain linguistic information not obvious at first, uneducated glance, but the density of this information as well as its value to the linguistic scholar, i.e., its relevance to their investigative aims, varies wildly depending not only on the individual scholar’s aims, but also on a wide array of the data’s characteristics, content, and context. Because this study focuses on the regional characteristics of productive augmentation and diminution, fertile linguistic conditions are sought after, which leads to the following rationale.

According to experience, single-comment threads are often just a superficial reaction to the contents of the original post, often just a phrase like “me, too!” It is generally when threads grow into ever longer conversational chains that more substantive conversations, likely with higher word counts, are found. And even within these longer, more likely to be linguistically fertile, conversations, low-word count responses typically contribute little and are in any case less likely to carry evidence of active linguistic processes.

But thread length is by no means a guarantor of linguistic quality. A large fraction of internet discourse, including initial engagement with original posts, been extensively meme-ified. That is, comprised of a byzantine semiotic parlance of heavily self-referent, incredibly context-dependent, templatic, and often grammatically anomalous language and multimedia gags riddled with niche cultural references, and which chronically have a noticeably short shelf life. “First responders” often submit comments directly under the original post snugly couched in this peculiar jargon, contributing to the reputation of the poor linguistic quality of single-comment threads, but this unique facet of internet culture, on Reddit, also chronically spawns highly ritualistic discourse chains centering on prefabricated phrases and call-and-response activities. A common example is reciting the internet memetic sensation “Never Gonna Give You Up” by Rick Astley, with Redditors taking turns contributing a single word at a time.

Thus, experience informs the first assumption that the larger the “area” of the thread, the more linguistically valuable the linguistic information contained within. That is, the higher the number of comments in a thread (i.e., height), while also having a higher average per-comment word count (i.e., width), the more likely the content within is a substantive, linguistically active conversation. This metaphor also extends to the third dimension: the larger the “volume” of the post (i.e., the larger the number of large-area comment threads the post contains), the more linguistically valuable information it is likely to contain. The ideal would therefore be to devise an algorithm[[8]](#footnote-9) to measure a combined statistic of average word count per thread comment and the length of said thread, then an average combined statistic for all the threads in the post. However, again due to constraints of time, labor, hardware, and knowledge, this ideal was not feasible, so a backup process approximating this ideal needed to be devised, using another, more readily available, metric.

The second assumption, related to the first, is that comment count, that is, the exact number of comments under a certain post, could roughly address the concerns above and function as a makeshift approximation of the linguistic quality of the post and its comment threads. The more comments a post contains, regardless of their structure, the more words the comment threads are likely to contain on average. The corollary third, and far more tentative, assumption is that the more comments a post contains, the more likely at least some of these comments will have arranged themselves in the linguistically fertile conversational format that makes finding active linguistic processes more probable. Thus, armed with a metric for approximating the likely volume of linguistically valuable information in each post, the scraping process could be proceeded with.

To scrape a subreddit, one must enter the name of the subreddit as well as a small set of parameters into a special script set up in R. This script then goes to Reddit and then the subreddit and collects a list of URLs that meet given parameters. The user then specifies which URLs the script should scrape data from, which it then does and spits out data and metadata onto a document, from which the user can draw whatever selection they desire.

The first process was run without any issue. The designated subreddits were combed for candidate URLs, which were collated into a table with various information, of which, as stated above, the primary concern was comment count. These URLs were then arranged by comment count, representing the number of comments in the given thread.

Due to the theoretical rationale detailed just above but especially technological limitations that precluded the use of all available data, a comment number threshold was implemented to ensure high linguistic quality in the collated data as well as the timely execution of the corpus creation process without extensive equipment damage. This threshold was not uniform but rather was determined for reach region individually according to the volume of available data (i.e., the comment threshold for Mexican URLs was much higher than that for Guatemalan URLs). URLs that met the threshold were scraped for linguistic data while those that did not were discarded. The thresholds are tabulated below, with the algorithm for its determination detailed in the Appendices.[[9]](#footnote-10)

Table – Cutoff Thresholds of Dialectal Sub-corpora

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***Dialect*** | **Argentina** | **Chile** | **Colombia** | **DR** | **Guat.** | **Mexico** | **Spain** |
| *Threshold* | 75 | 50 | 100 | 10 | 20 | 150 | 20 |

### 2.2.3 – Compilation

The above process resulted in text files that were fed directly into the user-friendly corpus creation tool featured on the SketchEngine site. All compilation, sorting, and tagging was directly and automatically handled by the site.

## 2.3 – Results

The corpus is a monolingual, synchronic, specialized Spanish-language corpus sourced entirely from 2021 Reddit posts by Spanish speakers from around the world. For reasons explained in Section 2.1.1, most or all these speakers are expected to be native Spanish speakers and likely of the Spanish dialect dominant in the targeted country. It holds 11.5 million tokens, of which 9.3 million are recognized as linguistically significant words, organized into 350,000 sentences. The final breakdown of tokens among the geographic origins is as follows:

* Argentina – 1.53 million (13.3%)
* Chile – 0.698 million (6.1%)
* Colombia – 1.75 million (15.2%)
* Dom. Rep. – 0.470 million (4.1%)
* Guatemala – 0.367 million (3.2%)
* Mexico – 5.18 million (44.9%)
* Spain – 1.53 million (13.3%)

## 2.4 – Discussion and Conclusions

A qualitative review of the corpus contents gives the impression that the vast majority of the one-off and the cyclical call-and-response memes that otherwise would have infiltrated the corpus was largely successfully filtered out, as the levels were considered noticeably lower than considered normal by the author. While the author cannot prove qualitatively the absence of these memes, there is some evidence to suggest that the filtration was not total, as this large sequence of comments solely comprising of the letter F demonstrates[[10]](#footnote-11):

|  |
| --- |
| * F
 |
| * F
 |
| * F
 |
| * F
 |
| * F
 |
| * F
 |
| * F
 |
| * F
 |
| * F
 |
| * F
 |
| * F
 |
| * F
 |
| * F
 |
| * F
 |
| * F
 |
| * F
 |

With regards to the substantivity of captured conversations, this qualitative analysis further suggests that the assumptions delineated in Section 2.2.2 were largely correct, as captured by this small exchange discussing the effects of food and nostalgia:

|  |
| --- |
| * It is the great power of food: it is memory and is tangible. I’ve wanted to make a cookbook with my grandmother[. S]he is 90 years old, and I don’t know how much more time she has and I want to take advantage of each and every minute I can while she’s here.[[11]](#footnote-12)
 |
| * + D\*mn, I just lost my grandma. I lived and grew up with my grandmother for a time because my parents worked. I miss her food a sh\*t-ton…[unintelligible]…how you got along with your granny, it’s something you’ll never have again. Enjoy your grandma.[[12]](#footnote-13)
 |

Again, a few counterexamples, nearly all memetic as demonstrated in the counterexample in the paragraph above, can be found, but a similarly healthy level of substantivity can be found in most conversations in the scraped data.

The targeting scheme detailed in Section 2.1.3 §2 was, again by qualitative analysis, largely successful in restricting the vast majority of linguistic data included in the dialectal sub-corpora to the intended dialect, as words and phrases such as “qué pedo” and “tío[[13]](#footnote-14)” are largely limited to their respective dialectal sub-corpora, the Mexican and the Spanish, respectively. However, as with the other filtration systems above, success was not complete, as this Mexican Spanish speaker posting in the r/Spain subreddit shows:

* Hahahahha, no puedo crearlo. ¡¿Qué pedo?! No mames me doy un montón de risa.[[14]](#footnote-15)

There were various examples and degrees of linguistic contamination from English as well, such as in the example provided for conversational substantivity above. However, because of English’s global status as *lingua franca* as well as its inextricable link with meme culture and internet culture in general, the total exclusion of English was never considered remotely feasible.

# 3 – Part II: Replication Experiment with Gaarder’s System

## 3.1 – Background Information

### 3.1.1 – Endings

The Spanish language constrains its strategy of augmentation and diminution largely to synthetic suffixation, and yet it, being a Romance language, is equipped with a wide variety of endings to do so. There is no wide consensus on which endings are currently used; the question hinges on what constitutes a different ending (allomorphy) as well as the counting of historical suffixes (productivity) and regional variants (regionalism).

As anyone moderately familiar with Spanish may know, the language utilizes a gender-based agreement system that is utilized in adjectives as well as nouns. This generally results in allomorphic pairs of endings, the masculine generally ending in -*o* and the feminine in -*a*. For simplicity and brevity, the feminine allomorphs have been folded under the masculine henceforth unless otherwise stated, as in per convention in Spanish-language dictionaries and most modern linguistic literature.

The set of endings studied here is based on that used by A. Bruce Gaarder in 1966, modified with the passage of time and the abovementioned factors in mind. The original set consisted of:

* -ito
* -cito
* -illo
* -cillo
* -ín
* -uelo
* -ejo
* -udo
* -acho
* -ota
* -zota
* -ón
* -ijo
* -ola
* -zaso
* -ajo
* -ote
* -cejo
* -ucho
* -azo
* -osa
* -ullón
* -oide
* -uza
* -uno
* -erío
* -iza
* -ana
* -ales
* -ada
* -ulón

In addition to its unwieldy length, the list includes several instances of context-based allomorphy, such as -*cillo* and -*illo*. Of this pair, -*illo* is arguably the base form, with -*cillo* appearing in certain contexts, not important for this study. Other endings are simply patently not what was being looked for, such as -*oide*, analogous to English -*oid*, or far too likely to be confused for encoding some other semantic meaning for large-scale annotation to be time-efficient, such as -*osa*, analogous to English -*ous*, and -*al(es)*, analogous to English -*al*. To reduce the scope of the project, a mix of the endings already provided along with the addition of other augmentative- or diminutive-adjacent endings were selected and the others discarded.

The chosen endings were thus:

###### Diminutives

* -ito
* -ico
* -illo
* -ín
* -iño
* -uelo
* -ejo
* -ijo
* -ete

###### Augmentatives

* -ote
* -ón
* -azo
* -udo
* -ucho
* –(er)ío
* -che
* -ísimo

One will notice that both Gaarder’s list and the list in this study include endings which are not normally considered “augmentative” or “diminutive,” but rather “intensifying,” “affectionate,” or “pejorative.” This is by design. There was some evidence presented, in Gaarder, Hagg, and Zuluaga specifically, that many suffixes of “adjacent” categories are often used to modify size, and many size suffixes are also utilized in “adjacent” roles, such as intensification and pejoration. These facts are also reflected in the rationale for later systems mentioned further below.

Other endings, namely -*che*, -*ico*, and -*iño*, were added in the interest of inclusivity and representation; these endings are characteristically regional, exist in at least one of the regions added, and ought to be compared to and contrasted against the more mainstream suffixes.

### 3.1.2 – Semantic Categorization

To draw any statistically testable and systematic conclusion beyond simple frequency analysis, a framework of annotation needs to be developed and underpinned by an abstract theoretical rationale. Because meaning and usage in context are this study’s key areas of interest, the rational used is founded in the fields of corpus linguistics, (lexical) semantics, and pragmatics, influenced also by morphology and conversational analysis.

Previous studies have concluded that augmentative and diminutive endings, both cross-linguistically and in Spanish in particular, are both incredibly productive and often highly expressive, the extent to which has, to the knowledge of the author, not been thoroughly quantified but which has nevertheless been widely observed. This high degree of productivity and expressivity lends itself to a wide variety of denotative, connotative, and pragmatic meanings, which have often been observed as cooccurring, thus leading to the observation that augmentative and diminutive endings often carry both semantic and pragmatic meanings and functions (Augustyn & Gniecka, 2011, p. 32).

A systematic quantification of these semantic and pragmatic meanings has been sought after for quite some time, never quite having been perfected. Gaarder utilized a sprawling but disorganized scheme in 1966, based on the earlier work of one Amado Alonso and geared towards exhaustiveness. Translated into English, this is the scheme upon which this study’s is based. The system employed in this study, detailed below, is primarily based on the Gaarder scheme.

To understand the modern incarnation, one must understand the Gaarder model. This system was, as stated before, built upon the work of the aforementioned Alonso, and was an attempt to enumerate the different functions of the diminutive and the augmentative in the written language. Because it was based on previous, non-systematized understandings of the role of augmentatives and diminutives in actual usage, the categorizations are haphazard and sometimes overlap, as shown in the explanatory material below.

Gaarder split the functions of the augmentative and diminutive endings into the *intellectual* and *emotional domains*, or, more roughly, between “objective” and “subjective” semantic functions. Using the diminutive -*ito* in *perrito* to mean a physically small dog (e.g., a Chihuahua) would be an example of an “objective” usage of the diminutive in the *intellectual domain*. By contrast, using the same ending on the same word but referring to a dog the speaker finds cute would instead be an example of a “subjective” usage in the *emotional domain*.

Intellectual categorizations were broken down into two further subdomains, the *physical* and the *evaluative*. The physical subdomain was self-explanatory. The categories within are *increase* or *decrease*. The first example above of the intellectual domain is an example of the physical subdomain, specifically the increase categorization. The evaluative subdomain was less obvious, differing between *positive* and *negative*. It was used for when the subject used their intellect to deduce whether the modified noun was good or bad and used a diminutive or augmentative to convey this evaluation.

The emotional domain was much more varied, containing thirteen different categorizations: appreciation, depreciation, impertinent familiarity, respect and courtesy, humility, strategic timidity, euphemistic softening, intensification, ironic or playful extravagance, tenderness with respect to surrounding reality, and baby talk. Gaarder also included a catchall slang categorization for instances he couldn’t explain. Some categorizations were more self-evident than others.

*Appreciation* represented like or affection towards the given object, whereas *depreciation* represented the opposite. *Impertinent familiarity* was used for a speaker using a familiar tone disrespectfully. *Humility* and *strategic timidity* were similar, with the distinction being that strategic timidity was used to manipulate to elicit a specific outcome. *Respect and courtesy* was used for deference usually towards social superiors. *Euphemistic softening* was often used to contrast against *intensification*, which itself was differentiated from *playful or ironic extravagance* by the intent: the former to emphasize and the latter for comedic or dramatic effect. *Tenderness with respect to surrounding reality* just encapsulated sympathy and empathy, which is what the category was renamed to in this experiment.

As one can easily see, some of these categories overlapped significantly with others. The overlap was something Gaarder himself mentioned in his paper, but it was decided to test his categories as is in this first experiment, and then use the results to craft a new set of categories in the second experiment. Based on feedback from annotators and to make some names more succinct, some category names were adjusted. The original name follows in parentheses.

* Increase
* Decrease
* Positive
* Negative
* Appreciation
* Depreciation
* Condescension (Condescending Superiority
* Impertinent Familiarity
* Respect/Courtesy
* Humility
* Strategic (Strategic Timidity)
* Euphemism (Euphemistic Softening)
* Intensification
* Extravagance (Playful or Ironic Extravagance)
* Sympathy/Empathy (Tenderness with Respect to Surrounding Reality)
* Baby Talk
* Slang

### 3.1.3 – Three Primary Layers of Meaning

A concept mentioned in the Categorization Guidelines, available in the Appendices below and prudent to review here, is the idea of **layers of meaning**. It is related to, and was used pedagogically with the annotators instead of, the Shingles of Linguistics (section 1.2.1), which can be visualized thus:

* Meaning of **utterance**
* Meaning of **sentence**

Meaning of **word** The explanation below is extracted from the Categorization Guidelines, included in Appendices:

In this study, we are focusing on how the addition of augmentative and diminutive suffixes affect the meaning of keywords. This is not to be confused with the overall tone and/or meaning of the sentence or the greater utterance, concepts which are related to, and influence, the meaning of keywords and the selection of applied suffixes. You will see examples in *Examples from the dataset* that illustrate quite clearly a separation between the three layers of meaning, their separation in practice, and getting to the root question of **“what is the difference between the suffixed and un-suffixed word in this context?” Or “why did the speaker choose to use the suffixed word instead of the un-suffixed word?”**

However, a useful example of that separation in the meantime is the sentence: *No me gusta el* ***perr-ón****, pero el* ***perr-ito*** *sí es muy lindo*. “I don’t like the **big dog**, but the **little dog** is very sweet.” The significance of the sentence is about dislike and affection, two of the categories we will be identifying, but the speaker uses the augmentative and diminutive suffixes to differentiate between the two dogs using AMPLITUDE: the feelings of dislike and affection stem from other sources within the sentence.

## 3.2 – Materials and Methods

### 3.2.1 – Data Sets

#### 3.2.1 §1 – Lexemic Sortition

Following in the footsteps of Gaarder, Zuluaga, and Hagg, lexemes made independent through lexicalization needed to be filtered out to leave only linguistically productive instances of augmentative and diminutive suffixation for analysis. Because the exact methods are enumerated in neither Gaarder’s nor Zuluaga’s studies but are in Hagg’s, the sortition method used here strongly resembles hers.

Utilizing SketchEngine’s wordlist search function, the corpus was sifted for incidences of nouns or adjectives ending in character sequences matching each allophone of the studied augmentative and diminutive endings. SketchEngine yielded large wordlists comprised of mixtures of coincidental matches (usually typos or verbs and adverbs missorted as nouns or adjectives), lexicalizations, and truly productive instances. Paralleling Hagg, these wordlists were manually scoured of lexicalizations and coincidences against four online Spanish resources (given in order of precedence): SpanishDict, Wiktionary, Google Translate, and Google Search. If a featured lexeme was found on any of the resources, starting with SpanishDict, the entry was immediately discarded to an *exclusion list*, a text file of specific words to be filtered from SketchEngine’s results when performing retrieval of the augmentative and diminutive instance data. If exclusions inundated the author, an “inclusion list” of words to be specifically sought out by SketchEngine was instead compiled. These exclusion and inclusion lists were by no means perfect or definitive and examples are available for review in Appendices.

#### 3.2.1 §2 – Data Set Creation

Wordlists on SketchEngine are sorted from highest to lowest by number of matching tokens. With these inclusion and exclusion lists in hand, each of the sub-corpora, corresponding to each of the studied dialectal regions, was combed through for tokens with productive augmentative and diminutive endings. The data arrays were constructed from a maximum of 10 concordances whenever possible from the top 10 most frequent keywords, repeated for each sub-corpus, yielding a theoretical maximum of 700 concordances, split into two data arrays.

### 3.2.2 – Annotation

Double annotation was performed by four annotators, all four being native speakers from the Mexican dialect region, with pairs each annotating half the total data. The annotators were trained for one hour according with the coding packets included in Appendices. Utilizing both this training, in which the annotator was, as stated above, urged to attempt to isolate and evaluate the function of the diminutive itself, as well as the quick reference coding key in the back, the annotators evaluated and annotated their assigned data arrays.

For reference, this is the view of one of the concordances given to an annotator for “perrito:”

Figure – Example of Concordance

One may note that there is a significant amount of context given to the annotator, who is told that the presence of a large number sequential marks the start of a new comment. Thus, the linguistic information between large numbers (generally) comprises one comment. The comment with the red word is taken to be the word’s immediate context.

## 3.3 – Results and Analyses

### 3.3.1 – Interrater Agreement Quotient

The interrater agreement quotient was calculated using the metric Cohen’s Κ , chosen instead of simple percent agreement to consider individual bias. It is used in conjunction with the confusion matrix, detailed below, to guide the design and execution of annotation and data collection. The calculation for the interrater agreement quotient is thus:

$$agreement= \frac{observed-expected}{1-expected}$$

Equation – Interrater Agreement Quotient

The interrater quotient for the tweaked Gaarder system was not very good, being only 0.23, barely meeting the requirement for “minimal” agreement, according to Stefanowitsch (2021, p.131).

### 3.3.2 – Confusion Matrix

The causes of the lack of cohesion can be seen in the confusion matrix below, taken from the data shared between raters A and D:

Table 4 – Confusion Matrix (Replication Experiment)

The following categories were omitted entirely from the confusion matrix because they were completely disused by both annotators:

* Respect/Courtesy
* Humility
* Strategic
* Euphemism
* Extravagance
* Baby Talk
* Slang

The cells represent the tally of responses where Rater A gave X and Rater D gave Y. For example, the cell Negative, Depreciation = 13 signifies that there were thirteen instances for which Rater A answered Negative and Rater D answered Depreciation. The ideal case would be that only the cells on the diagonal, where the two raters agree, would be filled, as all tallies that fall outside the diagonal indicate that the raters did not agree.

## 3.4 – Discussion and Conclusions

The first cells to notice are the two large agreements on the diagonal, that of Decrease and that of Depreciation. These suggest that the two raters had a clear idea of these two categories and that they largely agreed on this idea, implying that either the categories reflect clear semantic differences represented in the data or that the instructional materials were quite clear in their articulation, or both.

The first off-diagonal cell is the most bizarre. One of the two raters got confused during annotation and swapped Increase with Decrease or vice versa, leading to a systemic confusion between two diametrically opposed categories and contradictory annotation 13 times. This systemic error was probably a direct result of both the unwieldiness of the original list’s categorical structure and the overly elaborate coding system devised for quicker annotation.

Two additional off-diagonal cells are of note, implying a significant overlap: that between Negative and Depreciation and between Increase and Emphasis. While much of this can be attributed to a cultural conflation between the two concepts, it is believed that a not insignificant portion of the confusion was due to the abovementioned structural issues with the original category list or that the semantics of the Negative and Depreciation categories overlap so much that they are indistinguishable to the annotators.

Overall, Gaarder’s semantic categories appear to be too specific and inefficiently structured for annotators not linguistically educated to grasp and use easily and efficiently, with many distinctions being lost on the general public and sometimes confusing even to those with a linguistic education. These criticisms consequently called for the restructuring of the overall scheme into a more symmetrical and systematic framework with fewer, broader categories through the rebranding, elimination, merging, and rearranging of Gaarder’s functions into something new.

# 4 – Part III: Experiment with Revised Categories

## 4.1 – Creating Final Categories

As one could see from the discussion of the results and data above (Section 3.4), the interior arrangement of the old categories made usage difficult, indicating that structural changes were needed. Exactly what the categories should be was subject to much deliberation. As Jurafsky notes, there is a wide variety of roles and functions filled by the diminutive alone (Jurafsky, 1996), let alone the augmentative. This wide variety was simplified slightly by the above data, which gave a roadmap for what to do to establish the new categories. Seldom-used categories were to be eliminated or folded into other categories, resulting in broader categories that would need new names and descriptions. These new, broader categories could be arranged according to a series of functional axes, namely Magnitude, Emphasis, Affinity, and Reverence.

Magnitude would be the functional axis concerned with the physical or abstract modification of size of the concerned noun/adjective. Because the role of the augmentative or diminutive is almost always as expected (i.e., augmentatives almost always increase physical or abstract size whereas diminutives do the opposite), it was deemed unnecessary to make two opposing categories for this functional axis.

Emphasis would be concerned with denotationally and connotationally adding or subtracting intensity to the modified noun/adjective, but also, secondarily, highlighting the word within the discourse. Emphasis was clearly meant to be a functional axis from the beginning. *Intensification* comprised a disproportionate fraction of all datasets, usually used in an emphatic manner, to grab attention during discourse. In the new categorization model, this function has been renamed Accentuation. It would stand to reason, then, that an opposite function, a deemphasizer, would be necessary to balance out this axis of categories. This new category was branded Minimization, to represent the diversion of attention away, either from taboo or *Humility*.

While the above two axes would be more object-oriented, the following two would be more interlocutor-oriented.

Affinity would concern itself with the interlocutor’s attitudes of affinity or disaffinity towards the given object/adjective. This axis was already laid out by the seminal study, which featured a difference between *Appreciation* and *Depreciation*, now Affection and Dislike, respectively.

Reverence would be concerned with the interlocutor’s attitudes of reverence or irreverence. Its existence was made necessary because of the prevalence of *Condescending Superiority* in all annotators’ datasets. This category, now rebranded Contempt, conveyed a certain **lack** of respect that contrasted with *Respect/Courtesy*, now Respect,quite nicely. The seldom-used related terms *Impertinent Familiarity* and *Sympathy/Empathy* could then be folded into the newly minted Contempt category and then secondarily split into the new Respect, Minimization, and Affection categories, respectively.

This left a handful of categorizations without assignments. *Baby talk* and *Slang* were never really used and did not fit into the new framework. *Slang* itself was never really defined in the seminal study, used as a catch-all term for anything the author could not explain, and *Strategic* was not used even once in the entire dataset. *Extravagance* serves to call attention and emphasize something and was seldom used, so it could be folded neatly into the new Accentuation category.

## 4.2 – Materials and Methods

### 4.2.1 – Final Category List

Following the logic laid out in Section 4.1, the final list of categories was thus:

#### Magnitude

Modification to the physicality of the root word according to the nature of the ending itself. An *increase*, usually of size, in the case of augmentatives; a biggening. A *decrease*, also typically of size, in the case of diminutives; a smallening. Magnification can also, less commonly, concern quantity or duration.

#### Accentuation

This function is to intensify the intrinsic quality of the root word, either for emphasis or for playful exaggeration. If used on *amigo* (friend), it might emphasize the strength of the friendship. If it is used on *chocolate*, it might highlight its chocolatiness. This option is often open for purely physical, purely abstract, and partially physical, partially abstract root words. If the modifications exist to adjust the amplitude of the object or its abstract manifestation, accentuation exists to make either more intense.

#### Minimization

This function is to soften the intrinsic quality of the root word, often to distance or to euphemize. If used on *amigo* (friend), it might de-emphasize the strength of the friendship. If it is used on *chocolate*, perhaps it plays down the sense of chocolatiness. Perhaps you must use *el bañito* (the little boys’ room). As with accentuation, this option is often available for all three usual manifestations of nouns (purely physical, purely abstract, or a mixture). If the modifications exist to adjust the amplitude of the object or its abstract manifestation, minimization exists to blur either.

#### Affection

This function is to indicate a liking or a fondness, such as for soup, a friend, or a brother. Especially common with animate objects and common, useful items, usually with anthropomorphic tendencies.

#### Dislike

This function is used to indicate an antipathy or an aversion, perhaps even a disgust or revulsion. Maybe in reference to muddy socks, sticky heat, or blue cheese. Gross. Common to nouns at all levels of abstraction, but especially foods, people, or general categories.

#### Respect

This function is used to indicate a deference or reverence towards a big accomplishment or important figure. Martin Luther King, Alexander the Great if you are not Iranian, or Porfirio Díaz if you were Mexican and had money. You can respect your grandfather but probably not puppies. This option is generally only open to animate objects.

#### Contempt

This function serves to indicate disdain and even condescension, perhaps towards the newest generation of spoiled kids or, God forbid, hipsters. Oddly, this option seems to be open to both animate and inanimate objects, but generally not abstract nouns.

### 4.2.2 – Corpus, Exclusion/Inclusion Lists, and Annotators

The same corpora, exclusion/inclusion lists, and annotators as Section 3 were utilized, both in the experiment detailed in Section 4.2.3 below and in Section 4 in general.

### 4.2.3 – Category Calibration Experiment

Such vast and wide-reaching structural changes in the annotation system, though logical and rooted in empirical data, necessitated some sort of calibration. A much smaller dataset, consisting of what would become the Argentina and Chile sub-datasets, totaling 121 entries[[15]](#footnote-16), was used to calculate this second round of interrater agreement testing, which yielded a quotient of 0.7925, only 0.0075 away from a “strong” rating from Stefanowitsch (2021, p. 191). Being that the quotient basically indicated there was strong agreement between the two raters (thus implying a high degree of reproducibility), there was sufficient confidence in the system to move forward with implementation without delay.

## 4.3 – Results and Discussions

### 4.3.1 – Recap of Research Questions and Hypotheses

It would be useful at this point to revisit the guiding principles of this study. The intent of this study is to observe the usage patterns of augmentative and diminutive suffixation in the different dialectal regions of the Hispanosphere, secondarily surveying the existence and characterizing the usage of suffix reiteration as a productive linguistic process, to answer the following Research Questions, followed by their Hypotheses. These Questions will be restated at the beginning of every analysis section, as will their accompanying Hypotheses, in italics:

#### Research Questions:

1. Does regional variation exist in augmentative or diminutive use in Spanish?
	1. Concerning regional preference for various suffixes
	2. Concerning regional propensity to augment or diminish
2. Do augmentative or diminutive semantic functions vary across geographic regions?
3. What semantic functions do Spanish augmentatives and diminutives serve?

#### Hypotheses:

1. *Regional variation in augmentative and diminutive use is expected*
	1. *Regional preferences in suffix use are expected*
	2. *Regional variation in overall augmentation and diminution is expected*
2. *Augmentative and diminutive semantic functions are expected to vary by region*
3. *Different augmentatives and diminutives are expected to perform different semantic functions*

### 4.3.2 – Research Question 1

1. Does regional variation exist in augmentative or diminutive use in Spanish?
	1. Concerning regional preference for various suffixes
	2. Concerning regional propensity to augment or diminish
2. *Regional variation in augmentative and diminutive use is expected*
	1. *Regional preferences in suffix use are expected*
	2. *Regional variation in overall augmentation and diminution is expected*

#### 4.3.2 §1 – Results and Analyses

As mentioned previously, no distinction was made between gender and part of speech for a given ending. Counts for masculine and feminine suffixes were combined and referred to using the masculine form, e.g., -ITO. The following results and analysis do not include data from the Dominican Republic or Guatemala due to the paucity of data from these two regions but are instead included in the Appendices.

Because each regional sub-corpus was a different size, the raw count for each sub-corpus was normalized by being divided by the combined noun and adjective count for each sub-corpus and then multiplied by one million to get the normalized frequency per million adjectives and nouns (pman), tabulated in the tables below. Rather than dividing by the total word count of the sub-corpus, as is typical, this method would instead reflect the frequency that a noun or adjective in each region will contain a certain augmentative or diminutive. Given that augmentatives and diminutives are much more likely to modify nouns and adjectives as primitive bases, a presupposition that has already informed the design of the semantic categorization process, this measure would be far more representative of the propensity of speakers from certain regions to use the given endings, cutting out classes of words that see no or minuscule amounts of augmentation and/or diminution.

Table – Augmentative frequency per million nouns and adjectives, by geographic region

Table – Diminutive frequency per million nouns and adjectives, by geographic region

Even at this early point in the processing of data, some degree of regional variation is seemingly apparent. A cursory comparison of these international averages against the regional figures hints at regional variation of usage and distribution. However, before any trends may be inferred and accepted, it must first be established that these ostensible regional differences are, in fact, statistically significant, which must be done by rejecting the null hypothesis (H0,1) that these differences are instead due to chance. This is done using the Chi-Square Test, appropriate for two or more categorical variables.

The Chi Squared Test is done by comparing an array of actual values with a generated array of expected values and calculating a Chi Squared value, which is further processed to calculate a p-value, which represents the probability that the numbers found in the array of actual values are different due to chance, thus the probability that the null hypothesis is correct. There is a mathematical limitation with this test, that the results of the test are considered unreliable if the actual values array contains a value smaller than 5, thus the values found in were multiplied by 1000 to produced normalized values of frequencies per one hundred million words. At the same time, 5 was added to each and every cell, according to the supposition that the inserted sixes would be so *proportionally insignificant* as to *not affect the test result* almost at all. Applying the Chi Squared Test, with the added adjustments enumerated above, both augmentative and diminutive arrays yield infinitesimally small p-values (**p < 0.0001**), confirming the rejection of H0,1 and the validation of H1. The different frequencies with which the regions use the carious suffixes are not by chance. We can conclude that each region has its own preferences for augmentative and diminutive suffixes. Having accepted that the data represented in these new tables are of statistical significance, various regional trends are now in full relief.

Of note are the global minima and maxima of augmentation and diminution, informed by the total augmentation and diminution frequencies per noun or adjective in a region, found at the bottom of Table 2 and Table 3, respectively. These figures represent the frequency, per million words, that a random noun or adjective from the given region will be an augmentative or a diminutive, separately. From these total numbers, especially when compared to the international average at the lower right-hand side of their respective tables, interesting international trends become apparent.

The Argentines seem the most predisposed population to augment their bases, at 588 augmentations per million adjectives and nouns (augs pman), about 60% above the international average of 371 augs pman. Strangely enough, the country most predisposed to diminish lies just across the Andes Mountains. Chileans absolutely riddle their nouns and adjectives with diminutives, applying them to a whopping 3354 nouns and adjectives per million, an entire 65% more than the international average of 2027 dims pman.

The Spanish-speaking world’s largest party-poopers, as is the technical term, seem to be the Colombians. Both the least likely to augment and to diminish, the speech of Colombia remains, relative to the other surveyed regions, the least modified, only augmenting 225 nouns and adjectives per million (39% less than average) and applying diminutives to only 1169 words per million nouns and adjectives (42% lower than average).

To make the regional preferences for certain endings clearer, the frequencies of the endings have been converted in Table 3 and Table 4 below to percentages of each sub-corporeal total, conveying the share of augmentation or diminution each suffix represents. Each region’s most frequent augmentative and diminutive morpheme is boldfaced.

Table – Share of Augmentation

Table 4 - Share of Diminution

Subjecting these new data to another Chi Square Test, with the same modifications and exclusions as the other above, yields p-values below 0.0001 for both sets yet again, permitting the sound rejection of the Null Hypothesis (H0,1a) and validating the Hypothesis (H1a) that these specific differences in suffix usage are due to regional variation.

The undisputed international favorite diminutive is -*ito*, which comprises upwards of 80% of diminutives in almost every region sampled. However, this is not the case in the Spanish language’s country of origin, Spain, where it only forms a significantly reduced 57% of diminutives, lowering the international average to 77%, which is still quite dominant. The distant second, and the ending which siphons off most of the missing Spanish -*ito* volume, is -*illo*, making up around 10% of diminutives outside of the Iberian Peninsula, but 30% within.

The augmentatives demonstrate an altogether different dynamic than the oppressively consolidated field of diminutives. Each region’s representative country is characterized by a much more distinctive distribution of augmentative suffix usage. The Mexican and Colombian varieties exhibit remarkably similar frequency distributions, the only major difference being the Mexican usage of the purported -*che* suffix and its absence in Colombian speech. The Southern Cone is, again, the site of distinct divergence, with Argentine augmentation being concentrated in the hallmark -*udo* ending while also maintaining an even spread of the remainder throughout the other suffixes, while Chilean augmentation concentrates largely on -*azo* and the infamous -*ón* ending. Peninsular Spanish, ever the maverick, converges predominantly on -*ísimo*, an ending not even traditionally considered an augmentative.

Finally, to answer the question of whether there is a propensity to diminish more than to augment, one need no more than look at the total frequency of augmentation and compare it to the total frequency of diminution, in Tables 1 and 2, respectively. The ratio of augmentatives to diminutives, when calculated and tabulated, yields the following results:

Table – Ratio of Augmentation to Diminution



The average augmentation to diminution ratio internationally is 0.18, and each nation’s ratio is, without exception, smaller than 0.3. This means that, on average, a random Spanish speaker is more than 5.5 times as likely to use a diminutive than an augmentative. Therefore, it is indisputable that diminution is far more popular than augmentation, thereby validating Hypothesis H2b by rejecting H0,2b that this variation is due to chance. Generally, Spanish speakers, it would seem, are about five and a half times as likely to use diminutives than augmentatives.

#### 4.3.2 §2 – Discussion

Nothing within the data themselves gives direct evidence of why Argentina and Chile augment and diminish, respectively, so often, why these two neighboring countries differ so starkly, or why Colombia so reluctant to suffix at all. The author speculates that, much as evidenced in other areas with similar geography, the impressive height of the Andes may have had an insulating effect on the speech of Chile, which may have allowed a perhaps preexisting propensity to augment to shift into one to diminish, or vice versa, as the cultures of the two countries diverged over the centuries. The isolation of Chile is further evidenced by unique accentual hallmarks such as the intense use of “po’ ” and unique realizations of verbal conjugations because of accentual pronunciation differences.

As summarized in *Augmentative/Diminutive Suffix Usage* above, the most common augmentative suffixes are -*ísimo* and -*azo*, their diminutive counterparts being -*ito* and -*illo*, respectively. The table similarly portrays Spain as the maverick when it comes to ending preference, being the most likely to choose *-ísimo* (not unexpected), the least likely to choose *-azo*, the least likely to choose *-ito*, and the most likely to choose -*illo* instead.

There are a few possible explanations for the stark difference between Spain and the former colonies in suffix preference. The author posits that it may be historical, as Spain is currently balanced between North and South Peninsular Spanish varieties in the dataset, but the phonological and morphosyntactic tendencies of South Peninsular Spanish are what comprised the core of what would later become the Latin American varieties of Spanish. This fact is evidenced by the near universal lack of *distinción*, or distinction between the sibilants /s/ and /θ/ in minimal pairs such as *casa* and *caza*, in the Americas, the dialects of which instead exhibit *seseo*, or the collapse of both into /s/, previously limited only to the southern Iberian Peninsula.

Though not thoroughly investigated in this study, a cursory analysis of the exclusion lists for each ending coincides suggestively with this supposition, as -*illo* seems to be the diminutive ending by far the most likely to have been lexicalized. In regions of the world where the productive use of this ending is still relatively high, the derivational rationale for these otherwise lexicalized words may yet live in the minds and lexica of speakers, strongly influenced by the North Peninsular variety, whereas in regions without this regularity of use, the ending may survive in a reduced form, either fossilized in lexical items or, perhaps due to its reduced status, the use of this ending specifically for items up for lexicalization. Had the sample been more geographically specific, the author suspects that there would have been a higher frequency of this -*illo* ending in the north and a smaller one in the south, though no proof for this speculation is offered.

An interesting sidenote is the existence of productive occurrences of the ending -*iño* in Argentina and Chile. They occur in the countries closest to the major population centers of Brazil, which is to be expected, as the rough phonological equivalent of this ending in Portuguese, -*inho*, is highly productive in, at least Brazilian, Portuguese, being otherwise extremely rare outside of the Spanish-speaking periphery of the Lusophone world. The other region that would be expected to attest this ending is the Spanish region of Galicia, had sub-national varieties been specified in the data.

Another, salient, observation that can be made is the utter lack of success in capturing mainstream -*ito* volume displayed by the famously regional diminutive substitute -*ico*, especially in the expected home territory of Colombia. This could possibly be explained by a stigma against this ending as a mark of ignorance and lack of education, which may not be compatible with the cautious and defensive style so common to some online communities, concerned about judgment on their indelible comments online.

It is prudent to note again that the definitions of *augmentative* and *diminutive* have been expanded in this study to include morphemes with augmentative- or diminutive-*adjacent* meanings as well, such as intensifiers. This is, as stated far above, due to the presence of a systematic and strident conflation across the Spanish-speaking world between the physical or metaphorical size of an object and its intensity, importance, etc. This conflation does not solely exist in Spanish, but seems to be quite prevalent, hence the inclusion of these adjacent morphemes in the study.

Note is made of this because it relates to the distribution of augmentatives between the strictly intensifying and the more physically augmentative endings. In Spain, again the maverick, there seems to be a clearer distinction between these two ideas, as the intensifying ending *-ísimo* dominates the augmentative space, whereas in Colombia, this space is shared fairly equally among *-ísimo, -azo*, and *-ón.* This hints at a much more comprehensive conflation between the two concepts in Colombia and a partial offloading of the intensifying role from -*ísimo* and onto the other two endings. Even though this distribution is not shared exactly by any other country, with most countries showing a clear preference for at least one of these endings, averaging these figures internationally seems to corroborate the notion that this conflation is at least partially present in all five of the major regions, most of all in the Americas.

### 4.3.3 – Research Question 2

1. Do augmentative or diminutive semantic functions vary across geographic regions?
2. *Augmentative and diminutive semantic functions are expected to vary by region*

#### 4.3.3 §1 – Results and Analyses

The semantic categories from the annotated data set, broken down by semantic functions (combinations when present) and geographic subregion, are tabulated below, normalized as percentages of the total augmentation or diminution, respectively.

Table – Share of Augmentative Semantic Function

Table – Share of Diminutive Semantic Function

As noted above, the entire data set collected was annotated according to the renovated annotation system detailed in the project further above. However, due to the small size of the data set (thanks to limited time, manpower, and technological resources) the resulting semantic categorization data set was rather sparse, as seen by the sea of red directly above. Rather than drastically complicate the analysis of each combination of semantic category, as well as increasing the possibility of drawing statistically significant conclusions, the combinational annotations were aggregated by primary annotation (i.e., combinations marked “con/acc” and “acc/con” would be counted as CON and ACC, respectively). The results are tabulated below:

Table – Share of Augmentation by Primary Category

Table – Share of Diminution by Primary Category

The first step in establishing the effect of regional variation in the semantic distribution of augmentative and diminutive endings, at least in terms of the endings themselves, is the performance of a Chi Square Test. The above figures were enhanced by 1000 and 5 was added to all cells, following the smoothing process. As before, the probability of the differences between these numbers being due to chance (p) calculated from the above tables is so infinitesimally small that the author’s computer could only represent the numbers as zeroes. The final figures for both Chi Square tests are thus taken to be **p < 0.0001**.

Several interesting observations may be made about this aggregation of the semantic categories, the most striking and straightforward of which is the nearly complete and total lack of Minimization, Respect, and Dislike as primary functions in augmentation (all capturing nearly none of augmentation in all regions). This is contrasted against the most popular primary augmentative semantic category, Accentuation (43.7% of augmentation on average), which peaks in Mexico and Spain (54% and 51%, respectively) and dips in Chile (still a sizeable 29% of augmentation).

Turning to the diminutives, Accentuation (43% internationally) again takes first place. The Colombians are far and away the most prone to using diminutives in an Accentual fashion, almost 50% more likely than the international average, comprising the principal part of a whopping 61% of all diminution in the country All other countries use Accentuation as the primary element in about 40% of their diminutions, with Argentina being only slightly below as the least likely (37%) to utilize the diminutive in this fashion, even though Accentuation is their most preferred semantic usage of the diminutive. Interestingly, there seems to be a lack of both Dislike, as in augmentation, but also of Respect as primary semantic functions in diminution across the board.

A more granular approach may be deemed necessary than simply aggregating by the primary category. The sparsity of data can be minimized while increasing granularity by not focusing on the order of the annotation elements in the combination categories. Given a broader set of data, the author fully expects for the distinction between, say “con/acc” and “acc/con,” as given above, to be significant. But it was deemed necessary to instead collapse the combination categories such that order does not matter (i.e., “con/acc” and “acc/con” would both count as “acc+con”). The results of this aggregation are given below:

Table 8 – Share of Augmentative Annotations by Aggregated Combinations

Table 9 – Share of Diminutive Annotations by Aggregated Combinations

The Chi Square test was performed on the above data, again with the statistical smoothening of multiplying by 1000 and then adding 5, to yield two more p-values so infinitesimally small as to be considered **p < 0.0001**.

Beginning with augmentation, pure Magnitude and Accentuation tie for most common of all (23% internationally), with Mexico utilizing both most often (29% and 30% of augmentation, respectively). Nearly all other areas are tightly grouped around the average, with the exceptions of Colombia for Magnitude (13%) and Spain for Accentuation (18%). Respect is found completely absent in its pure form in the augmentation of *all* tested regions.

This is almost perfectly parallel with the situation for the pure diminutive semantic functions, of which Accentuation is the most frequent (on average comprising 28% of diminution internationally), being favored by Colombia (45% of diminution, a little over 60% more than the international average). The other regions again group tightly around the international average, apart from Mexico, which dips to 20% of diminution. Pure Respect seems to only exist in Colombia.

The retention of the combination categories brings some much-needed nuance. Of these collapsed combination categories, Accentuation+Contempt reigns supreme in augmentation (15% internationally) but varies insanely from 27% in Colombia to only 10% in Argentina and Spain. Accentuation+Contempt and Accentuation+Affection are the most common diminutive combination annotations, comprising 9% and 8% of total diminution, respectively. Any combination involving minimization is totally absent from augmentation. The diminutives are also marked by a total lack of the combination annotations of Magnitude+Respect, Magnitude+Dislike, and Minimization+Respect, with Minimization+Dislike not far behind.

All four of the abovementioned p-values, both by primary category and aggregated combinations, are much, much smaller than alpha (0.05), so the author can, on all counts, confidently reject the null hypothesis H0,2, permitting the acceptance of Hypothesis H2: that regional variation exists with respect to semantic distribution.

#### 4.3.3 §2 – Discussion

Accentuation far and away appears to dominate both augmentative and diminutive suffix usage across all regions of study, regardless of if aggregated by primary categorization or by unified combinations. Magnitude similarly dominates augmentative usage regardless of aggregation strategy. Conversely, Respect is represented the most poorly of all semantic categories in all environments, followed extremely closely by Dislike.

The fact that both augmentative and diminutive suffixes exhibit Accentual semantic functions most often in all environments challenges the author’s previously held hypothesis that the confusion in augmentatives between increase in size and intensification in Part 2 (illustrated by the confusion matrix in Section 3.3.2) was primarily due to conflation between the large and the important. The similar frequencies between the two successor categories, Magnitude and Accentuation, suggests that, had annotators been allowed to coincide them[[16]](#footnote-17), these categories would have been strongly co-occurrent, indicating a strong link. But the asymmetry in the phenomenon’s intensity between augmentatives and diminutives suggests either that this is not the mechanism of confusion or that another mechanism, perhaps another metaphorical scheme, takes precedence regarding diminutives.

This asymmetry does instead reaffirm the decision to maintain a distinction between magnitude and accentuation in the newest semantic categorization scheme, as, with the above data, the average international primary categorization aggregates for referential categories can be rearranged[[17]](#footnote-18) and heat-mapped to illustrate the proposed mechanism visually[[18]](#footnote-19):

Figure ­– Skew in Referential Suffix Usage, by Primary

An asymmetrical mechanism of a different sort, perhaps a misalignment of the annotation system, may also be causing the enormous rift between Respect, almost absent in all environments, and its negative counterpart, Contempt, which is itself healthily used, even being tied for the most used diminutive combination when paired with Accentuation (Accentuation+Contempt). This phenomenon is not limited to Respect/Contempt but also extends in mirrored form to Affection/Dislike: Dislike is almost absent, like Respect, in all environments whereas Affection is one of the most often used semantic elements, tying with the abovementioned Accentuation+Contempt when also in conjunction with Accentuation (Accentuation+Affection). When primary category aggregates are tabulated and heat mapped as done above, the skew becomes visible to the naked eye.

Figure – Skew in Expressive Suffix Usage, by Primary

The new semantic categorization system is only equipped with Magnitude as a semantic category to code both increases and decreases in physical or metaphorical size. This is because the number of instances where a suffix performs a magnitudinal modification in the opposite direction (i.e., an augmentative used to make an object smaller, a diminutive to make an object bigger) is so minuscule as to be safely ignored. It is for this reason, as well, that the opposite magnitude (decrease for augmentatives and increase for diminutives) has been assumed to be 0% in the table above.

Also of note is the fact that, when the difference between pure and combinational semantic categorizations is upheld, pure semantic categorizations, at least qualitatively, seem to be slightly favored by all regions for both augmentatives and diminutives, though the author has no ready explanation for why.

Upon qualitative examination of the above data, both augmentatives and diminutives favor pure annotations rather than combinational ones.

### 4.3.4 – Research Question 3

1. What semantic functions do Spanish augmentatives and diminutives serve?
2. *Different augmentatives and diminutives are expected to perform different semantic functions*

Though partially addressed in proving Hypothesis H2 in Section 4.3.3 §1, the answer to such a broad question as RQ3 requires a more general look at the data and a novel aggregation of the semantic categories to investigate the resulting contrasts, such as positivity vs. negativity and referentiality vs. expressivity. To examine these differences, one must determine what “positive,” “negative,” “referential,” and “expressive” mean in this context.

#### 4.3.4 §1 – Positivity vs. Negativity

##### 4.3.4 §1.1 – Results and Analyses

Gaarder refers to “favorable” vs. “unfavorable” semantic meanings, making it the primary distinction between his Positive and Negative Evaluation categories and between his Appreciation and Depreciation categories. This corresponds nicely with the differentiation between this study’s semantic categories of Affection and Respect, and Dislike and Contempt, respectively. Thus, the categories Positive and Negative are comprised of the fine-grained categories of Affection and Respect, and Dislike and Contempt, respectively. The resulting aggregation of positivity (Affection and Respect) and negativity (Dislike and Contempt) is tabulated below:

Table – Share of Positivity and Negativity in Augmentation (by Primary Category)

Table – Share of Positivity and Negativity in Diminution (by Primary Category)

Argentines are the most predisposed to use augmentatives in a primarily positive light, more than twice as likely than the average Spanish-speaker, which contrasts strongly against Colombians, who apparently never use augmentatives to convey a primarily positive connotation. Columbians, by contrast, are about 67% more likely compared to the international average to use augmentatives in a primarily negative fashion.

The Chileans again make an appearance as the most likely to use diminutives in a primarily positive fashion (17%), along with the Mexicans (17%), both significantly more likely than least, the Spaniards. In addition to never using augmentatives in a primarily positive light, the Colombians simultaneously also have the peculiar honor of not using diminutives in a primarily negative fashion at all, a tendency Spain vehemently disagrees with (22%, almost three times the international average of 8%).

When performing the same aggregation to the unified combinations (i.e., counting when positivity or negativity is a part of the annotation, regardless of whether it’s the primary semantic meaning or not), we derive the following tabulations:

Table – Share of Positivity and Negativity in Augmentation (any Inclusion)

Table – Share of Positivity and Negativity in Diminution (any Inclusion)

When any hint of positivity or negativity is counted, either primary or secondary, the data becomes far less dramatic than before. Spain once again distinguishes itself as being the most likely to use an augmentative in some positive fashion, almost *four times* as likely as the least prone to such usage, Colombia, and almost twice as likely as the international average. Spain is also the least likely to use the augmentative in a negative fashion of any kind, which Colombia does almost *three times* as often.

All regions are almost equally as likely to use diminutives positively, with the recurring notable exception of Colombia, which drops far below the international average of 29% to 21%. Here Spain is again an oddball, being the most likely to use diminutives in a negative fashion of some kind, contrasting sharply against the tendency of Chile.

##### 4.3.4 §1.2 – Discussion

Examination of the positivity vs. negativity dichotomy becomes much clearer after calculating and tabulating the ratio of positivity to negativity for both augmentatives and diminutives, as shown below:

Table – Ratio of Positivity to Negativity in Augmentation and Diminution (by Primary Category)

\* - Diminutive international average calculated without Col, Dim (6/0 = undefined)

The ratio data above suggest that, with few exceptions, diminutive suffixes are, on average, 2.3 times more likely to be used primarily positively than primarily negatively, whereas augmentatives are 2.3 times more likely to be used primarily negatively than primarily positively, almost exactly the opposite. Colombia is notable for being the most extreme of the studied countries, with augmentative use being skewed almost entirely primarily negatively and diminutives being skewed almost entirely primarily positively.

Table – Ratio of Positivity to Negativity in Augmentation and Diminution (Any Inclusion)

All regions seem to agree to imbue augmentatives with some sort of negative connotation more often than a positive one, merely disagreeing on how much, from the nearly ambivalent Argentina to the almost fanatically negative Colombia. However, Spain, ever the special snowflake, tears away violently from this agreement, skewing about 10 times as forcefully in the direction opposite of Colombia. By contrast, all regions agree that diminutives are usually somewhat positive. Colombia again remains the most conservative, and Chile is the most convinced of the goodness of small things.

Upon examining the above data, one can see the trend of augmentatives being used mostly primarily negatively and diminutives the opposite, as noted before. Colombia features prominently on these tables, being the most likely to use augmentatives negatively, the least likely to use them positively, the least likely to use diminutives negatively, and the nation whose diminutive usage skewed most positively for diminutives and most negatively for augmentatives.

The fact that Colombia’s usage is the most skewed of all, both positively for diminutives and negatively for augmentatives, contrasts interestingly with the abovementioned finding that Colombia is the most conservative of all studied nations to use augmentatives and diminutives in the first place. It suggests that Colombian Spanish speakers use augmentatives and diminutives sparingly but with vigor and clarity of purpose when doing so.

Figure – Skew in Expressive Suffix Use, by Primary

These data, when combined with the Affection/Dislike, Respect/Contempt skew (illustrated again above), suggests, tentatively and qualitatively, that when an object, usually small or cute, is described in a positive expressive manner, the general Spanish-speaking public will convey positivity Affectionately (and with a diminutive) rather than Respectfully, but when this evaluation shifts negatively, often towards a large or menacing object, this negativity will be expressed through Contempt rather than through Dislike, more often than not with an augmentative. It is no mere coincidence that, when the categories are arranged as above, along the axes explained in Section 4.2.1 (Final Categories), the top row of expressive categories coincides nicely with the “Positive” grouping of semantic categories and the bottom with the “Negative” grouping.

#### 4.3.4 §2 – Referentiality vs. Expressivity

##### 4.3.4 §2.1 – Results and Analyses

Examining the contrast between referentiality and expressivity requires a further aggregation of the semantic categories, collapsing them until only the distinction between referential (denotative, Magnitude plus Accentuation/Minimization) and expressive (connotative, Affection/Dislike and Respect/Contempt) semantic categories remains. Below are the results for the aggregation by primary categories:

Table – Share of Referentiality and Expressivity in Augmentation (by Primary)

Table – Share of Referentiality and Expressivity in Diminution (by Primary)

Mexico has proven itself exceptional by using the augmentative most often for primarily referential purposes, almost 100% of the time. This contrasts with Colombia, which only uses the augmentative primarily referentially around 70% of the time. Mexico, on the other hand, uses augmentatives expressively the least, at less than half the international average, contrasting with Argentina.

The Colombians continue their trend of skewing sharply, this time the most likely to use diminutives referentially, disagreeing most with the Spaniards, by almost 50%. Spain and Colombia continue their disagreement into expressive diminutive use, being the most likely to use diminutives in an expressive fashion while the Colombians are the most reluctant, by almost the exact same margin as above.

It was considered of possible interest to extend the above analysis parallel to the “any inclusion” method applied to the other analysis sections before. However, since combination annotations contain both referential and expressive elements, a decision had to be made on how to count said combinations. Since the above data demonstrate that primary referentiality is overwhelmingly present, it was decided that, for ease of calculation, the collapsed combination annotations would be counted as “somewhat expressive,” yielding the following table, which contrasts pure referentiality against expressivity of any kind.

Table – Pure Referentiality vs. any Expressivity in Augmentation

Table – Pure Referentiality vs. any Expressivity in Diminution

Among augmentatives, Mexico still seems to be the most likely to use augmentatives purely referentially and Colombians retain their reluctance to do the same, but in much more pronounced fashion, almost half as likely as Mexican Spanish-speakers. The Colombians are, instead, tied with the Spaniards for the most likely to use augmentatives in an expressive fashion of some sort.

Among diminutives, Colombians stand out again by being the most likely to use them purely referentially. However, this time they disagree most with the Spanish, which exhibit this behavior about 50% less, and are instead the most likely to use diminutives with some degree of expressivity, again disagreeing with the Colombians, who exhibit this behavior about 50% less, mirroring the abovementioned level of contrast.

##### 4.3.4 §2.2 – Discussion

Table – Ratio of Referentiality to Expressivity in Augmentation and Diminution (by Primary)

As with Positivity vs. Negativity, the trends become clearer when the data is converted into ratios. When aggregating by primary category, Spanish speakers in general use augmentatives and diminutives more primarily referentially, that is, with regards to the size or intensity of the modified object, than expressively, or to reflect the speaker’s attitudes towards the object. This skew, however, is much more pronounced among the augmentatives than the diminutives, almost doubly so. This finding is somewhat surprising, considering the angle Gaarder takes in his study that most augmentative and especially diminutive use is expressive.

Table – Ratio of Pure Referentiality vs any Expressivity

When aggregating by counting combinations as “somewhat expressive,” Colombians are confirmed to skew most extremely between augmentatives and diminutives, both towards expressive augmentative use and towards purely referential diminutive usage. While their rival from above, Spain, holds the highest skew towards expressive diminutive usage, Mexico skews more strongly than them towards purely referential augmentative usage. Overall, however, regional differences aside, the international Spanish speaker is almost equally as likely to use both augmentatives and diminutives purely referentially as with some expressive function. This is most exemplified by Argentina and Chile, who hover near an even split between pure referentiality and some expressivity for both augmentatives and diminutives.

The data suggest that, when referentiality and expressivity are placed on equal footing by aggregating by primary category, Spanish speakers are far more likely to use augmentatives and diminutives in a primarily referential fashion, very especially augmentatives. The ratio data demonstrates that the most eager to do so, especially with augmentatives, is, by leaps and bounds, Mexico, with the least being Spain.

However, when pitting all categories with at least some expressive semantic value against purely referential suffix usage, the data does something interesting. While the overall minima and maxima remained largely the same or changed expectedly (Mexico retaining the role of most likely to use augmentatives referentially, Spain remaining the most likely to use diminutives expressively, and Colombia becoming the most skewed both referentially and expressively), the unequal footing given to expressive forms evened out the ratios between referentiality and expressivity for both augmentatives and diminutives, making diminutives more expressive for the first time (more in line with expected results). In addition, Spain went from being much more likely to use both diminutives and most especially augmentatives referentially rather than expressively to leaning towards expressivity for both augmentatives and diminutives rather equally.

## 4.3 –Final Discussion

It is interesting that Mexico in particular, when observing its corpus share qualitatively, seems to be the most open to innovative suffix usage. For example, it seems to be the only studied region in which -*ucho* shows some productivity. *Infraestructurucho* (infrastructure-*pejorative*), *periodistucho* (reporter-*pejorative*), and *presidentucho* (president-pejorative) all count at least one instance in the Mexican sub-corpus, and it is of special note that these are **not**normal or typical forms of these nouns. Interesting as well is the largely intact conservation of -*ucho*’s nominal pejorative (i.e., contemptuous) usage in these instances, though with a degree of playful exaggeration that Gaarder was trying to get at in his original semantic categories.

On a more important note, the nature of the corpus itself may impact the irregularities, skews, or at least some asymmetries observed. Since the corpus mines an almost entirely anonymous internet forum, there is (here the author speaks from experience) a much more liberating freedom when interacting with other users. Social norms and inhibitive forces are far weaker because of the ephemerality of Reddit interactions and the lack of any tangible retribution, like physical violence, or punition, like legal consequences, against the perpetrators of any except the most egregious of social breaches. This lack of inhibition, akin to the one found among the inebriate, often results in aggression, contempt, and a lack of respect, which might have impacted the amount of contempt and respect found in the sub-corpora positively and negatively, respectively. A different type of corpus, such as a live recording of impromptu in-person conversation, would likely see the reverse; a stark decrease in contempt and a dramatic increase in respect, as in-person relationships are much more durative and consequential, especially among friends and family.

# 5 ­– Conclusions

## 5.1 – Final Conclusions

The corpus analyzed in this Honors thesis is an original, synchronic, conversational corpus created by the author in 2021 by using R to scrape Reddit to capture genuine, linguistically valuable conversational data between users from around the Hispanophone world. Through the use of SketchEngine, 10 million tokens were compiled and annotated from seven nations representing seven major dialectal regions of the Spanish language (Mexico, the Dominican Republic, Guatemala, Colombia, Chile, Argentina, and Spain) and were sorted by their origin into sub-corpora. Hosted on SketchEngine, this online data resource will be available for other scholars to use in the future.

The experiment detailed in Bruce Gaarder’s key study was replicated by double annotating a much larger, more representative, corpus using his semantic categorization scheme. By statistically analyzing the data produced, new categories were developed and arranged into an entirely new categorization scheme for these semantic functions in Spanish, with influences from newer studies and methodological frameworks from both the English- and Spanish-language academic mainstreams. These new categories were verified, and the scheme was then applied to a sample of the corpus, yielding results that confirmed without much doubt that different dialectal regions of Spanish speakers do exhibit statistically significant regional variations in the augmentative and diminutive suffixes they prefer, the rate at which they augment or diminish, and the meanings they most and least often convey with such suffixes.

Notable among the findings is the distinct Colombian trend of using augmentative and diminutive suffixes sparingly but with focused purpose, the near-universal propensity of Spanish speakers to express positive attitudes through Affectionate diminutive use while expressing negative ones through Contemptuous augmentation, the monopoly of Magnitude and Accentuation in both diminutive and especially augmentative usage, and the cross-dialectal conflation of Magnitudinal and Accentual augmentation, perhaps due to metaphorical mapping. An important corollary finding is that Peninsular Spanish maintains distinct currents from Magnitude and Accentuation to traditionally dedicated augmentatives ( -*ón*/*ote/azo* and -*ísimo*, respectively) whereas in all studied American dialects, to varying degrees, the distinction has eroded, and the semantic flows comingle among the augmentative suffixes.

## 5.2 ­– Areas of Future Study

The author can imagine a myriad of investigations spawning from the work done here in this thesis. The final SpanRed corpus alone could potentially contribute to many studies, but there are more concrete ideas as well, such as replicating this project with a larger dataset, more annotators (at least two from every examined region), and potentially more granular sub-regions, or flipping the project on its head, i.e., giving the same corpus to an array of annotators from around the Hispanophone world and recording the differences in how they assign annotations. Perhaps a closer look at what endings are most likely to lexicalize and how much?

Just working with the data gathered and annotated here, an analysis could be done to see if the semantic category of the augmentative or diminutive suffix in context depends on the particular suffix used and if there is regional variation. As mentioned above, Spain maintains a distinction between Magnitude and Accentuation in Augmentatives. Perhaps looking at a diachronic corpus could illuminate why, when, and where this distinction first eroded, or if it was even a post-colonial development within Spain alone.

Most exciting of all would be a broader replication study but with a special focus on regional variations in handling reiterated suffixation, in what contexts they occur, how often they do, and what their purpose is. It has been posited that reiterated suffixation does not occur, or only very rarely, but the author found a handful of instances even in the small annotation datasets, with a distinct proclivity towards Accentuation and entirely within the Americas, specifically in Mexico.

## Appendices

### Appendix B – Part I

*Appendix B1 – Subreddit List (caps-sensitive!)*

* r/argentina
* r/chile
* r/Colombia
* r/Dominican
* r/guatemala
* r/mexico
* r/spain

#### Appendix B2 – Cutoff Algorithm

The harvesting script was to slice the list at a certain cut off and only use the URLs above said index, e.g., the “top one hundred” or “top 238.” Knowing no other approach, the author flipped the table of URLs with accompanying comment counts and, going down from smallest to largest, searching for when a certain comment count threshold, e.g., ten comments, was reached. Once it was, the URL index, e.g., #335, was noted. Theoretically, this meant that, of the, say, 500 URLs in the table, 335 contained at least ten comments. The index matching the following, larger, comment count threshold was then sought out and recorded. The thresholds were usually separated by tens, and the process was repeated until the end of the list, with the URLs of the posts with the largest comment counts, was reached. The table was flipped right way up, and the data were plotted out on a graph.

Due to the nature of the data and of Microsoft Excel, the X-axis of the graph needed to represent the comment threshold and the Y-axis represented the URL index in the table. The area under the curve was then interpreted, perhaps erroneously, to be equal to the combined comment count of all the URLs in the table. Thus, given a horizontal line representing the cutoff at table index Y, e.g., Y = 200, then the area both under this line and the curve was interpreted to represent the combined comment count of the posts represented by the top Y (in this example, 200) URLs in the list. Equipped with this tentative assumption, cutoff points were selected that would maximize the amount of high-yield Reddit posts harvested.

### Appendix C – Part II

###### Appendix C1 – Annotator Groupings

Annotators A, B, C, and D were collected into the following Groups according to their living arrangements: Group 1 (A, B) and Group 2 (C, D). They were then separated across households into the following pods to limit cooperation: Pod 1 (A, C) and Pod 2 (B, D).

#### Appendix C2 – Coding Packet

Today, you will be evaluating what role(s) is/are being played by the words marked and in the middle of context. We will go over what *augmentatives* and *diminutives* are and their functions. You are probably more familiar with these than you may think. First are *diminutives*. An example of a diminutive is the suffix -ito, as in *perrito*. *Perro*, as you know, when unmodified, simply means “(male) dog.” When modified with -ito, the meaning becomes *little dog*. As you are aware, -ito is not the only suffix to exist in Spanish that makes things smaller. This class of suffixes is what linguists call *diminutive suffixes*, and it is the real-world usage of these and of the opposite class, the *augmentatives*, which I wish to evaluate, with your help.

Second, we will go over their functions. As you know, making things smaller is not the only use for a diminutive suffix. When someone says “mira esa casita,” they might not always be talking about the smallest house being looked at, but often the house that they deem “cute.” In such a context, the diminutive’s role is not *making things smaller* but *making things cuter*. This is one of the functions we will be evaluating into categories. There are two domains of categories, housing eighteen functions in total. We will go over each of the broad categories and then break them down into their constituent categories, then delve into the functions, with examples for each.

This study splits the functions of augmentatives and diminutives into the intellectual and emotional domains, or, more roughly, between “objective” and “subjective” evaluations. An example of an intellectual evaluation is using the diminutive -ito in *perrito* to describe an actually small dog (a Chihuahua or a Dachshund, for instance). The above context of using the diminutive to describe a house as “cute” would be an example of an *emotional* evaluation.

Intellectual evaluations are broken down further into two sub-domains, or “categories:” Physical and Evaluative, which are relatively simple. The Physical category is easy. In the example of *perrito*, a Physical categorization means that the dog was physically considered small. *Perrote* or *perrón* would indicate that the dog is physically considered large. These are examples of the two functions in the Physical category, Decrease of size, quantity, intensity, etc. and Increase of size, etc., respectively.

Evaluative categorizations are a bit tricky. The valuation may be Positive or Negative. They reflect a valuation of the object on part of the speaker, either positive or negative. This may be difficult to determine in the limited context given and may be conflated with or co-occur regularly with categorizations of the emotional domain. There is an example that may prove illustrative: a man, without having met the subjects, states, “Qué grandes amigotes tenía Chente entre ellos.” Without having met the friends, the speaker evaluated that they must be good friends.

The Emotional domain is quite complex. We will go over these in order, with examples.

First is Appreciation, fairly straightforward. This comes from many sources, most commonly affection for the modified object. An example: “Le dijo cariñosamente: —Ahora, amiguito, a trabajar.” Affection is even made explicit by the word *cariñosamente*.

Next is Depreciation, mostly from disgust, resentment, or hatred. For example: “¡Quita este viejillo de mi vista!” The content of the sentence indicates some sort of resentment toward the old man, so the diminutive in this context is seen as a belittlement.

The subsequent category is Condescending Superiority, often because of gender or some other gatekeeping function. An example: “—Muchachita, por ser mujer no entiendes la complexidad del arte.” The disdain and superiority towards the female in this context are obvious.

Impertinent Familiarity is when the suffix is used in a familiar context without the familiarity. An example is a child going over to his friend’s house for the first time and naming his friend’s parents by name, with the diminutive suffixes to indicate a non-existent familiarity

Respect and Courtesy is a function that is often used towards social superiors. “---Buenas tardes, amo don Juanito” is a good example of this, someone showing reference to the master of wherever they are and showing some level of care and maybe even affection. Another example is referring to a priest as *padrecito*.

Humility is another common function. “Les hice estas tortillitas” said by a woman who spent some time making tortillas for her family is an example of a level of humility. She might be proud of them but does not want to come off as boastful, so reduces them in metaphorical size.

Strategic Timidity is another function, harder to determine, but quite common, usually springing forth from deceit, but sometimes for another ulterior motive. An example found in the literature is a man in prison saying to a large, convicted murderer: “—¡Manuelito, por favor, cálmate!” His intention is to use this diminution to calm the man down rather than reflecting affection or the man’s physical stature.

Related to Strategic Timidity is Euphemistic Softening, often from a place of not wanting to hurt another’s feelings. An example found is, when discussing a friend, a woman once said “---Mira qué tanto la camiseta no le cubre la pancita.” This is probably so as to not hurt the man’s feelings or to soften the blow so as to not invite too much ridicule.

Intensification is another oft-used function, one you may be familiar with. But it is broken into two parts. The first is the simplest. The suffix is used to emphasize the meaning of the base. “—Se lo dije lo primeritito que le vi” is an example of this emphatic intensification. The other part, Compensatory Intensification, is to mark neutral instances where the suffixed word form has become the norm through consistent use. An example of this is most uses of the word “chiquito,” which is more common than the word “chico” that it was based on, with almost no alteration in meaning.

Ironic or Playful Extravagance is often used in lively banter among friends. An example found is a man, when speaking of a fart, telling a friend celebrating his birthday, “!He aquí, el regalazo que te dejé!”

Less festively, there is also a Tenderness with Respect to Surrounding Reality. Not very self-evident, this function is often tied to grief and tragedy. When discussing the premature death of a child to its distraught parent, someone is recorded to have said, “-Diosito se lo llevó.” This is an effort, I believe, to soften the blow as much as possible of the death of the child while still giving the news to the parent.

Speaking of parents, the subsequent function, Baby Talk, is pretty simple. For example, a man referring to his girlfriend might say, “¿Tienes hambre, amorcita?”

The final category, rather nebulous and probably not of much use to us during this investigation, is Slang. Some instances of diminutives or augmentatives are just instances of Slang usage, without much rhyme or reason.

Outside of both domains, and which I will try to minimize as much as possible, are so-called Lexicalized Instances, which, despite my best efforts, you may encounter anyway. These are instances where the diminutive or augmentative is so common that they just become part of the word, or through such common usage begin to mean something different. The best example of this is the word *bonito*, which used to be a construction of *bueno* with the suffix -ito, meaning “little good.”

Below is a reference list with the codes we will be using for evaluating the instances, along with the relevant examples.

1. Intellectual Domain (can be metaphorical, mark with \*)
	1. Physical modification
		1. Increase of size, quantity, intensity, etc.
			1. “Le trajo a Juanito un *carrito*…”
		2. Decrease of size, quantity, intensity, etc.
			1. “…es una *casota* gigantesca…”
	2. Evaluative modification
		1. Positive valuation
			1. “Qué grandes *amigotes*…” (example from above)
		2. Negative valuation
			1. “Va a ser un *papeleazo*”
2. Emotional Domain
	1. Affective modifications
		1. Appreciation
			1. “Qué lindo *perrito*…”
		2. Depreciation
			1. “…ese huerquillo terrible.”
		3. Condescending Superiority
			1. “Mirá, hombrecito…”
		4. Impertinent Familiarity
		5. Respect and/or Courtesy
			1. “Por favor ayúdame, diosito”
		6. Humility
			1. “Nos lo dio el padrecito…”
		7. Strategic
		8. Euphemism
			1. “Sólo va durar un tantito.” It did not.
		9. Intensification
			1. Emphasis
			2. Compensation
		10. Extravagance
			1. “Jugué con los borrachotes allá”
		11. Sympathy/Empathy
			1. “Se le murió el perrito.”
		12. Baby Talk
			1. “Mira que perrititito que eres, amor”
		13. Slang
3. Lexicalized Instances (if in a saying, mark with two asterisks)
4. Name

S for Sarcasm

##### Animacy: X/Y/Z = feminine human/masculine human/other animate

Remember to ask yourself: *What is the* ending *doing in the word?*

### Appendix D

*Appendix D1 – Categorization Guidelines*

###### Introduction

Today, you will be evaluating what role(s) is/are being played by the augmentative and diminutive suffixes of Spanish keywords collected from Reddit when evaluated in context.

###### What are augmentatives and diminutives?

First, we will refresh your memory over what augmentatives and diminutives are. They are affixes, suffixes in the case of Spanish, which nominally adjust the size of, usually, nouns and adjectives. An example of a diminutive is the suffix -ito, as in *perr-ito*[[19]](#footnote-20). A common augmentative is -ote, as in *perr-ote. Perro*, as you may know, when unmodified, simply means “(male) dog.” When modified with -ito, the meaning often becomes “littledog.” Additionally, as you may already be aware, -ito is not the only diminutive suffix to exist in Spanish. A (non-exhaustive) list of the augmentative and diminutive endings evaluated in this study are provided in the appendix.

###### Semantic functions

Next, we will go over what a semantic function is. As you know, making things smaller is not the only function of a diminutive suffix, although it is its primary one. When someone says *mira esa* ***casita***, they might not always be talking about the smallest house being looked at, but often a house that they deem “cute.” In such a context, the diminutive’s function may not necessarily be **making things smaller** but making **things cuter**. This function (which we term affection) is one of the eight functions we will be evaluating, detailed in *Category descriptions* below, with relevant illustrative examples in the following *Examples from the dataset* section.

###### Description of data

Finally, the nature of the data is very important to your evaluation of them. These data were harvested from the social media site Reddit, which some of you may be familiar with. For those that are not, the format is forum-like, with a main post consisting of text, an image, a link to a video, etc., followed by comments about that topic, comments to comments, comments to comments to comments, and so on. Knowing exactly which comment is in response to which is a key piece of information, as is knowing what the overall post is about. However, due to the technical limitations of the harvesting procedures, these data are lost, and it is up to the rater to glean whatever information they can to piece together from the surrounding context and evaluate the meaning of suffix appended to the keyword.

In the dataset, different Reddit comments are often separated by sequences of numbers or nonsensical characters. The relation of the comment containing the keyword to the original surrounding Reddit comments is not explicitly provided in your tasks but can often be deducible through the surrounding context that is provided for you.

###### Layers of meaning

A concept relevant to your work in this study is the idea of **layers of meaning**, which can be visualized thus:

* Meaning of **utterance[[20]](#footnote-21)**
* Meaning of **sentence**
* Meaning of **word**

In this study, we are focusing on how the addition of augmentative and diminutive suffixes affect the meaning of keywords. This is not to be confused with the overall tone and/or meaning of the sentence or the greater utterance, concepts which are related to, and influence, the meaning of keywords and the selection of applied suffixes. You will see examples in *Examples from the dataset* that illustrate quite clearly a separation between the three layers of meaning, their separation in practice, and getting to the root question of **“what is the difference between the suffixed and un-suffixed word in this context?” Or “why did the speaker choose to use the suffixed word instead of the un-suffixed word?”**

However, a useful example of that separation in the meantime is the sentence: *No me gusta el* ***perr-ón****, pero el* ***perr-ito*** *sí es muy lindo*. “I don’t like the **big dog**, but the **little dog** is very sweet.” The significance of the sentence is about dislike and affection, two of the categories we will be identifying, but the speaker uses the augmentative and diminutive suffixes to differentiate between the two dogs using amplitude; the feelings of dislike and affection stem from other sources within the sentence.

###### Category descriptions

The categories of semantic functions are diagrammed below:

1. Affective
	1. Positive
		1. Affection
		2. Respect
	2. Negative
		1. Dislike
		2. Contempt
2. Neutral
	1. Amplitude
		1. Amplitude
	2. Intensity
		1. Accentuation
		2. Minimization
3. Lexicalization

You will be inputting the **first three letters** of the categories as codes into the Excel sheets, to make evaluating easier. Each evaluation can consist of an Affective category, a Neutral category, or a combination of Affective and Neutral categories (only one from each). The more important of the two should be placed first. Only mark a combination evaluation if both are strongly present.

Example: Acc/Dis

###### Amplitude

Modification to the physicality of the root word according to the nature of the ending itself. An *increase*, usually of size, in the case of augmentatives; a biggening. A *decrease*, also typically of size, in the case of diminutives; a smallening. Amplitude can also, less commonly, concern quantity or duration.

If we start out with a brick, modifying amplitude would be a physical enlargement or a shrinking of this brick.

###### Accentuation

This function is to intensify the intrinsic quality of the root word, either for emphasis or for playful exaggeration. If used on a friend, it might emphasize the strength of your friendship with them. If it is used on chocolate, it might highlight its chocolatiness. This option is often open for purely physical, purely abstract, and partially physical, partially abstract root words. If the modifications exist to adjust the amplitude of the object or its abstract manifestation, accentuation exists to make either more intense.

If we have a camera focused on the brick, this would be like a sharpening of the focus on the brick itself, intensifying the image.

###### Minimization

This function is to soften the intrinsic quality of the root word, often to distance or to euphemize. If used on a friend, it might de-emphasize the strength of your friendship with them. If it is used on chocolate, perhaps it plays down the sense of chocolatiness. Perhaps you must use the little boys’ room. As with accentuation, this option is often available for all three usual manifestations of nouns (purely physical, purely abstract, or a mixture). If the modifications exist to adjust the amplitude of the object or its abstract manifestation, minimization exists to blur either.

minimization serves to blur the focus on the brick or to adjust the lighting on the brick to lighten its shadow.

###### Affection

This function is to indicate a liking or a fondness. Like for soup, a friend, or a brother. Especially common with animate objects and common, useful items, usually with anthropomorphic tendencies.

Use it on the brick to display affection for it if it is cute or of sentimental value. Maybe you just like bricks in general.

###### Dislike

This function is used to indicate an antipathy or an aversion, perhaps even a disgust or revulsion. Maybe in reference to muddy socks, sticky heat, or blue cheese. Gross. Common to nouns at all levels of abstraction, but especially foods, people, or general categories.

Perhaps someone has thrown a brick through your window. Or maybe you just don’t like the look of this particular brick. Or bricks disgust you.

###### Respect

This function is used to indicate a deference or reverence towards a big accomplishment or important figure. Martin Luther King, Alexander the Great if you are not Iranian, Porfirio Díaz if you were Mexican and had money. You can respect your grandfather but probably not puppies. This option is generally only open to animate objects.

Perhaps this brick is a war hero. The cornerstone of a grand cathedral. Maybe it foiled an assassination attempt.

###### Contempt

This function serves to indicate disdain and even condescension. Perhaps towards the newest generation of spoiled kids or, God forbid, hipsters. Oddly, this option seems to be open to both animate and inanimate objects, but generally not abstract nouns.

Perhaps the brick voted for a political party you disagree with. Maybe bricks are a sign of the lower class and marble is the superior building material.

###### Lexicalization

There are several cases of lexicalized instances of augmentative and diminutive suffixes in Spanish. The funniest example is the lexicalization of the word *gans-ito*, which should mean “little goose,” but in Mexico and in other Latin American countries often refers to a packaged sugary confection instead.

The exclusion lists provided do the utmost to ensure that no lexicalized instances are included in the dataset. However, as always, some will slip through. If you keep evaluating the same word in the same way over and over, or if you get the feeling that the word means something besides ROOT + ENDING, look it up on SpanishDict or Wiktionary. If you find that it is a lexicalization, mark the first entry under the word as “4” and skip the rest of the instances.

###### Examples from the dataset

###### Note about the following examples

I am a Mexican-American heritage speaker of Spanish. The conclusions to which I arrive below may not be fully informed and may deeply conflict with your own. That is okay! That is the point of the study! These examples are just to highlight thought processes that will be helpful in evaluating your keywords.

###### Accentuation

* Discussing the moral decay of Latin American youth, especially in reference to American influence: *Son* ***huerqu-illos*** *mocosos que pretenden ser más* cooles[[21]](#footnote-22) *[*sic*]* *y maduros que sí son…* “They are snot-nosed **little ogres** that pretend to be cooler and more mature than they really are…”
	+ The overall tone is one of contempt towards the younger generation, but the semantic function of the diminutive in this case is one of either Accentuation of the belittlement already carried lexically by the word *huerco* and reinforced by the adjective *mocoso[[22]](#footnote-23)* or modifying the amplitude of the target, specifying the younger and/or smaller end of the spectrum defined by the word *huerco*, which often encompasses very small children up to mid-teenagers. The diminutive probably does not introduce contempt, it probably accentuates it instead.
		- Final answer: accentuation
* Discussing a show wherein two characters are mortal enemies: *Creía que Brayan y él eran muy* ***amig-otes****. No… he visto [el programa] en ya hace mucho tiempo. ¿Qué cambió?* “I thought that Brian and he were real buddies. I haven’t watched [the show] in a really long time. What happened?”
	+ There does not seem to be an evaluative undertone in the utterance or the sentence, merely confusion. The usage of *muy* as an intensifier (“real”) suggests that the augmentative, too, acts as an accentuator. It is unlikely to be a biggening, as there is not a sense of size involved at all. The diminutive seems instead to be highlighting, or accentuating, the level of friendship that existed between these two characters before the betrayal.
		- Final answer: accentuation

###### Minimization

* Discussing a woman who got a tattoo that reads “Daddy” on her inner lower lip: *“Abre sus* ***al-itas****…”* “ ‘She opens her **little wings**…’ ”[[23]](#footnote-24),[[24]](#footnote-25)
	+ Given that the entire phrase refers to a euphemistic subtext, it is logical to conclude that the tone of the comment is likely to be one of contempt. The metaphor in the song itself also includes the diminished form, which is likely meant to soften or minimize the impact of the metaphor.
		- Final answer: minimization
* Discussing the up-and-coming generation of politically active voters: *Lo que me da miedo es que todos los* ***joven-citos*** *vienen demasiado* ***zurd-itos****…un* ***poq-uito*** *más centrista estaría mejor.* “What scares me is that all of the **little young’uns** come in too **lefty**…a **little bit** more centrist would be better.”
	+ This is an excellent utterance that has three separate instances of diminutive, which we will unpack in turn. Overall, the tone of the utterance is neutral if a bit negative.
		- The first instance is perhaps a highlighting of just how young the new voters are, especially contrasting with the age (and probably the implied accompanying experience) of the commenter. This is an excellent example of accentuation.
			* Final answer: accentuation
		- The second instance is more loaded, especially from a socio-historical perspective. In much of Latin America, often due to the influence of the United States and its Central Intelligence Agency, the Communists and all such associated “leftists” are widely reviled, and the term is tainted with a negative connotation. To some, this taint is no longer held. It is likely, considering the context and the location (Argentina), as well as the word choice (*zurdo* means “left-handed”), that the commenter regards Communists and leftists negatively, suggesting a Minimizing function on the diminutive, to suggest that the young people don’t come onto the political scene as full-blown “lefties,” but that they do emerge a little more left than they would like.
			* Final answer: minimization
		- The third instance can be interpreted in a few ways. *Poco* already means “little,” so the inclusion of a diminutive could be an Accentual, emphasizing how small the increment to the right the commenter would like the future voters to enter the political scene as. Or the addition of the diminutive could be softening the request or notion that the youth should be a bit to the right, something that may not find widespread appeal among said youth, who might react poorly against the commenter, who might be thus attempting to preempt or immunize themselves from said potential backlash. This one is quite close, but I lean towards a minimization of the request for rightward movement among the youth.
			* Final answer: minimization.

###### Affection

* Concerning a photo of dogs: *¡Mira los* ***perr-itos*** *chul-itos!* “Look at the cute **little** **puppies**!”
	+ This one can either be concerning small dogs (so, Magnitude) or cute (Affection) dogs. Because the root *perro* does not have any lexically embedded evaluation (i.e., it is “neutral”), the addition of the diminutive is likely Affection, reinforced by the subsequent, similarly diminished, word *chulo*, which means “cute.” Especially since there is no semantic meaning to Accentuate (it is probably not conveying “more dog”).[[25]](#footnote-26)
		- Final answer: affection
* Thread concerning meat-related recipes: *Mmmm…huelo* ***carn-ita***, “Mmm…I smell **meat**”
	+ Because of the inclusion of several meat recipes, it is likely that this commenter is using “meat” as a general class noun and not referring to any meat in any individual recipe. This makes example unlikely to be Physically diminishing. The inclusion of the “smells good” interjection or conversational particle at the beginning strongly suggests affection.
		- Final answer: affection
* Thread concerning photos of an Old-World mansion and surrounding countryside (in Italy): *¡Guau! ¡Qué* ***cas-ota****!* “Wow! What a (big) **house**!”
	+ The photos, from the other comments, depict quite a beautiful landscape and mansion. The exclamatory tone of the comment suggests a positive overall evaluation. Because the word *casa* is used and not a lexically charged one like *choza* (“hut” or “shack”), it is likely that this is a co-occurrence of Magnitude and Affection, or even Respect (less likely.)
		- Final answer: magnitude/affection

###### Dislike

* When discussing genres of music: *No me gusta el* ***rap-ito***gangsta*. No es tanto que me disguste más que no me relaciono con él.* “I don’t like *gangsta* **rap**. It’s more that I just can’t relate to it than that it disgusts me.”
	+ The speaker seems not to have an obviously intense emotional overtone, they are simply stating, with an added caveat appended, that they simply do not like *gangsta* rap because they don’t relate to it, implying that there is no judgment of either the subgenre or the people who enjoy it. This could be Dislike, but the inclusion of the diminutive may be softening in nature, minimizing the judgment in anticipation of backlash, supported by the added caveat at the end.
		- Final answer: dislike/minimization

###### Respect

* Discussing the musical history of the band Soda Stereo: *Y ¡qué* ***tem-azo*** *es [*De música ligera*]!* “And what a song [*De música ligera*] is!”
	+ The exclamation sets up a good case for the utterance undertone to be positive. However, the interpretation of this ending may be confusing. It may be referring to how large of a hit the song was (it was and remains one of the most treasured songs to emerge from Argentina), meaning a magnification, even if abstract; it could be that the commenter has an Affection for this song. Or it can be an instance of Respect, especially since, later down the thread, the song is contrasted against later music of “less musically talented” regions of the Hispanosphere.
		- Final answer: magnitude/respect

###### Contempt

* Discussing international politics in Latin America: *Yo no soy* ***perr-ito*** *obediente de los yanquis anglosajones…* “I’m not an obedient **little dog** of the Anglo-Saxon Yankees…”[[26]](#footnote-27)
	+ The tone of the long comment of which this excerpt is a part is quite aggressive in its rejection of American interventionism in the Western Hemisphere, especially the assumption by the “Yankees” that Latin America does not have to be consulted about major policy decisions that will impact the region. The *perrito* portion is meant to distance the commenter from other collaborationist elements within his own country, which they regard with aggressively evident Contempt. This contempt also comes with an aggressive accentuation of one of the features that the author deems contemptuous about dogs (and thus brought by the marked root itself): their blind obedience.
		- Final answer: contempt/accentuation

###### Procedures

###### Excel spreadsheet

An Excel spreadsheet with your assignments will be provided to you via email, with the following format:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** |
| *Perrito* |  |  |  |  |
| *Gansito* |  |  |  |  |
| *Casita* |  |  |  |  |

The word on the left will represent the word you are evaluating in the *Concordances* tool of Sketch Engine, and the number will indicate which item down the list you will be evaluating. If you do not have an answer to one, skip both the number and the cell, making sure your evaluations and their indices agree.

###### Accessing the data

It is advised that you evaluate the data one rater at a time, as every time one loads the page while the other is on the account, you will be prompted to sign back in. However, once you have a screen open, it will not automatically kick you out of the account, only if you cause the interface to try to contact the Sketch Engine servers (usually to fetch new information). So, either block out separate times to evaluate, or evaluate at the same time, taking care not to press on anything too often. Blocking out separate times is *highly* recommended.

##### Login information

The login information for my Sketch Engine account is:

###### Username: cago4316@colorado.edu

Password: gmzSKEN@2022

Wordlist

Once logged into Sketch Engine, you will (usually) be redirected to the Dashboard. From there, head to the option labeled “Wordlist.” You will select “words” and “ending in” from the scroll menus, before entering the suffix you are evaluating. The complete list of suffixes you will be evaluating will be provided in the appendix at the end of the packet.

From there, you will click the checkbox marked “exclude these words.” You will have been emailed an assortment of text files with the correct suffix labeled as the name of the file. If you are evaluating the suffix -ito, access the file marked “N -ito exclusion list.” Copy the entire file to your clipboard and paste it onto the line below the checkmark. These files are provided as an attempt to cut down on the number of lexicalized instances muddying your data. They are by no means complete or even available for all the endings. If you believe you have encountered a lexicalized ending, check online. If so, simply move onto the next word.

Set the “frequency min” to zero. Then click “Go.”

Concordances

Once brought to the wordlist, go to the three dots to the right of the topmost word in the list. Select the little box with the arrow to the right of “Concordance” to be brought to the data you will be evaluating.

You will evaluate the keywords marked in red (known as concordances), with the context of the word displayed. If the displayed context is insufficient, you may press directly on the keyword itself and press the up and down arrows for more context to be loaded.

###### Submitting the data

If the information is on Google Docs, please make sure that I am in the list of editors. If the data is on an Excel spreadsheet, please email it to my school email address: cago4316@colorado.edu.

###### Category list

1. Positive
	1. Affection
	2. Respect
2. Negative
	1. Dislike
	2. Contempt
3. Neutral
	1. Physical change
	2. Abstract change
	3. Intensity
		1. Accentuation
		2. Minimization

###### List of augmentatives and diminutives

Augmentatives (and augmentative-adjacents)

* -ote, -ota
* -ón, -ona
* -azo, aza
* -udo, uda
* -ucho, -ucha
* -erío
* -ío
* -che
* -ísimo, -ísima

Diminutives (and diminutive-adjacents)

* -ito, -ita
* -illo, -illa
* -ín, -ina
* -uelo, -uela
* -ejo, -eja
* -ico, -ica
* -ete, -eta
* -ijo, -ija

*Appendix D2 – Summary Tables*

###### Augmentation and Diminution

|  |  |  |
| --- | --- | --- |
|  | **Augmentatives** | **Diminutives** |
| **Maximum** | arg | chi |
| **Minimum** | col | col |

Augmentative/Diminutive Suffix Usage

|  |  |  |
| --- | --- | --- |
|  | **Augmentatives** | **Diminutives** |
|  | **max** | **next** | **max** | **next** |
| **max** |  -ísimo | -azo | -ito | -illo |
| **max** | **min** | **max** | **min** | **max** | **min** | **max** | **min** |
| spa | chi/arg | chi | spa | else | spa | spa | else |
| **min** | -che | -iño |
| mex | arg/chi |

|  |
| --- |
| **Semantic Annotation by Primary Categorization** |
|  | **Augmentatives** | **Diminutives** |
| **max** | mag/acc | acc |
| **max** | **min** | **max** | **min** |
| mex/spa | chi | col | arg |
| **min** | min/res/dis | dis/res |

|  |
| --- |
| **Semantic Annotation by Any Inclusion** |
|  | **Augmentatives** | **Diminutives** |
|  | **pure** | **combo** | **pure** | **combo** |
| **max** | mag/acc | > | acc+con |  | acc | > | acc+con, acc+aff |
| **max** | **min** | **max** | **min** | **max** | **min** | **max** | **min** |
| mex | col, spa | col | arg,spa | col | mex | spa | col |
| **min** | res | min+X | res/dis | mag+res/min+dis/min+res |

|  |
| --- |
| **Positivity vs. Negativity by Primary Category** |
|  | **Augmentatives** | **Diminutives** | **ratio** |
| **max** | neg | pos | dim |
| **max** | **min** | **max** | **min** | **max** | **min** |
| col | mex | mex | spa | col\* | spa |
| **min** | Pos | neg | aug |
| **max** | **min** | **max** | **min** | **max** | **min** |
| arg | col | spa | col | arg | col |
| **Positivity vs. Negativity by Any Inclusion** |
|  | **Augmentatives** | **Diminutives** | **ratio** |
| **max** | neg | pos | dim |
| **max** | **min** | **max** | **min** | **max** | **min** |
| col | spa | else | col | chi | col |
| **min** | pos | neg | aug |
| **max** | **min** | **max** | **min** | **max** | **min** |
| spa | col | chi | col | spa | col |

|  |
| --- |
| ***Referentiality vs. Expressivity by Primary Category*** |
|  | **Augmentatives** | **Diminutives** | **ratio** |
| **max** | ref | ref | aug |
| **max** | **min** | **max** | **min** | **max** | **min** |
| arg | col | chi | col | mex | arg |
| **min** | exp | exp | dim |
| **max** | **min** | **max** | **min** | **max** | **min** |
| arg | mex | chi | col | col | spa |
| ***Pure Referentiality vs. any Expressivity*** |
|  | **Augmentatives** | **Diminutives** | **ratio** |
| **max** | EXP | ref | dim |
| **max** | **min** | **max** | **min** | **max** | **min** |
| col | mex | col | spa | col | spa |
| **min** | ref | EXP | aug |
| **max** | **min** | **max** | **min** | **max** | **min** |
| mex | col | spa | col | mex | col |

# Works Cited

Alonso, A. (1951). Noción, emoción, acción y fantasía en los diminutivos. In A. Alonso, *Estudios lingüísticos. Temas españoles* (pp. 195-229). Madrid: Editorial Gredos.

Augustyn, R., & Gniecka, A. (2011). Irony behind diminutives: a cognitive linguistic analysis of popular technical terms. *Lublin Studies in Modern Languages and Literature*(35), 31-49.

Gaarder, A. B. (1966, Dec.). Los llamados diminutivos y augmentativos en el español de México. *PMLA, 81*(7), 585-595.

Hägg, A. (2016). *A contrastive study of English and Spanish synthetic diminutives.* Master's Thesis, University of Oslo.

Lipski, J. M. (2012). Geographical and Social Varieties of Spanish: An Overview. In J. Hualde, A. Olarrea, & E. O'Rourke (Eds.), *The Handbook of Hispanic Linguistics* (pp. 1-26). Blackwell Publishing, Ltd.

Maíz-Arévalo, C. (2018). 'Solo un poquito'. El uso y funciones del diminutivo en español peninsular en dos grupos de Facebook. *Círculo de Lingüística Aplicada a la Comunicación, 73*, 33-52.

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: A Guide to the Methodology.* Cambridge University Press.

Stefanowitsch, A. (2020). *Handbook of Corpus Linguistics: A Guide to the Methodology.* Language Science Press.

Zuluaga Ospina, A. (1970). La función del diminutivo en español. *Thesaurus, 25*(1), 23-48.

1. Translator’s note: this seems to be referring to a colonel’s eagle. [↑](#footnote-ref-2)
2. Translator’s note: The “attritional” aspect is expounded upon and partly explained below, in the antepenultimate paragraph of this article [↑](#footnote-ref-3)
3. Translator’s note: this is probably a catch-all category for lexicalized instances of the suffixes. Cf. *bonito* (pretty) as its own lexical item instead of being analyzed as *bueno* (good) +*ito* (little). [↑](#footnote-ref-4)
4. Translator’s note: lit. “language of the students’ type” [↑](#footnote-ref-5)
5. Translator’s note: basically, nonce words and lexicalized instances [↑](#footnote-ref-6)
6. Here Zuluaga means that the diminutive meaning is already expressed by some other syntactic or contextual element, such as an adjective. [↑](#footnote-ref-7)
7. Dominican Republic was chosen because Cuba has an even smaller Reddit presence. [↑](#footnote-ref-8)
8. Likely already worked on and available to social media platforms such as YouTube to measure things such as “engagement” [↑](#footnote-ref-9)
9. It is notable that these specific metrics may be used to approximate the level of engagement from a given country on Reddit. Even more information could be gleaned when the volume is measured against population size. [↑](#footnote-ref-10)
10. “F” in these comments is shorthand for “paying respects,” a memetic response usually to a severe, often comedic, mistake. [↑](#footnote-ref-11)
11. Original: “Es el gran poder de la comida, es memoria y es tangible. Con mi abuela he querido hacer un recetario, tiene 90 años no sé cuánto tiempo más la tenga y quiero aprovechar cada minuto que pedo este aquí.” [↑](#footnote-ref-12)
12. Original “D\*mn, yo apenas perdí a mi abuela y yo que viví y crecí con mi abuela un tiempo por que [*sic*] mis papás trabajaban se extraña un chingo la comida y el relajo como te llevabas con tu abuelo ta [*sic*], pff, es algo que jamás conseguirás, disfruta tu Abuela.” [↑](#footnote-ref-13)
13. When referring to a male individual; cf. Eng. “guy.” [↑](#footnote-ref-14)
14. The use of both “qué pedo” and “no mames” are extremely recognizable hallmarks of Mexican Spanish. [↑](#footnote-ref-15)
15. Later revisions would increase the combined size of the Argentine and Chilean sub-datasets to 127. [↑](#footnote-ref-16)
16. A reminder that only up to one referential and one expressive semantic category can be given simultaneously in the latest annotation system. [↑](#footnote-ref-17)
17. The rearrangement is made along the axes mentioned in 4.2.1 (Revised Categories) [↑](#footnote-ref-18)
18. Magnitude is on opposite sides below because there is no separate category between sizing up and sizing down; previous experimentation found the frequency of augmentatives sizing down and diminutives sizing up minuscule enough to ignore entirely; thus, the value for cross-sizing is assumed to be 0% [↑](#footnote-ref-19)
19. Hyphenated for clarity. All subsequent augmentative- or diminutive-modified nouns will be so hyphenated in the instructional packet. [↑](#footnote-ref-20)
20. **Utterance** is the linguistic term for a single unit of speech, a stretch of words, sentences, and phrases preceded and followed by silence or a change in speaker [↑](#footnote-ref-21)
21. This is cool because the commenter seems compelled to cram the English loan adjective *cool* into the Spanish agreement system, pluralizing the adjective when convention leaves most recent English loanwords unmodifiable [↑](#footnote-ref-22)
22. The reason this is accentuating in this circumstance is because the root word *huerco* is almost exclusively used as a synonym for children and young adults. [↑](#footnote-ref-23)
23. The comment ends here; there is nothing else after the ellipsis [↑](#footnote-ref-24)
24. This is very likely a reference to a song by a Mexican rock band wherein the metaphor of a butterfly opening its wings is used to illustrate the “opening of the legs,” a euphemism for sex [↑](#footnote-ref-25)
25. Because of the nature of the extracted data, it is not possible to see what the thread author initially posted nor the hierarchy of the individual comments, so topics must be gleaned from context. [↑](#footnote-ref-26)
26. This quote *does* continue after the ellipsis, but is much too long to include [↑](#footnote-ref-27)