

Frecuencia y distribución geográfica del diminutivo *-ito/-ita* en la lengua oral y escrita del CORPES XXI

Katharina Gerhalter

Karl-Franzens-Universität Graz (Austria)  <https://www.doi.org/10.5209/clac.92605>

Recibido: 19 de noviembre del 2023 • Aceptado: 14 de septiembre del 2024

ES Resumen: Este artículo analiza la distribución diatópica de los diminutivos en *-ito/-ita* en el mundo hispanohablante. Exploramos de manera exhaustiva el CORPES XXI, contando un total de 267 000 casos (*tokens*) de diminutivos no lexicalizados. Para el análisis cuantitativo se tienen en cuenta tres parámetros: (i) la zona geográfica, (ii) la clase de palabra y (iii) el medio escrito o hablado. El estudio muestra que los diminutivos son más frecuentes en el subcorpus oral y en registros informales. Es también en el subcorpus oral donde se observan diferencias diatópicas más marcadas. De modo global, se confirma la mayor frecuencia de diminutivos en América frente a España, incluso si a *-ito/a* se añaden *-ico/a* e *-illo/a*. En el subcorpus oral, la frecuencia relativa más elevada se encuentra en las zonas andina, antillana y mexicano-centroamericana (con cifras muy semejantes entre sí). Con algo más de diferencia siguen el Caribe continental, el Río de la Plata y Chile. En cambio, en la lengua escrita es el subcorpus rioplatense el que muestra la mayor frecuencia relativa de diminutivos. Contrastando los resultados escritos y orales llegamos a la conclusión de que la lengua escrita rioplatense es más permeable a fenómenos coloquiales y familiares que la de otras variedades. En cuanto a las clases de palabra, predominan diminutivos en sustantivos comunes (especialmente en la variedad rioplatense), seguido de nombres propios y adjetivos. El uso más productivo en categorías menos prototípicas (adverbios, cuantificadores, artículos, numerales, demostrativos, posesivos, relativos y gerundios) se encuentra en México y Centroamérica, en la zona andina y en el Caribe continental.

Palabras clave: lingüística de corpus, diminutivos, variación diatópica, lengua hablada y escrita, variedades americanas.

ENG Frequency and geolinguistic distribution of the diminutive *-ito/-ita* in the spoken and written language of CORPES XXI

Abstract: This article analyzes the geolinguistic distribution of diminutives in *-ito/-ita* in the Spanish-speaking world. We exhaustively explored CORPES XXI, counting a total of 267,000 cases (*tokens*) of non-lexicalized diminutives. For the quantitative analysis, three parameters are taken into account: (i) geographical area, (ii) part of speech and (iii) written vs. spoken medium. This study shows that diminutives are more frequent in the spoken subcorpus and in informal registers. Also, in the spoken subcorpus, differences between varieties are more pronounced. Overall, the higher frequency of diminutives in America compared to Spain is confirmed, even if *-ico/a* and *-illo/a* are added to *-ito/a*. In the spoken subcorpus, the highest relative frequency is found in the Andean region, very closely followed by the Antillean and Mexican and Central American regions. The continental Caribbean, the Río de la Plata and Chile follow with slightly more difference. On the other hand, in the written language, it is the Río de la Plata subcorpus that shows the highest relative frequency of diminutives. Contrasting the written and oral results, we conclude that the written language of the Río de la Plata is more permeable to colloquial and familiar phenomena than other varieties. As for part of speech, diminutives predominate in common nouns (especially in the Río de la Plata variety), followed by proper names and adjectives. The most productive use of diminutives with less prototypical categories (adverbs, quantifiers, articles, numerals, demonstratives, possessives, relatives and gerunds) is found in Mexico and Central America, in the Andean area and in the continental Caribbean.

Keywords: corpus linguistics, diminutives, geolinguistic variation, spoken and written language, Spanish varieties in America.

Sumario: 1. Introducción. 2. Estado de la cuestión e hipótesis. 2.1. Diferencias diatópicas en el uso de diminutivos. 2.2. Estudios comparativos entre variedades. 2.3. Hipótesis para el presente estudio. 3. Breve descripción del CORPES XXI. 4. Metodología. 4.1. Lemas y formas. 4.2. Configuración de la búsqueda: el uso de comodines y operadores. 4.3. Configuración de la búsqueda: categorías gramaticales. 5. Resultados de la búsqueda por clases de palabras. 5.1. Nombres propios. 5.2. Sustantivos comunes. 5.3. Adjetivos y participios.

5.4. Adverbios. 5.5. Cuantificadores, artículos, numerales, demostrativos, posesivos y relativos. 5.6. Gerundios. 6. Distribución geográfica de los resultados. 6.1. Frecuencia global de diminutivos y agrupación de países en zonas. 6.2. Frecuencia de diminutivos según temáticas y tipología (variación diafásica). 6.3. Frecuencia de diminutivos según clases de palabra por zonas. 6.4. Distribución geográfica de los resultados en función del medio (escrito vs. oral). 6.4.1. Variación diatópica en el subcorpus oral. 6.4.2. Variación diatópica en el subcorpus escrito. 6.4.3. Divergencias entre datos escritos y orales. 7. Conclusiones. Referencias bibliográficas.

Cómo citar: Gerhalter, K. (2026). Frecuencia y distribución geográfica del diminutivo *-ito/-ita* en la lengua oral y escrita del CORPES XXI. *Círculo de Lingüística Aplicada a la Comunicación* 105 (2026): 161-186. <https://www.doi.org/10.5209/clac.92605>

1. Introducción

Los sufijos apreciativos y más en concreto los diminutivos son un tema ya clásico de la lingüística hispánica, que se viene analizando desde diversas perspectivas (variación morfológica, desarrollo histórico, lexicalización, valores semántico-pragmáticos, variación diatópica, diastrática y diafásica, etc.). En el presente artículo, nos centramos en la perspectiva diatópica y la relacionamos con la variación diafásica (lengua escrita vs. oral) y con las clases de palabras que pueden llevar sufijos diminutivos. Hasta el momento, los estudios cuantitativos que comparan variedades distintas no muestran un panorama claro, aunque muchos autores coinciden en que los diminutivos son, por lo general, más frecuentes y productivos en América que en España (véase apartado 2.1). Analizaremos si, efectivamente, tal distribución geográfica se puede observar en el CORPES XXI. Además, pretendemos averiguar en cuál de las zonas lingüísticas americanas (mexicano-centroamericana, caribeña continental, antillana, andina, chilena, rioplatense) se documenta la frecuencia más elevada de diminutivos, ya que estudios previos muestran resultados contradictorios al respecto.

El aporte novedoso de este artículo consiste, sobre todo, en la gran cantidad de datos analizados, ya que para dar una respuesta sólida a la pregunta planteada es preciso un corpus amplio. Para que el estudio cuantitativo pueda ser realmente representativo, es necesario recuperar —en la medida de lo posible— todas las ocurrencias de diminutivos no lexicalizados en *-ito(s)* e *-ita(s)* en el corpus investigado. Así, pretendemos mostrar, en un plano metodológico, el potencial y los límites de un corpus tan amplio como el CORPES XXI, sobre todo en lo que concierne la anotación automática y la representatividad de los materiales incorporados.

2. Estado de la cuestión e hipótesis

El presente estudio se centra en los diminutivos no lexicalizados, es decir, en aquellas formas en las que el sufijo es productivo: la forma diminutiva alterna con la no disminuida. Quedan excluidas, pues, formas lexicalizadas como *bonito* o *cigarrillo* (Real Academia Española 2009: § 9.3). En los diminutivos no lexicalizados, el sufijo diminutivo aporta distintos matices semánticos al lexema base. Desde los estudios de Alonso (1930, 1935) sobre los valores emocionales, relacionales y afectivos de los diminutivos, son innumerables los análisis de los diversos matices semánticos y pragmáticos que aportan los sufijos llamados ‘apreciativos’ (entre muchos otros, Hasselrot 1957, Fontanella de Weinberg 1962; Zuluaga Ospina 1970; Náñez Fernández 1973; Hummel 1997; Reynoso Noverón 2005). Los diminutivos no lexicalizados suelen expresar afecto, en especial familiaridad o cercanía, ironía, cortesía o incluso menosprecio. En muchos casos, estos valores afectivos se unen y se añaden al significado base de disminución de tamaño o importancia (Real Academia Española 2009: 651).

Vista la gran cantidad de bibliografía respecto a las funciones semántico-pragmáticas, no entraremos en detalle, sino que nos limitamos a citar la categorización de Reynoso Noverón (2005), que fue la base de muchas otras (por ejemplo, Malaver 2018; Malaver & Paredes García 2020; Jara 2021). Los valores del diminutivo se ordenan en una escala de más a menos objetivo:

- i. Valoración cuantificadora: Valoración de la dimensión de la entidad disminuida
 - a) Cuantificadora (uso referencial)
 - b) Descentralizadora (mitigadora)
 - c) Centralizadora (intensificadora)
- ii. Valoración cualificadora: valoración de las cualidades de la entidad disminuida
 - a) Negativa
 - b) Positiva
- iii. Valoración relacional: valoración de las relaciones del hablante con las entidades del discurso
 - a) Irónica
 - b) Amortiguadora
 - c) Respetuosa

(Reynoso Noverón 2005: 81)

No pretendemos llevar a cabo un análisis funcional-cualitativo exhaustivo, sino que simplemente nos referiremos a estas categorías para la interpretar los ejemplos citados. Nos centraremos, a continuación, en el aspecto geográfico, que constituye el factor de mayor interés para nuestro estudio.

2.1. Diferencias diatópicas en el uso de diminutivos

Los sufijos diminutivos suelen usarse con mayor frecuencia en América en comparación con España. Las diferencias entre el español en América y el de España no son solamente cuantitativas, sino que también

conciernen las clases de palabras: en América, se han documentado formas diminutivas de prácticamente todas las categorías gramaticales (Aleza Izquierdo 2010: 195). En este sentido, los diminutivos son más productivos en América que en España. Por ejemplo, es más fácil encontrar diminutivos formados sobre adverbios en América que en Europa: *ahorita*, *ahicito*, *allacito*, *alredorcito*, *nomasito*, etc. (Real Academia Española 2009: 632). Diminutivos en demostrativos (*estito*, *aquellito*) y en posesivos (*suyita*, *tuyito*) se aceptan en la lengua popular andina, caribeña y centroamericana (Real Academia Española 2009: 634). Numerales con sufijo diminutivo (*unito*, *dosito*, *cuatrito*) se documentan particularmente en el área andina (Real Academia Española 2009: 633; 1398), donde el diminutivo llega a usarse incluso con pronombres clíticos en imperativos: *bájemelito* (Aleza Izquierdo 2010: 196).

Precisamente el español andino es una variedad que repetidamente se destaca en la bibliografía como especialmente inclinada hacia un uso abundante de sufijos diminutivos. Esta variedad resalta no solo por la alta frecuencia de uso de tales sufijos y la aplicación a bases que no permiten estas terminaciones en otras variedades del español, sino también por los numerosos valores pragmáticos que presentan (Escobar 1998; Ramírez Luengo 2016: 112). Esta característica se ha relacionado con posibles influencias de las lenguas quechua y aimara, lenguas muy ricas en morfemas que expresan no sólo tamaño pequeño, sino también emociones, afecto, desprecio, cortesía, etc., especialmente en la lengua hablada. Ciertos sufijos del quechua y del aimara presentan valores pragmáticos como los que presenta el sufijo español *-ito* actualmente en la región andina (Escobar 1998, 2000: 96). Destaca particularmente el uso del diminutivo con valor de cortesía y modestia, incluso en registros no familiares y no afectivos, es decir: como señal de cortesía deferencial (Aleza Izquierdo 2010: 197; Escobar 1998, 2016: 358-359).

La hipótesis de la proliferación de los diminutivos en las variedades andinas como consecuencia del contacto lingüístico con lenguas indígenas como el quechua y el aimara es ampliamente aceptada (Fernández Lávaque 1998; Escobar 2000; Reynoso Noverón 2001; Ramírez Luengo 2016; Jara 2021). Para ello, Ramírez Luengo (2019) aporta un estudio de caso anecdótico de Bolivia: la comparación entre cartas y diarios de una hablante mono- y otro bilingüe del siglo XIX muestra un uso más productivo de diminutivos por parte del segundo.

Además, al lado de la variedad andina, Reynoso Noverón (2001) destaca también la variedad mexicana como especialmente productiva de diminutivos. En este caso, se explicaría por una posible influencia del náhuatl, que también muestra un uso abundante de marcas morfológicas de apreciación y disminución (Reynoso Noverón 2001: 63-64).

2.2. Estudios cuantitativos comparativos entre variedades

El estudio empírico de Reynoso Noverón (2001) se basa en datos orales y escritos de cuatro variedades; dos de carácter mestizo con contacto prolongado con lenguas indígenas, concretamente con el náhuatl (México) y el aimara y quechua (Bolivia y Perú), y dos variedades que carecen de un adstrato indígena (Buenos Aires y Madrid). Los datos cuantitativos se basan en aquellos diminutivos que son productivos (Reynoso Noverón 2001: 77) y muestran que las primeras dos variedades usan más diminutivos que los últimos: en el corpus analizado, hay un porcentaje de 0.53 % de diminutivos en México y un 0.31 % en los Andes, frente a un 0.17 % en Madrid y un 0.13 % en Buenos Aires (los porcentajes se refieren al porcentaje de *tokens* de diminutivos sobre el total de palabras; véase Reynoso Noverón 2001: 103-106). El uso de sufijos diminutivos es especialmente elevado en las sociedades mestizas urbanas de México y de los Andes, que muestran también la gama semántico-pragmática más amplia y un mayor empleo de usos subjetivos (Reynoso Noverón 2001: 228, 230). La hipótesis de una posible influencia indígena defendida por Reynoso Noverón (2001) se limita a dos zonas concretas y, obviamente, no se puede extender a todas las lenguas indígenas de América. Además, el estudio se restringe a cuatro variedades seleccionadas y no tiene en cuenta otras áreas como el Caribe (antillano y continental) o Chile para probar la hipótesis planteada.

Los datos varían mucho de un estudio a otro. Por ejemplo, Hu (2022) analiza series de televisión de España, México, Colombia, Argentina y Chile (veinte horas por cada país). La distribución del total de los diminutivos no lexicalizados por orden geográfico es: Colombia (31 %), Chile (25 %), Argentina (20 %), México (16 %) y, por último, España (8 %), siempre contando *tokens* (Hu 2022: 131). Se confirma, pues, que en España se usan considerablemente menos diminutivos que en América, pero entre las regiones americanas, el último puesto de México contradice a los resultados de Reynoso Noverón (2001). Probablemente, la muestra analizada no es suficientemente representativa (1-2 series por zona) como para que estos datos sean significativos. De hecho, los resultados muestran que, más que depender del país, el uso de diminutivos depende, sobre todo, del género de serie: es más frecuente en comedias familiares y menos en series policíacas (Hu 2022: 132).

El estudio de Criado de Diego & Andión Herrero (2016) en un corpus propio y en los corpus CREA y CORPES XXI (en una versión anterior) también constata una frecuencia más baja de diminutivos en España. En la muestra sacada del CREA, la frecuencia relativa (de *tokens*) más alta de diminutivos se documenta en el español rioplatense, el caribeño y el chileno, mientras que los resultados para la zona andina y mexicana son más bajos y se aproximan a los de España (Criado de Diego & Andión Herrero 2016: 100). Sin embargo, creemos que la selección de datos también es demasiado reducida como para ser representativa, ya que se basa en una selección previa de 50 diminutivos (*types*) no lexicalizados, lo cual es un número muy bajo (compárense con los 6 000 *types* —sin contar los nombres propios— detectados en el presente trabajo, ver sección 5).

Un corpus que gracias a su diseño permite comparaciones entre variedades orales distintas es el corpus de entrevistas PRESEEA, para el que se ha propuesto un esquema de análisis que diferencia entre

diminutivos lexicalizados y productivos. El estudio de Malaver & Paredes García (2020), basado en las entrevistas de las ciudades de Medellín (en los Andes), Caracas (en el Caribe) y Madrid (España), revela lo siguiente: La frecuencia relativa más alta de diminutivos no lexicalizados (*tokens*) se encuentra el corpus de Medellín (proporción 0.499), seguido de Madrid (0.342) y Caracas (0.293; cf. Malaver & Paredes García 2020: 324). Las dos variedades americanas son más productivas en cuanto a la creación de distintos tipos de diminutivos (es decir: se usan diminutivos en más palabras distintas/*types*), por ejemplo, en el caso de los adverbios (Malaver & Paredes García 2020: 327, 329-330).¹

2.3. Hipótesis para el presente estudio

Mientras que casi todos los estudios citados coinciden en que España es claramente la zona de la frecuencia más baja de diminutivos, no parece haber una imagen clara en cuanto a qué variedades americanas usan más diminutivos que otras. Por ejemplo, en el estudio de Reynoso Noverón (2001) se documentan notablemente más diminutivos en México y en los Andes que en Buenos Aires. En cambio, el estudio de Criado de Diego & Andión Herrero (2016) invierte por completo esta tendencia. En Hu (2022), la variedad que más diminutivos documenta es la colombiana. Por último, en Malaver & Paredes García (2020), la ciudad andina de Medellín sobrepasa a la ciudad caribeña de Caracas.

Creemos que estos resultados contradictorios se deben a problemas metodológicos. Las cifras de diminutivos parecen ser particularmente sensibles a factores pragmáticos, individuales y diafásicos, tales como la selección de informantes, de contextos más formales o informales, y de registros (familiar, etc.). De hecho, las diferencias diastráticas entre los informantes no son especialmente significativas para los diminutivos (Malaver & Paredes García 2020), mientras que factores como el grado de conocimiento entre los interlocutores influyen gravemente en la frecuencia de diminutivos. Por ejemplo, en un corpus escrito sacado de redes sociales, “el diminutivo se prefiere en aquellas relaciones en las que los interlocutores gozan de una mayor cercanía y solidaridad [0.65 % de diminutivos], mientras que su uso es marcadamente inferior entre aquellas cuya relación es de carácter más distante [0.06 % de diminutivos]” (Maíz-Arévalo 2018: 48). Asimismo, los diminutivos en nombres propios son significativamente más frecuentes en cartas que en otros textos escritos (Calderón Campos 2024: 72).

Teniendo en cuenta el factor diafásico, probablemente los resultados de las entrevistas semiformales de PRESEEA, con un bajo grado de conocimiento entre los interlocutores, sean más representativos para la comunicación semiformal y menos para la comunicación familiar. De hecho, el uso de diminutivos aumenta ligeramente a medida que avanzan las entrevistas y aumenta la familiaridad entre los interlocutores (Paredes García 2015: 146-147). Asimismo, los diminutivos son más frecuentes en comedias familiares que en series policíacas (Hu 2022: 132) y, en México y Buenos Aires, son más frecuentes en lengua oral que en la escrita (Reynoso Noverón 2011: 122, 129).

Por lo tanto, en diferentes configuraciones situacionales (medio, situaciones comunicativas, relaciones entre hablantes, grados de formalidad, intenciones de los interlocutores, etc.) se obtienen resultados muy distintos acerca del uso de diminutivos. En este sentido, los resultados cuantitativos de los estudios citados quizás dependen simplemente de la elección de fuentes, que en ocasiones no llegan a ser muy extensas. También es posible que las fuentes elegidas para contrastar variedades distintas no sean totalmente equivalentes en cuanto a su configuración situacional. Para contrarrestar estas posibles carencias, creemos necesario contar con una base de datos mucho más amplia, procedente de un mayor número de informantes y basada en una gran variedad de registros para así cubrir la amplia gama de configuraciones situacionales. Para ello, se ofrece el CORPES XXI, que dentro de los corpus disponibles en este momento se considera el corpus de referencia por excelencia gracias a su tamaño, equilibrio, diversidad de fuentes y calidad de anotación (Moreno Sandoval 2022: 410). Precisamente por la gran cantidad de fuentes, el CORPES XXI permite comparar distintas subsecciones entre sí (Rojo 2016) y es el indicado para investigaciones como la que pretendemos aquí.

Como muestra el repaso bibliográfico, la hipótesis de que los diminutivos son más frecuentes en América que en España no es nada nueva ni sorprendente. Más bien pretendemos ponerla a prueba y respaldarla en un corpus grande como el CORPES XXI. Para la distribución de los diminutivos dentro de América, ponemos a prueba la hipótesis de Reynoso Noverón (2001) y Escobar (2000) de que los diminutivos son especialmente frecuentes y productivos en la región andina y mexicana. Obviamente, los datos del CORPES XXI sólo pueden constatar la frecuencia de uso actual en distintas variedades, pero no nos permiten probar posibles causas (como las supuestas influencias indígenas en dichas dos zonas).

Con respecto a las clases de palabras, nuestra hipótesis consiste en que se documente una mayor variedad de diminutivos en América que en España, o sea, esperamos un mayor número de *types* distintos en América (por ejemplo, en adverbios, numerales, demostrativos, gerundios, etc.).

Una tercera hipótesis concierne el medio: debido al carácter coloquial, familiar, cercano e informal de los diminutivos (véase, por ejemplo, Paredes García 2015: 143-144; Maíz-Arévalo 2018), esperamos una frecuencia más elevada de diminutivos en la lengua oral que en la lengua escrita. Aunque no se pueda equivaler la

¹ El estudio de Malaver & Paredes García (2020), además, contradice a la propuesta de Company Company (2012), basada a su vez en los datos de Reynoso Noverón (2001), de que la variedad mexicana usa más diminutivos con funciones subjetivas que la variedad española. Según Malaver & Paredes García (2020: 332-334), la función cuantificadora es más frecuente en Caracas y en Medellín que en Madrid; en cambio, Madrid sobrepasa notablemente a las ciudades venezolanas en cuanto a la función descentralizadora. En el presente estudio, es imposible analizar los valores semántico-pragmáticos dada la gran cantidad de ejemplos. Por lo tanto, este aspecto queda todavía por analizar.

variación diafásica con la variación entre escrito y oral –véase especialmente Koch y Oesterreicher (2011) sobre oralidad y escrituralidad conceptual y medial– creemos que los datos relativos al medio (escrito u oral) del CORPES XXI pueden dar indicios sobre variación diafásica.

3. Breve descripción del CORPES XXI

El *Corpus del español del siglo XXI* es un corpus de referencia elaborado por la Real Academia Española en colaboración con la *Asociación de Academias de la Lengua Española*. Para esta investigación, se usó en la versión 1.0. (noviembre de 2023), que cuenta con más de 395 millones de formas ortográficas (es decir, ‘palabras’), en su gran mayoría de origen escrito. Sólo un 10 % de las fuentes son orales. Se trata, pues, de un corpus enorme en cuanto a la cantidad de palabras. Como subrayaba Rojo (2017: 125) ya para la versión anterior del CORPES XXI (con todavía menos palabras), es deseable aspirar al mayor tamaño posible de un corpus para poder obtener documentación de elementos con frecuencias muy bajas. En el caso que nos ocupa en este artículo, el tamaño de COPRES XXI probablemente permite documentar también diminutivos menos frecuentes como, por ejemplo, en artículos, posesivos, numerales, etc.

Desde el inicio, el diseño del CORPES XXI buscó una representación dialectal proporcional (Rojo 2016; Moreno Fernández 2022: 301). La distribución geográfica de los textos incorporados refleja aproximadamente la proporción de hablantes por país, pero todavía con una sobrerrepresentación de España: un 30 % del corpus proviene de España y un 70 % de América (la proporción real del número de hablantes estaría más bien entorno al 10 % vs. 90 %). Las zonas lingüísticas establecidas por la *Asociación de Academias de la Lengua Española* para América son la zona Andina (Bolivia, Ecuador, Perú), las Antillas (Cuba, Puerto Rico, República Dominicana), el Caribe continental (Colombia, Venezuela), Chile, México y Centroamérica (Costa Rica, El Salvador, Guatemala, Honduras, Nicaragua, Panamá) y el Río de la Plata (Argentina, Paraguay, Uruguay).

Excluimos del presente estudio los datos de Estados Unidos, Filipinas y Guinea Ecuatorial. La representación del español estadounidense es problemática, ya que no se trata de una variedad homogénea, sino que altamente variable. En el caso de Filipinas y Guinea Ecuatorial, el número de fuentes es probablemente demasiado escaso como para poder considerarse representativo (Moreno Fernández 2021: 934).

La identificación de países enteros con una determinada zona lingüística no está totalmente libre de problemáticas. Por ejemplo, podría ser cuestionable en el caso de Argentina y quizás Chile: el norte de Argentina no pertenece a la variedad rioplatense, sino a la andina, como también lo hace el norte de Chile. Asimismo, las costas de Venezuela y Colombia sí forman parte de la variedad caribeña costera (tierras bajas), pero los interiores montañosos formarían parte del español andino (tierras altas). Para una discusión de las distintas divisiones propuestas para el español en América y una división basada en la percepción propia de los hablantes, remitimos para más detalles, entre otros, al trabajo de Quesada Pacheco (2014). No obstante, por simples razones metodológicas, es comprensible que, dentro del CORPES XXI, un país se adscriba a la zona lingüística que mayoritariamente representa. En nuestros datos, tendremos en cuenta tanto los resultados por zonas como por países particulares. A través del ejemplo de los diminutivos, verificaremos si la agrupación en zonas se mantiene.

4. Metodología

El tamaño del CORPES XXI es, principalmente, una ventaja: a mayor tamaño, mayor representatividad. Al mismo tiempo, sin embargo, la gran cantidad de resultados imposibilita leer y corregir todos los resultados y distinguir los usos lexicalizados de los productivos. El objetivo de esta sección consiste en encontrar la manera más exhaustiva posible para poder contar todos (o al menos casi todos) los casos registrados de diminutivos no lexicalizados en el CORPES XXI (sobre distintas soluciones para buscar diminutivos en diversos corpus del español, véase también Calderón Campos 2024: 60-61). No buscamos por formas concretas (*abuelito*, *bajitas*, etc.), ya que esto supondría que hay que saber de antemano la lista completa de formas documentadas. A través de las búsquedas esquemáticas por las terminaciones *-ito*, *-itos*, *-ita* e *-itas* que ilustraremos a continuación, el CORPES XXI revela aproximadamente 267 000 ocurrencias o casos en total (*tokens*) de unas 14 000 formas distintas (*types*) (véase la tabla 2 en la sección 5). Vistos estos números, no sería factible buscar manualmente forma por forma (*type* por *type*). A continuación, describimos nuestras búsquedas para que puedan ser reproducidas en un futuro (por ejemplo, en una futura versión más ampliada del mismo corpus o en una subsección específica).

4.1. Lemas y formas

El CORPES XXI es un corpus lematizado, lo cual significa que a cada forma ortográfica se la ha asignado su lema base. Esta lematización supone una gran ventaja respecto a versiones anteriores como el CREA. Por ejemplo, el estudio de Criado de Diego & Andión Herrero (2016: 99) en el corpus CREA se basa en la selección previa de 50 diminutivos (*types*), lo cual ciertamente reduce la representatividad de sus resultados. En comparación, la lematización del CORPES XXI permite prescindir de listas y buscar una gama mucho más amplia de formas.

En el caso de los diminutivos, las formas disminuidas tales como *abuelito* o *abuelitas* se han lematizado como *abuelo*. Las formas con diminutivo se consideran, pues, variantes o alternativas de las formas sin diminutivo. Por ejemplo, el lema *abuelo* agrupa las formas *abuela*, *abuelo*, *abuelos*, *abuelas*, *abuelita*, *abuelito*, *abuelitos*, *abuelitas*, *abuelete*, *abueletes*, *abuelín*, *abuelillo*, *abuelilla*, *abuelota*, *abuelillos*, *abuelica*, *abuelote* (lista exhaustiva de todas las formas asociadas al lema *abuelo*, ordenada según frecuencia y recuperada a través de la función “inventario”). Nos limitamos al sufijo *-ito/a*, ya que es, con diferencia, el más frecuente

del mundo hispanohablante (véanse por ejemplo los estudios de Criado de Diego & Andión Herrero 2016 y de Malaver & Paredes García 2020), dejando de lado las variantes *-illo/a*, *-ico/a*, *-ete*, *-ín*, *-ote/a*, que sin duda merecerían otro estudio de enfoque diatópico. De manera ilustrativa, presentaremos en el apartado 6.1. una breve comparación cuantitativa de *-ito* con *-ico* e *-illo*.

En este trabajo, diferenciamos entre $type_{FORMA}$ y $type_{LEMA}$, de acuerdo con la denominación *lemas y formas* en el CORPES XXI: las variantes flexionadas como *bajito*, *bajita*, *bajitos*, *bajitas* son cuatro $types_{FORMA}$ distintos. Juntos pertenecen a un solo $type_{LEMA}$: *bajo*.

Para filtrar solamente aquellos diminutivos que consideramos productivos y no lexicalizados, nuestra estrategia consiste en buscar las formas que terminan en *-ito/s* o *-ita/s*, pero cuyo lema no contiene tal terminación, como es el caso de *abuelito*, lematizado como *abuelo*. Es decir, la lematización del CORPES XXI impone el criterio de qué formas se consideran lexicalizadas y cuáles no. No es una decisión nuestra, sino que esta decisión queda necesariamente relegada de antemano al corpus. Generalmente, los anotadores automáticos como los del CORPES XXI se basan en diccionarios para la lematización (Moreno Sandoval 2022). Por lo tanto, suponemos que los diminutivos que se lematizan como tal en el CORPES XXI son aquellos que cuentan con entradas propias en los diccionarios de referencia y pueden considerarse lexicalizados (sobre la lexicalización de diminutivos especialmente en el español americano, véase, por ejemplo, Aleza Izquierdo 2017). Entre ellos se encuentran casos como *bonito/a*, *mesita*, *granito*, *mosquito*, *carrito*, *señorito/a* o *pajarito* y muchos más. Estos se excluyen necesariamente de nuestro análisis, aunque sería cuestionable si no habría que incluir a algunos de ellos (ciertas documentaciones de *pajarito*, por ejemplo) en el presente estudio.

La lematización de los diminutivos no siempre es del todo coherente: los casos de *viejita* y *viejitas* están lematizados con el lema *viejita*, mientras que *viejito* y *viejitos* llevan el lema *viejo*. Por lo tanto, *viejita/s* se interpreta como un diminutivo lexicalizado que se excluye en nuestra búsqueda, mientras que *viejito/s* se considera un diminutivo productivo y se incluye. Esa asimetría habría que arreglarla manualmente, pero frente al gran inventario de casos que habría que repasar manualmente, no es una tarea factible dentro de lo que nos proponemos para el presente artículo. Considerando el tamaño del CORPES XXI, nos conformamos con la lematización propuesta.

Por último, los nombres propios con sufijo diminutivo son una excepción, ya que se han lematizado sistemáticamente como tales. Por ejemplo, *Anita* y *Juanito* están lematizados como *Anita* y *Juanito*, respectivamente, y no como *Ana* y *Juan*. Consecuentemente, en el caso de los nombres propios no se pudo usar la metodología general de excluir toda forma en *-ito* que fuera también lema.

4.2. Configuración de la búsqueda: el uso de comodines y operadores

El buscador del CORPES XXI conoce una serie de comodines que facilitan la búsqueda: el asterisco *** representa una serie de caracteres de cualquier tipo. Así pues, buscar **ito* como *forma* recupera todas las palabras que terminan en *-ito*. La combinación de varias búsquedas en una sola búsqueda se puede efectuar mediante la barra larga. De este modo, la búsqueda por **ito|*ital|*itos|*itas* en la casilla *forma* recupera simultáneamente las cuatro variantes. Al especificar sólo la terminación de las formas, se incluyen posibles casos de acumulación de varios sufijos en una palabra (como *chiquitito*), así como variantes morfológicas en *-cito*.

Ahora bien, la simple búsqueda por las formas **ito|*ital|*itos|*itas* recupera todos los ejemplos de cualquier palabra que termina en una de estas cuatro secuencias. Entre los resultados se encuentran sustantivos y adjetivos como *propósito*, *gratuito*, *favorita* y diminutivos lexicalizados como *bonito* o *mesita*. Para analizar el uso realmente productivo de los sufijos diminutivos, es necesario excluir estos casos. Para ello, aprovechamos el hecho de que estas formas están lematizadas como tales. Excluimos, pues, en la casilla *lema* las cuatro terminaciones mediante el operador *!* (que significa 'no'): *!*ito|*ital|*itos|*itas*. Así, solucionamos el problema de búsqueda en el CORPES XXI al que se enfrenta, por ejemplo, Calderón Campos (2024) en su estudio de diminutivos en diversos corpus.

En el caso de los nombres propios que terminan en *-ito*, tuvimos que excluir mediante el operador de exclusión aquellos lemas que no son pertinentes (por ejemplo, *Margarita*, *Quito*, *Tito*, *Benito*, etc.).

4.3. Configuración de la búsqueda: categorías gramaticales

Cada forma ortográfica ('palabra') está etiquetada con una categoría gramatical (sobre el funcionamiento de etiquetadores morfosintácticos, véase Moreno-Sandoval 2022). Al igual que en el caso de algunas lexicalizaciones inconsistentes, hay que advertir también sobre algunas clasificaciones categoriales inconsistentes. Dada la gran cantidad de datos en el CORPES XXI, la anotación de las categorías gramaticales solo se ha podido hacer de manera semiautomática y esquemática. Por ejemplo, los participios se clasifican sistemáticamente como verbos, sin considerar el contexto sintáctico (véase el §5.3 sobre los participios con función adjetival). El adverbio *poquito* en la locución *un poquito* se ha anotado como sustantivo, probablemente debido al contexto inmediato: por regla general, a un determinante como *un* le sigue un sustantivo. Esto se debe al funcionamiento de la anotación automática en un corpus tan grande como el CORPES XXI: el reconocimiento de la clase de palabra se orienta por el contexto sintáctico de aparición, siendo locuciones formadas por más de una palabra uno de los problemas persistentes de anotadores automáticos (Moreno-Sandoval 2022: 404). Otra categorización discutible es la clasificación de la forma *m'hijita* (forma de tratamiento o vocativo) como interjección. Por último, también hay categorías transversales dentro de los determinantes: tanto *unito/a* como sus bases *uno* y *una* se clasifican de artículo, numeral o cuantificador, es decir, pertenecen a tres categorías paralelas.

En todo caso, la separación por categorías gramaticales conviene desde un punto de vista metodológico, ya que los lemas a excluir varían en función de la clase de palabra. Para la presente investigación buscamos

las categorías gramaticales con sus correspondientes restricciones tal y como aparecen en la tabla 1. Las restricciones específicas en la casilla *lema* surgieron de un proceso retroalimentado: a medida que íbamos comprobando los resultados de las búsquedas en el “inventario de formas”, iban surgiendo ciertas formas no pertinentes para cada categoría gramatical que había que ir excluyendo, hasta que llegábamos finalmente a un “inventario de formas” limpio. Para aquellas categorías gramaticales que no aparecen en la tabla no hubo resultados pertinentes.

Tabla 1. Búsquedas específicas por clases de palabras

Categoría gramatical	Lema	Forma	Comentario
Sustantivos comunes	!*ito!*ita!*itos!*itas	*ito *ita *itos *itas	Busca todas las formas terminadas en <i>-ito</i> , <i>-ita</i> , <i>-itos</i> , <i>-itas</i> , excepto las que tienen tal terminación en su lema.
Adjetivos	!*ito!*ita!*itos!*itas	*ito *ita *itos *itas	Íbid.
Adverbios	!*ito!*ita!*itos!*itas	*ito *ita *itos *itas *itamente	Íbid.
Cuantificadores		*ito *ita *itos *itas	No hay restricciones necesarias
Numerales		*ito *ita *itos *itas	Íbid.
Demostrativos		*ito *ita *itos *itas	Íbid.
Artículos		*ito *ita *itos *itas	Íbid.
Posesivos		*ito *ita *itos *itas	Íbid.
Relativos		*ito *ita *itos *itas	Íbid.
Sustantivos propios	!quito!margarita!tito!benito!san benito!quitos!hipólit!*rita!santa rita!fito!cáritas!afrodita!zita!caperucita	*ito *ita *itos *itas	Los nombres propios con sufijo diminutivo están lematizados como tales; por eso, hay que excluir manualmente cada nombre propio no pertinente
Participios (a través de “verbo” > “tiempo”)	!*scribir!*freír	*ito *ita *itos *itas	Excluye formas como <i>escrito</i> , <i>descrita</i> , <i>prescritos</i> , <i>transcrito</i> , <i>frito</i> , <i>refrito</i> , <i>sofrito</i> , etc.
Gerundios (a través de la categoría gramatical “desconocido”)		corriendito llegandito andando pasandito saliendito caminandito cantandito co leandito conversandito des eandito ensaliendito jugan dito leyendito oscureciendi to oscurociendito quedand ito resbalandito tirandito	Búsqueda manual por todas las formas, ya que los gerundios con sufijo diminutivo no están lematizados/etiquetados
Interjecciones		*ito *ita *itos *itas	Filtro manual: recuento de los casos de <i>cuidadito</i> y <i>m'hijita</i>

La anotación automática del CORPES XXI, obviamente, tiene sus ventajas y desventajas. Por un lado, permite procesar grandes cantidades de texto, lo cual aumenta la representatividad de los resultados: cuanto más material disponible, más representativos son los datos cuantitativos. Por otro lado, la anotación automática tiene sus límites: en ocasiones, los resultados requieren corrección manual. Es el caso de formas menos frecuentes, especialmente de diminutivos más ‘raros’, que se escapan de las instrucciones de anotación automática. Los anotadores automáticos generalmente funcionan sobre la base de textos anotados manualmente que les sirven de material de entrenamiento (Moreno Sandoval 2022). Si en estos textos de entrenamiento ciertos diminutivos no están documentados, el etiquetador automático del CORPES XXI no los reconoce y los etiqueta como “desconocido”. Por eso, repasamos también los resultados de *ito|*ita|*itos|*ita* en la categoría “desconocido”.

Un caso relevante son los gerundios con sufijo diminutivo: en contra de lo esperable, la búsqueda de la categoría “verbo gerundio” con la terminación **ito|*ita|*itos|*itas* no da ningún resultado. Al parecer, por ser una estructura poco frecuente, el lematizador automático del CORPES XXI ha sido entrenado con material en el que estas formas son escasas o ausentes. Solamente *callandito* aparece como adverbio y se incluye, pues, en los resultados de adverbios. En realidad, *callandito* no se ha anotado con su clase de palabra (que sería “verbo gerundio”) sino con su función en la oración (función adverbial). La anotación automática confunde, pues, el adverbio como clase de palabra con la función adverbial de un constituyente en una oración.

Al contrario de *callandito*, muchas otras formas como *corriendito*, *andandito*, *llegandito*, *caminandito*, *cantandito*, etc. llevan la marca gramatical “desconocido” y no están lematizados. Por lo tanto, el rastreo se tuvo que hacer manualmente. La solución propuesta en la tabla 1 se basa en una búsqueda previa de la forma *ndito|*ndita|*nditos|*nditas en la clase gramatical “desconocida”. Entre los resultados del “inventario de formas”, se han apuntado todos los casos de gerundios para unirlos en una sola búsqueda (véase tabla 1). Así, contamos con una búsqueda completa y exhaustiva de todos los gerundios con diminutivo documentados.

Como ya se apuntó anteriormente, los nombres propios (*Pablito*, *Juanita*, etc.) están lematizados con sufijo diminutivo. Es decir, para el CORPES XXI, *Juanito* y *Juan* son dos lemas (dos nombres) distintos. Optamos por incluir los nombres propios en nuestro estudio, ya que consideramos que en ocasiones son alternativas a los nombres sin diminutivo (*Pablo* vs. *Pablito*), aunque también es posible que una determinada persona casi siempre se llame *Pablito* y nunca *Pablo*. A pesar de que algunos nombres propios con diminutivo podrían considerarse lexicalizados, nos interesa incluirlos para ver su distribución diatópica. Para contar los nombres propios con sufijo diminutivo, hay que excluir las formas no pertinentes que aparecen en el “inventario de formas”, es decir, aquellos lemas de nombres propios que también terminan en *-ito/s* e *-ita/s* sin ser diminutivos (por ejemplo, *Quito*, *Margarita*, *Benito*, etc.). Sin embargo, la lista es demasiado larga y solo lo pudimos hacer para las formas más frecuentes. En este caso, aceptamos, pues, que los resultados no sean del todo limpios con el fin de que sea una búsqueda todavía manejable.

A pesar de algunas incongruencias discutidas en la sección anterior (por ejemplo, la exclusión de *viejita/s* a la par que se incluye *viejito/s*) y a pesar de que algunas correcciones son imposibles de hacer manualmente (como el repaso manual del inventario de todos los nombres propios), creemos que las búsquedas presentadas en la tabla 1 son la manera más limpia y —hasta el momento— la única factible desde un punto de vista metodológico para filtrar los diminutivos sin tener que pasar manualmente por la lista de todas las 14 000 formas (*types*). Sirva este ejemplo para llamar la atención sobre limitaciones y futuros retos de la anotación automática del CORPES XXI. Por el momento, hay que aceptar que los datos recuperados mediante nuestra búsqueda no son del todo coherentes ni al 100 % exhaustivos. De hecho, según Moreno Sandoval (2022: 416), los actuales corpus anotados automáticamente, como el CORPES XXI, tienen una tasa de precisión de alrededor al 90 %.

5. Resultados de la búsqueda por clases de palabras

En esta sección, analizaremos los resultados según clases de palabras. La tabla 2 presenta el número de *tokens*, *types*_{FORMA} y *types*_{LEMA} para cada clase de palabra.

Tabla 2. Número de resultados (*tokens*, *types*, *lemas*) por categoría gramatical

Categoría gramatical	<i>tokens</i> (casos)	<i>types</i> _{FORMA}	<i>types</i> _{LEMA}
Sustantivos comunes	143 513	7 019	4 501
Nombres propios	90 686	8 296	8 260
Adjetivos	22 993	2 089	1 191
Participios (adjetivos)	1 592	392	271
Adverbios	5 856	50	40
Cuantificadores	2 332	35	10
Interjecciones	238	2	2
Numerales	38	5	2
Gerundios	28	18	18
Artículos	13	3	1
Demostrativos	13	5	2
Posesivos	2	2	2
Relativo	2	1	1
Suma	267 306	17 917	14 301

En total, nuestro estudio se basa, pues, en unos 267 000 casos u ocurrencias (*tokens*). Estas pertenecen a aproximadamente 14 000 *types*_{LEMA} distintos. Este último número cuenta por separado palabras polifuncionales: por ejemplo, *poquito* —el diminutivo más frecuente de todos que hemos registrado, con 7 505 *tokens*— aparece en las categorías “sustantivo”, “cuantificador” y “adverbio”. Corresponde, pues, a tres *types*_{LEMA} distintos en el total. Ya que esta polifuncionalidad afecta sobre todo a sustantivos y adjetivos (por ejemplo, *viejito* como sustantivo y adjetivo), probablemente habría que corregir hacia abajo el total de *types*_{LEMA}. Si asumimos radicalmente que, en el ‘peor’ de los casos, todos los lemas de adjetivos son a la vez también lemas de sustantivos, habría que restar unos 1 000 lemas y el número total quedaría en unos 13 000. Además,

el número de *types* en los nombres propios está distorsionado por el hecho de que nombres completos se cuentan por separado (por ejemplo, el CORPES XXI cuenta por separado los lemas *Juanito* y *Juanito Pérez*). Si ignoramos los nombres propios, las demás clases de palabras suman unos 6 000 *types*_{LEMA} distintos, o probablemente 5 000 si eliminamos posibles casos de polifuncionalidad.

Respecto a la frecuencia *token*, la categoría gramatical que más frecuentemente acepta sufijos diminutivos son los sustantivos comunes, seguidos por los nombres propios y los adjetivos. Les siguen, con más distancia, los adverbios y los cuantificadores. Las demás clases de palabras muy raramente toman sufijos diminutivos. En términos de frecuencia *type* (es decir, de lemas distintos que toman sufijos diminutivos), hay una mayor variación en los nombres propios que en los sustantivos comunes. Esto quiere decir que hay más nombres propios distintos que aparecen con sufijos diminutivos que nombres comunes, pero que estos nombres propios se usan con menos frecuencia (hay una frecuencia *token* más baja que en los sustantivos comunes). Aparte de la distorsión ya mencionada respecto a la lematización de nombres completos, interpretamos esta divergencia como indicio de una mayor productividad de sufijos diminutivos con nombres propios que con sustantivos comunes: es más fácil añadir un sufijo diminutivo a un nombre propio cualquiera que a un sustantivo común cualquiera. De hecho, en teoría, todos los nombres propios de personas pueden llevar un sufijo diminutivo. En cambio, no todos los sustantivos comunes pueden aceptar sufijos diminutivos: por ejemplo, los que denotan características, cualidades y estados físicos o anímicos (*equilibrio*, *altura*, ...) los rechazan (Real Academia Española 2009: 634).

5.1. Nombres propios

En total, hay 90 686 casos (*tokens*) registrados de nombres propios con sufijo diminutivo (excluyendo las formas precisadas arriba, véase la tabla 1), que corresponden a 8 296 formas (*types*_{FORMA}) distintas, las cuales a su vez se reducen a 8 260 lemas (*type*_{LEMA}). En otras palabras, hay unos ocho mil nombres distintos documentados. En el caso de los nombres propios no esperábamos ninguna diferencia entre *types*_{FORMA} y *types*_{LEMA}: las formas *Juanita*, *Juanito*, *Anita*, etc. están lematizadas como *Juanita*, *Juanito*, *Anita*, etc. Sin embargo, en unos treinta casos, dos o más formas se fusionan en un lema. Hojeando el inventario “formas-lema” no hemos descubierto todavía cuál podría ser esta discrepancia (muy mínima) entre los dos números.² Los resultados de los nombres propios son algo preliminares y no del todo representativos: habría que repasar todo el inventario de más de 8 000 lemas para excluir realmente todos los nombres que no son diminutivos (recordamos que sólo llegamos a excluir los más frecuentes del inventario). Además, la alta frecuencia de ciertos nombres propios, por ejemplo, de protagonistas de novelas o piezas de teatro, puede distorsionar el panorama global.

Las 12 formas más frecuentes según el “inventario” son *Anita*, *Carlitos*, *Juanito*, *Conchita*, *Chiquita*, *Rosita*, *Miguelito*, *Evita*, *Pedrito*, *Juanita*, *Pepita* y *Paquita*. Llama la atención que los sufijos diminutivos parecen usarse más frecuentemente con nombres femeninos (de los doce, ocho nombres son femeninos). Esta tendencia se confirma cuando se buscan por separado: hay 43 100 casos de nombres propios terminados en *-ito* o *-itos*, frente a 47 586 casos de nombres propios terminados en *-ita* o *-itas*. Esta diferencia se agrava aún más teniendo en cuenta que, entre los primeros, se incluyen nombres como *Charito*, *Amparito* y *Rosarito*, que en realidad son nombres femeninos; al contrario, no parece haber nombres propios masculinos que terminen en *-ita(s)*. El ámbito de los nombres propios confirma, pues, que el diminutivo se asocia más a estereotipos femeninos que masculinos (por ejemplo, varios estudios en PRESEEA muestran constantemente un mayor uso de diminutivos por parte de las mujeres entrevistadas, cf. Malaver & Paredes García 2020; León-Castro Gómez 2020; Malaver 2021).

En el siguiente diálogo se observa un cambio en la forma de tratamiento: del nombre no disminuido, aunque con un diminutivo en el tratamiento (*señorita Juana*), la interlocutora pasa a la forma con diminutivo (*Juanita*), que expresa probablemente un grado más elevado de cercanía en el momento de emitir una advertencia:

- (1) – Buenos días, *señorita Juana*.
 – Buenos días, señora Pamela.
 – A ver cuándo nos pasea el perro, mi marido está enfermo.
 – Cuando deseé, señora.
 – Cuidado, *Juanita*, puedo tomarle la palabra.
 – Adiós.
 (CORPES XXI: Aridjis, Homero: *La zona del silencio*. México, 2005)

En este ejemplo, el valor del diminutivo es, sobre todo, relacional: expresa familiaridad con la interlocutora en una intervención comprometida.

5.2. Sustantivos comunes

La búsqueda por sustantivos comunes con sufijo diminutivo recupera 143 513 casos (*tokens*) de 7 019 formas flexionadas (*types*_{FORMA}) distintas, que a su vez se agrupan en 4 501 lemas (*type*_{LEMA}). Por ejemplo, las cuatro formas flexionadas *abuelito*, *abuelita*, *abuelito* y *abuelitos* (cuatro *types*_{FORMA}) pertenecen a un solo lema (el *type*_{LEMA} *abuelo*).

² Según la sugerencia de un/a revisor/a anónimo/a, podría tratarse de formas plurales (por ejemplo, en *todos los Juanitos*) que se deberían lematizar como *Juanito*. Sin embargo, no es el caso: los nombres propios en forma plural parecen haberse lematizado como tales (por ejemplo, *Juanitas* como *Juanitas*). También variantes como *Juanito* y *Juancito* se han lematizado con dos formas distintas. Probablemente, el desajuste surge de alguna confusión en variantes con mayúsculas o minúsculas, o en el recuento automático de nombres propios que aparecen lematizados junto al apellido (por ejemplo, *Juanito Valderrama*).

Las formas (*types*_{FORMA}) más frecuentes son *poquito*, *abuelita*, *cajita*, *ratito*, *jovencita*, *hermanito*, *hermanita*, *salita*, *trocitos*, *papelito*, *carita*, *pedacitos*, *pueblito*, *chorrito* y *bolsita*. Llama la atención el primero de la lista, *poquito*: está anotado como sustantivo cuando forma parte de la secuencia *un poquito*, que sería una locución pronominal o adverbial (Real Academia Española 2014: s.v. *poco*). La alta frecuencia de esta forma etiquetada erróneamente distorsiona, pues, los resultados de esta categoría.

Para ejemplificar el uso de sustantivos con diminutivo citemos el ejemplo (2):

- (2) ¿Elizabeth, de parte de quién?... ¡Ah!, Sr. Fernández. Qué bueno que llama. Permítame un *momentito*. (CORPES XXI: Durón, Hispano: Anita, la cazadora de insectos. Guión cinematográfico. 2002, Honduras)

En este ejemplo, *momentito* es una disminución subjetiva de *momento*: expresa una intensificación del componente semántico de 'poco tiempo' (*momento*) a 'muy poco tiempo' (*momentito*). Este ejemplo reúne, pues, una valoración cuantificadora-centralizadora con una valoración relacional: es una estrategia de atenuación de la petición y, por lo tanto, un diminutivo de cortesía orientado hacia el interlocutor.

De origen sustantival son también los dos tipos de interjecciones documentados con sufijo diminutivo: *cuidadito* (185 tokens) y *m'hijita* (52 tokens).

5.3. Adjetivos y participios

Hay 22 993 adjetivos (*tokens*) con sufijo diminutivo que corresponden a 2 089 formas distintas (*types*_{FORMA}). Estas variantes flexionadas, a su vez, se reducen a 1 191 lemas (*type*_{LEMA}) distintos. Los *types*_{LEMA} más frecuentes son *pequeñito*, *bajito*, *solito*, *pobrecito*, *chiquito*, *igualito*, *jovencito*, *suavecito*, *clarito* y *fresquito* (incluidas variantes flexionadas).

A modo de ejemplo, el adjetivo *jovencito* en (3) expresa tanto una valoración cuantificadora centralizadora (intensificación, expresada también por *muy* + adjetivo) como una valoración cualificadora positiva (emocional-afectiva y, más en concreto, compasión):

- (3) Alguno de los otros perros, Facundo, por ejemplo, olisqueaba el poste de la luz, la de la esquina o el árbol de níspero a la orilla del río donde Canelo había depositado su húmedo almizcle. Facundo, luego de semejante ceremonia, „meaba como hembra“, decía Juana; „como un *chiquitín*“, afirmaba Emperatriz, „es muy *jovencito* todavía“ -sostenían las dos- pero eran disculpas de Juana y Emperatriz, encariñadas con los bichos cuidanderos. (CORPES XXI: Martínez Espinosa, Jorge: «Canelito, Canelito, pobrecito Canelito», *El final de los milagros*. Colombia, 2001)

La secuencia contiene también la forma *chiquitín* (el sufijo *-ín* no cuenta en este estudio), que igualmente expresa tanto una intensificación del tamaño pequeño como una valoración afectiva-emocional. Además, el título mismo del relato muestra una acumulación de diminutivos: en el nombre propio *Canelito* y en otro adjetivo que expresa compasión (*pobrecito*).

A los adjetivos se suman los participios, que están anotados como verbos (concretamente bajo la subcategoría “tiempo”). En el corpus hay 1 592 casos (*tokens*) de 392 formas flexionadas (*type*_{FORMA}) de participios con sufijo diminutivo, que corresponden a 271 lemas distintos (*type*_{LEMA}). Los *types* más frecuentes son *calladito*, *picadito*, *pegadito*, *sentadito*, *dormidito*, *heladito*, *doradito*, *cortadito*, *entradito* y *pasadito* (incluidas variantes flexionadas). Comprobamos que, en todos los casos, estos participios tienen función adjetival; nunca forman parte de una forma verbal compuesta. Mediante la función “+ distancia” pusimos como filtro el lema *haber* a una palabra de distancia a la izquierda. De este filtro resultarían casos como **has calladito*. Esta combinación no da ningún resultado. Por lo tanto, en los resultados finales sumamos manualmente los participios a los adjetivos.

Por ejemplo, en (4), *calladito* tiene función adjetival:

- (4) ¿Te cae de madre?, se sorprende Rubén. Por Dios que sí. Mira nomás, qué *calladito* te lo tenías, ¿y cuántos cumpleaños? 38. (CORPES XXI: Serna, Enrique: «El matadito». *El orgasmógrafo*, México 2001)

En este ejemplo, la función del diminutivo es probablemente una valoración relacional hacia el interlocutor: o bien expresa ironía o bien atenúa el reproche.

5.4. Adverbios

Los 5 856 casos (*tokens*) de adverbios con sufijo diminutivo corresponden a 50 formas distintas (*types*_{FORMA}) y a 40 lemas distintos (*type*_{LEMA}). El descenso cuantitativo de formas a lemas se explica por variantes como *lejito* ~ *lejitos* ~ *lejecitos* (tres formas, pero un lema), *luegucito* ~ *lueguito* (dos formas, pero un lema) y por la acumulación de sufijos: *poquitito* ~ *poquito* (dos formas, pero un lema).

El adverbio más frecuente es *ahorita* (con 2 211 casos), seguido con mucha diferencia por *despacito* (972 casos). Los siguientes en frecuencia son *bajito*, *cerquita*, *clarito*, *tantito*, *tempranito*, *poquito*, *nadita* y *lueguito*. En el caso de *bajito* y *clarito* se trata de formas que también se documentan entre los adjetivos. De nuevo, es el contexto el que causa la anotación automática como adverbios (en este caso, correctamente): cuando son precedidos o seguidos por un verbo, se etiquetan como adverbios. Por ejemplo, *bajito* se documenta con *hablar* o *decir*, y *clarito* con *decir*, *ver* o *escuchar*. En (5), el adverbio *bajito* modifica el verbo *saludar* y creemos que su valor es meramente cuantificador (centralizador/intensificador):

- (5) Jamás se lo ve soltar una lágrima o una carcajada. Muy de vez en cuando es posible que deje escapar una sonrisa en la intimidad de su casa; en la calle o en el trabajo, jamás. De ahí que parezca una estatua de expresiones rígidas en movimiento.
Saluda cortésmente y muy *bajito*, aunque siempre se lo escucha.
(CORPES XXI: Oviedo, Jorge Luis: «EL SICARIO». *El Cazabrujas y otros personajes*. Honduras, 2001)

La anotación automática no sólo considera el contexto inmediato, sino también el más amplio (el verbo *saludar* está a cuatro palabras de distancia de *bajito*). No obstante, no diferencia siempre correctamente entre adjetivos y adverbios cortos (adjetivos adverbiales): hay casos de adjetivos etiquetados erróneamente como adverbio y al revés. Al parecer, la etiquetación de estas formas como adverbios sólo se ha hecho para algunos *types* (*bajito*, *clarito*, *derechito*, *rapidito*, *segurito*), pero no para otros como *justito*: tanto en ejemplos como adverbio de modo como en ejemplos de adverbio de foco, *justito* se ha anotado como adjetivo. En todo caso, a pesar de que la anotación como adjetivo o adverbio de formas como *bajito*, *clarito*, *justito*, etc. vacile y no sea coherente, tanto las formas etiquetadas como adjetivo como las formas etiquetadas como adverbio se suman en nuestros resultados finales y por lo menos no se pierden en el resultado global.

Son muy escasos los adverbios en *-mente* con sufijo diminutivo (Real Academia Española 2009: 573): en el CORPES XXI se documentan solamente tres formas: *viejecitamente* (1 caso), *jovencitamente* (1 caso), *chiquitamente* (2 casos). Estas se recuperan a través de la búsqueda por la forma **itamente*.

5.5. Cuantificadores, artículos, numerales, demostrativos, posesivos y relativos

Hay un total de 2 332 casos (*tokens*) de cuantificadores con diminutivo, que corresponden a 35 formas flexionadas (*types*_{FORMA}) distintas, que a su vez se reducen a 10 lemas (*types*_{LEMA}). Los más frecuentes son *poquito*, *mismito*, *todito*, *tantito*, *muchito*, *alguito*, *bastantito* y *nadita* (incluyendo las variantes flexionadas). Como ya mencionamos anteriormente, *poquito* es la forma diminutiva más frecuente de todo el corpus (7 505 casos, de los cuales la gran mayoría está etiquetada como sustantivo).

En el caso de los cuantificadores, el sufijo diminutivo suele enfatizar e intensificar. Tiene, pues, un valor cuantificador-centralizador. Por ejemplo, en el caso de *toditos*, la interpretación sería algo como 'realmente todos y cada uno'. Es más, se puede intensificar aún más con la acumulación de múltiples sufijos diminutivos, aunque en este caso el efecto es probablemente de burla hacia el interlocutor (por exageración):

- (6) – «Pa'qué», compadre – me respondió azorado.
– Pues para el fiestón de la cosecha.
– ¡Ah! «pus» ya «nomás» como doce días.
– A ver, a ver, pero cuéntame, ¿van todos?, ¿todos los de casa?
– *Todititos*.
– Oye, y el patrón qué tal toma.
– No, hasta eso; él casi no toma ese día. Sí se «echa» sus buenos tragos, pero «aluego» se va a su cuarto.
(CORPES XXI: Hernández Rodríguez, Rafael: *La muerte de un cardenal*. México, 2001)

El uso de artículos con sufijo diminutivo apenas se detecta en el CORPES XXI. Hay, en total, 13 casos (*tokens*) que corresponden a 3 formas flexionadas (*types*_{FORMA}: *unita*, *unitos*, *unitas*), que a su vez forman parte del mismo lema: *uno*. En casos como (7), el sufijo diminutivo tiene un valor cuantificador-centralizador: *ni unita* intensifica la valoración de *ni una* ('ni siquiera una sola esquina'):

- (7) No había rincón allí para ella, ni *unita* esquina dónde esconderse, donde no la pudieran espiar.
(CORPES XXI: García Huevo, Oscar: «Mientras, en España». *Lo que pasó en San Lorenzo y otros cuentos*. El Salvador, 2001)

De manera similar, entre los numerales con sufijo diminutivo se encuentran variantes de *unito* y *primerito*. Estos dos lemas, en 5 formas flexionadas (*types*_{FORMA}) distintos, son representados con 38 casos (*tokens*) en total. En el caso de *primero*, el sufijo *-ito* también intensifica subjetivamente, expresando énfasis ('realmente la primera de todos'), y además hay también afecto emocional hacia la persona referida:

- (8) Al escucharla, trataba de ver a mi abuela dentro de otro destino. Y pensé en lo que Mima decía de su madre cuando se sentía benévola: Esa Virginia se las trae. ¿Tú te imaginas lo que hubiera logrado con un poco de estudio? Niño, de haber llegado joven a este país hoy sería millonaria. Y de haber nacido aquí, hoy sería presidenta de Estados Unidos. ¡La *primerita*!
(CORPES XXI: Miguel Muñoz, Elías: *Vida Mía*. Cuba, 2006)

A su vez, los 13 tokens de demostrativos pertenecen a 5 formas flexionadas (*type*_{FORMA}) de 2 lemas (*type*_{LEMA}): *aquello* y *esto*. En el siguiente ejemplo, *aquellita* tiene probablemente un matiz despectivo (valoración cualificadora negativa):

- (9) – Entonces tú sí sabías dónde encontrarme. En cualquier ejemplar del periódico viene el teléfono.
– Pero no te iba a llamar nada más porque sí, mi querube. Después que te juntaste con *aquellita* preferí no intervenir. Pero como me imagino, si me buscaste no fue porque anden muy bien tus cosas con ella. Y como ya sé que en eso no me voy a meter, mejor ven, para que conozcas mi recámara, está arriba.
(CORPES XXI: Güemes, César: *Soñar una bestia*. México, 2011.)

De los posesivos se documentan solo dos casos de dos lemas (*suyita* y *tuyita*). Para los relativos hay dos casos de *cuantito*, pero ambos corresponden a la locución *en cuantito*, que al igual que muchas otras locuciones no se ha categorizado correctamente.

5.6. Gerundios

Hay 28 casos de gerundios con diminutivo (*tokens*), que pertenecen a 18 formas (*types*), etiquetados como “desconocido”. En este caso, el número de formas es igual al número de lemas. Solo cinco de estos *types* (*corriendito*, *llegandito*, *andandito*, *pasandito*, *saliendito*) ocurren más de una vez. A estos habría que añadir los 8 *tokens* de *callandito*, anotados erróneamente como adverbios (que sería su función sintáctica, pero no la clase de palabra). Los restantes 13 tipos (*cantandito*, *deseandito*, *jugandito*, etc.) tienen una sola ocurrencia por *type*.

En el ejemplo (10), el diminutivo *corriendito*, por un lado, puede intensificar la acción y, consecuentemente, el acto de la orden urgente. Por otro, expresa un matiz afectivo (o bien de cariño o bien de reproche condescendiente).

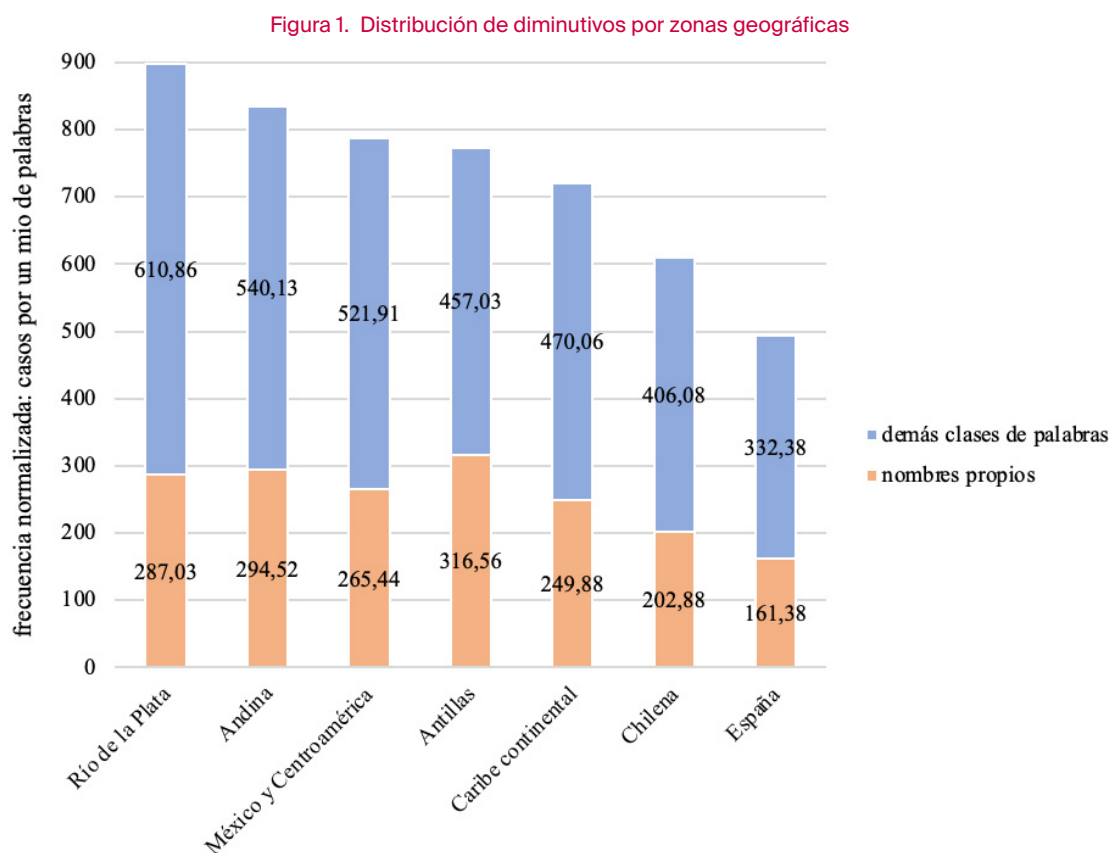
- (10) — ¿Abriste la puerta del corral de las gallinas? Acabo de sacar una clueca de la casa...
 — No, doña Luz, yo no, se lo prometo — pero en la cara de Toña había más una confesión que una disculpa, suerte que no se la vio doña Luz, que le hubiera puesto ahí mismo una regañiza.
 — Ve a decirle al joven que lo busca una señorita que es mecanógrafa. Y *corriendito* te me vas a revisar si está cerrada la puerta del corral, janda!
 (CORPES XXI: Boullosa, Carmen: *Las paredes hablan*. México, 2010).

6. Distribución geográfica de los resultados

En esta sección analizaremos la frecuencia normalizada de los diminutivos por zona geográfica. Primero discutiremos los resultados generales y la agrupación de países en zonas (6.1), después los resultados según tipología textual y temáticas (6.2), según las clases de palabras por zona (6.3) y finalmente la diferencia entre los subcorpus oral y escrito (6.4).

6.1. Frecuencia global de diminutivos y agrupación de países en zonas

La figura 1 resume los resultados globales de la frecuencia normalizada de diminutivos por zonas. Para simplificar la visualización, diferenciamos solamente entre los nombres propios y las demás clases de palabras, ya que la división por clases de palabras se analizará más adelante.



De acuerdo con las hipótesis planteadas, esperábamos que España fuera el país con la frecuencia más baja y probablemente la zona andina o mexicano-centroamericana la de la frecuencia más elevada. Sin embargo, los datos solo cumplen parcialmente las expectativas: la zona de mayor uso de diminutivos en el CORPES XXI

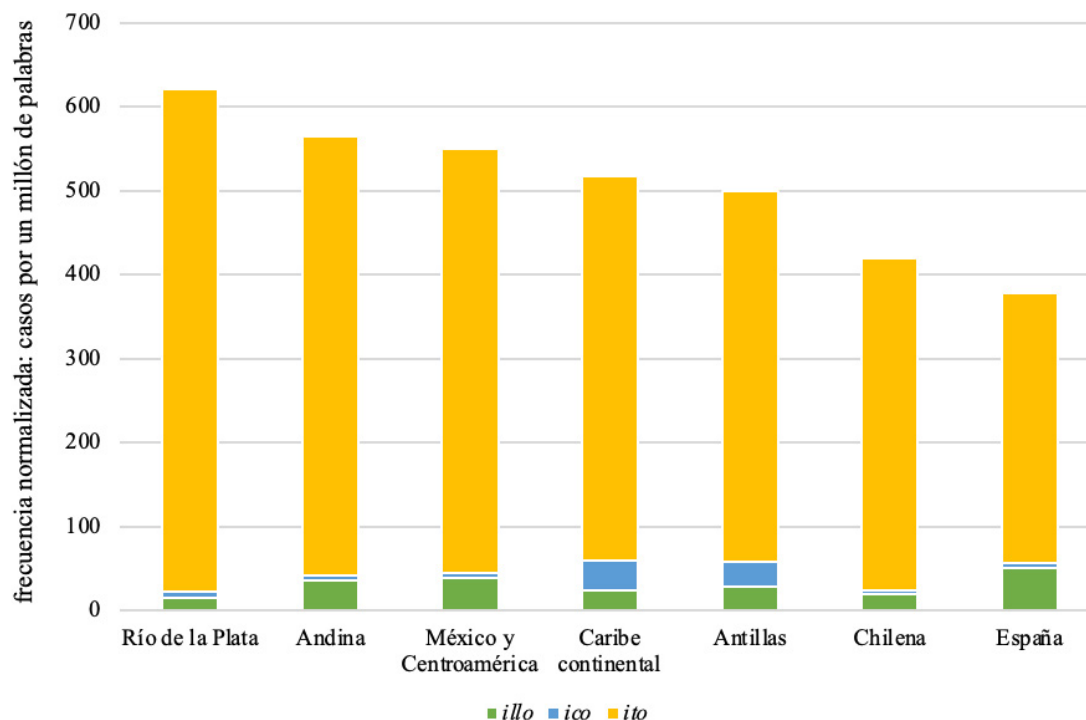
es el Río de la Plata, seguido por la zona Andina y México y Centroamérica. En último lugar y con una diferencia notable queda España, con casi sólo la mitad de casos documentados que en el Río de la Plata. Los datos confirman, pues, claramente el panorama general de un mayor uso de sufijos diminutivos en *-ito/-ita* en América que en España. También se confirma la alta frecuencia de uso en regiones como México y los Andes. Sin embargo, sorprende que el primer puesto sea el del Río de la Plata, una tendencia que contradice los resultados de Reynoso Noverón (2001). Sin embargo, como veremos más adelante (6.4.3), hay que matizar este resultado.

Para comparar los datos de *-ito/a* con los de otros sufijos relativamente frecuentes como *-ico/a* e *-illo/a* lanzamos también una breve búsqueda por estos dos sufijos según el modelo presentado en 4.3, es decir, excluyendo los lemas con tales terminaciones para excluir los casos lexicalizados. Nos limitamos a la búsqueda por nombres comunes, adjetivos y adverbios, ya que estas tres categorías se pueden agrupar con facilidad en una sola búsqueda y cubren la gran mayoría de formas diminutivas (ver apartado 6.3). La tabla 3 y la figura 2 resumen los resultados de la búsqueda paralela de estos tres sufijos diminutivos.

Tabla 3. Frecuencia relativa (por un millón de palabras) de los diminutivos en *-ito/a*, *-illo/a* e *-ico/a* en sustantivos comunes, adjetivos y adverbios en el CORPES XXI

Zona	<i>-ito/a</i>	<i>-illo/a</i>	<i>-ico/a</i>	suma
Río de la Plata	598.03	15.54	7.42	620.99
Andina	522.93	36.21	5.26	564.40
México y Centroamérica	505.47	39.17	5.46	550.10
Caribe continental	457.04	23.97	36.26	517.27
Antillas	441.83	28.93	29.29	500.05
Chilena	394.87	20.05	4.65	419.57
España	321.16	50.87	5.63	377.66

Figura 2. Distribución diatópica de los diminutivos en *-ito/a*, *-illo/a* e *-ico/a* en sustantivos comunes, adjetivos y adverbios



Añadiendo *-ico/a* e *-illo/a* a nuestros datos, el panorama general presentado en la Figura 1 no cambia sustancialmente y el orden apenas se altera, con la única excepción de que el Caribe continental y las Antillas —que son las regiones donde más se usa *-ico/a*— intercambian sus puestos. En todas las zonas, *-ito* es, con diferencia, el sufijo diminutivo más usado. España, a pesar de ser la zona donde más se documenta *-illo/a*, sigue estando claramente en el último puesto.³ Visto el uso muy reducido de *-illo* e *-ico* frente a *-ito*, creemos que es justificable limitarnos solamente a *-ito/-ita* en este trabajo.

³ En España, el sufijo *-illo* sigue siendo productivo en las zonas centro-sur, así como lo son *-ico* en la zona oriental (Navarra, Aragón, Murcia, Granada), *-iño* en Galicia e *-ino* en la zona occidental (Asturias, León, Extremadura). Esta distribución se comprueba, por

Para comprobar si los países que conforman una zona lingüística se comportan de manera similar en cuanto a la frecuencia de diminutivos en *-ito/a*, recurrimos a los datos por países, representados en la tabla 4 y en la figura 3.

Tabla 4. Frecuencia normalizada (por un millón de palabras) de diminutivos en *-ito/a* según países

País (Zona)	Nombres propios	Demás categorías	Suma
Perú (Andina)	352.32	786.29	1138.61
Argentina (Río de la Plata)	294.84	679.81	974.65
Honduras (Centroamérica)	406.70	558.07	964.77
Nicaragua (Centroamérica)	298.96	536.36	835.32
Puerto Rico (Antillas)	317.94	505.65	823.59
Cuba (Antillas)	315.25	479.54	794.79
Guatemala (Centroamérica)	220.38	566.67	787.05
México	245.55	541.02	786.57
Panamá (Centroamérica)	350.63	409.05	759.68
El Salvador (Centroamérica)	255.16	499.05	754.21
Venezuela (Caribe Continental)	260.08	483.40	743.48
Paraguay (Río de la Plata)	316.37	426.17	742.54
Uruguay (Río de la Plata)	231.45	502.84	734.29
Colombia (Caribe Continental)	244.40	462.92	707.32
República Dominicana (Antillas)	318.01	386.20	704.21
Costa Rica (Centroamérica)	273.16	345.91	619.07
Chile	202.88	406.08	608.96
Ecuador (Andina)	285.74	295.40	581.14
Bolivia (Andina)	191.64	353.08	544.72
España	161.38	332.38	493.76

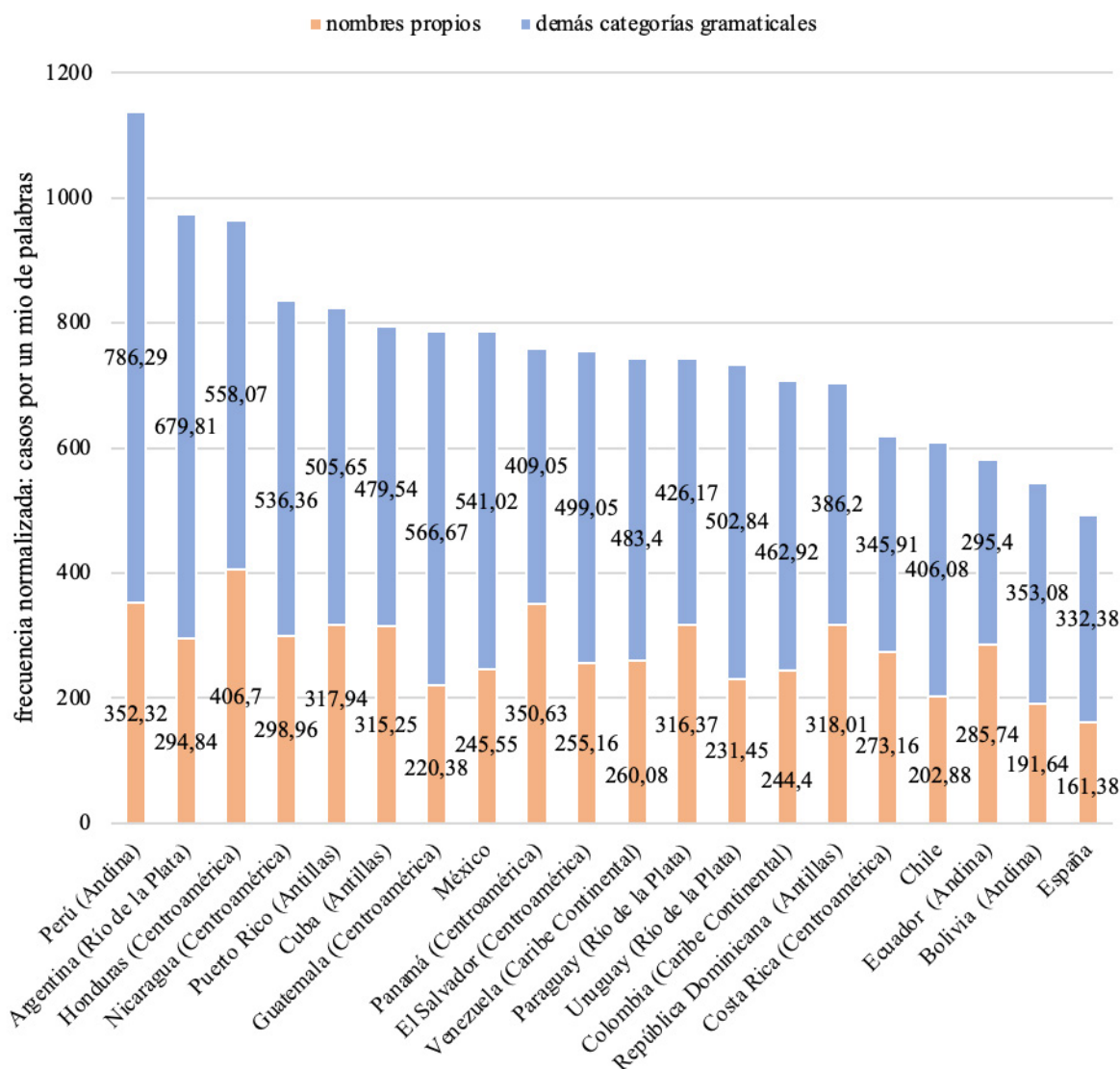
Para algunas de las zonas, la agrupación de países parece justificada: las Antillas, el Caribe continental, y México con los países centroamericanos muestran una imagen bastante homogénea entre sí. En los resultados por países, España se mantiene claramente en el último puesto, con menos de la mitad de casos de diminutivos que Perú. Es decir: en el CORPES XXI, se documentan más del doble de diminutivos en Perú que en España.

No obstante, los datos separados por países también parecen mostrar que la agrupación en zonas reúne en ocasiones países con frecuencias no coincidentes en el caso de los sufijos diminutivos: el primer país —con diferencia— es Perú, un país andino. Sin embargo, los otros dos países andinos, Ecuador y Bolivia, muestran las frecuencias más bajas dentro de América. También el Río de la Plata se dispersa, aunque no tanto como la zona andina: Argentina está en segundo lugar, mientras que Uruguay y Paraguay ocupan puestos intermedios. Tanto los resultados globales del Río de la Plata como de la zona andina se deben, pues, a dos países con una muy elevada frecuencia de diminutivos, que destaca frente a los demás países de su zona.

¿Quiere esto decir que la agrupación en zonas lingüísticas —concretamente, en la zona lingüística andina y, en menor medida, en la zona rioplatense— es un fallo metodológico? No lo creemos. La explicación más bien se encuentra en el tipo de textos que representan cada uno de los países, como mostraremos a continuación.

ejemplo, en mapas generados en el *Corpus Oral y Sonoro del Español Rural* (= COSER), que permite búsquedas esquemáticas parecidas a las efectuadas en este trabajo (buscando por **ito*, **illo*, etc.), con la ventaja de que los sustantivos y adjetivos con sufijo diminutivo se han etiquetados con una etiqueta especial de “diminutivo”. Además, según los datos de estudios en el corpus PRESEEA, *-ito* es el sufijo diminutivo más frecuente, seguido de *-illo*, tanto en Madrid (Paredes García 2015: 127) como en Sevilla (León-Castro Gómez 2020: 117), mientras que en Granada *-illo* predomina sobre *-ito* e *-ico* (Malaver 2021: 384). Para un estudio comparativo entre *-ito*, *-illo* e *-ico* en diversos corpus históricos y dialectales (con especial atención a Andalucía), véanse también los datos presentados por Calderón Campos (2024).

Figura 3. Visualización de la frecuencia normalizada de diminutivos por países



6.2. Frecuencia de diminutivos según temáticas y tipología (variación diafásica)

En esta sección analizamos la composición del CORPES XXI en cuanto a temáticas y tipologías textuales para ver si estos factores influyen en la documentación de diminutivos.

Veamos primero en qué tipo de textos predominan los diminutivos. Para comprobarlo, buscamos conjuntamente las categorías gramaticales más frecuentes: sustantivos comunes, adjetivos, adverbios (se pueden agrupar hasta tres categorías en una búsqueda y la búsqueda en “lema” y “forma” son iguales para estas tres categorías). Los resultados de estos diminutivos muestran que las temáticas en las que se documenta el mayor uso de diminutivos son el teatro (f.norm. 1 576.41), el relato (1 165.50), el guion (1 016.94) y la novela (935.03); en cambio, las temáticas en las que menos se documentan son los temas de salud (143.63), las ciencias y tecnologías (119.57), y política, economía y justicia (119.57). La distribución por tipologías apunta en la misma dirección y muestra discrepancias aún más pronunciadas: el mayor uso de diminutivos se documenta en magazines y variedades (2 000.12), sorteos y concursos (1 783.8), retransmisiones deportivas (1 155.98), ficción (1 022.49), mensajes en redes sociales (1 022.02) y tertulias (1 009.41). En cambio, la frecuencia más baja de diminutivos se documenta en noticias (130.73), editoriales (120.66), textos académicos (76.71) y jurídicos-administrativos (2.18). Se confirma, pues, que la vertiente diafásica es crucial: el uso de diminutivos depende significativamente de la configuración situacional: es más frecuente en géneros orales, en situaciones comunicativas más informales y espontáneas, así como en temas más cotidianos.

El perfil de los documentos recopilados para Argentina, Paraguay, Uruguay (que juntos forman la región rioplatense), Perú, Ecuador y Bolivia (que juntos forman la zona andina) y España se muestra en la tabla 5.

Tabla 5. Composición temática y tipológica del CORPES XXI (se mencionan solo los tres bloques temáticos más importantes)

	País	Distribución temática	Tipología textual	F. norm de dimin. (total)
Esp.	España 102 mil documentos 156 mio de palabras	23 % novela 17 % política, economía y justicia 12 % ciencias sociales, creencias, pensamiento	40 % ficción 16 % divulgación 11 % noticias	493.76
Río de la Plata	Argentina 24 mil documentos 40 mio de palabras	23 % novela 15 % política, economía y justicia 12 % ciencias sociales, creencias, pensamiento	42 % ficción 12 % divulgación 10 % noticias	974.65
	Uruguay 9,5 mil documentos 10 mio de palabras	23 % política, economía y justicia 15 % ciencias sociales, creencias, pensamiento 12 % actualidad, ocio y vida cotidiana	34 % ficción 17 % noticias 14 % académico	742.54
	Paraguay 11 mil documentos 9 mio de palabras	24 % política, economía y justicia 19 % novela 14 % ciencias sociales, creencias, pensamiento	37 % ficción 24 % noticias 13 % académico	742.54
Andina	Perú 13 mil documentos 16 mio de palabras	30 % novela 17 % política, economía y justicia 13 % ciencias sociales, creencias, pensamiento	42 % ficción 16 % noticias 15 % académico	1138.61
	Ecuador 15 mil documentos 10 mio de palabras	22 % política, economía y justicia 16 % actualidad, ocio y vida cotidiana 13 % artes, cultura y espectáculo	32 % noticia 23 % ficción 14 % divulgación	581.14
	Bolivia 11 mil documentos 8 mio de palabras	27 % política, economía y justicia 16 % actualidad, ocio y vida cotidiana 13 % ciencias sociales, creencias, pensamiento	30 % ficción 23 % noticia 19 % académico	544.72

De acuerdo con los datos de la tabla 5, la composición de los subcorpus de España, Argentina y Perú es muy parecida en cuanto a las temáticas y tipologías, con un porcentaje importante de novelas (23-30 %) y de ficción (40-42 %). En cambio, el perfil de los subcorpus de Uruguay, Paraguay y más todavía de Ecuador y Bolivia es distinto: la novela ni siquiera está entre los tres bloques temáticos con más peso en los subcorpus de Uruguay, Ecuador y Bolivia, y sólo representa un 19 % en Paraguay. Asimismo, la ficción ocupa considerablemente menos peso en el subcorpus de estos países (37 % en Paraguay, 34 % en Uruguay, 30 % en Bolivia y solo 23 % en Ecuador).

Ahora bien, si las tipologías textuales no se representan con la misma proporción para cada país y región, esto influye en la frecuencia de diminutivos. Precisamente, el corpus de Ecuador y Bolivia, como también el de Uruguay y Paraguay, tiene una porción más elevada de temáticas y tipologías que desfavorecen el uso de diminutivos (por ejemplo, “política, economía y justicia”). En cambio, los subcorpus de España, Argentina y Perú muestran una porción más elevada de temas y tipologías que favorecen el uso de diminutivos (como la novela y la ficción). Estos tres países son bastante comparables en cuanto a su composición temática, por lo que los resultados de los diminutivos para España, Argentina y Perú (véase la última columna de la tabla 5) se pueden considerar realmente comparables entre sí, pero menos con Bolivia y Ecuador por un lado, y Uruguay y Paraguay por otro. La composición no perfectamente balanceada de los subcorpus, además, explica por qué la agrupación en zonas no se mantiene en la Figura 3 presentado en el apartado anterior.

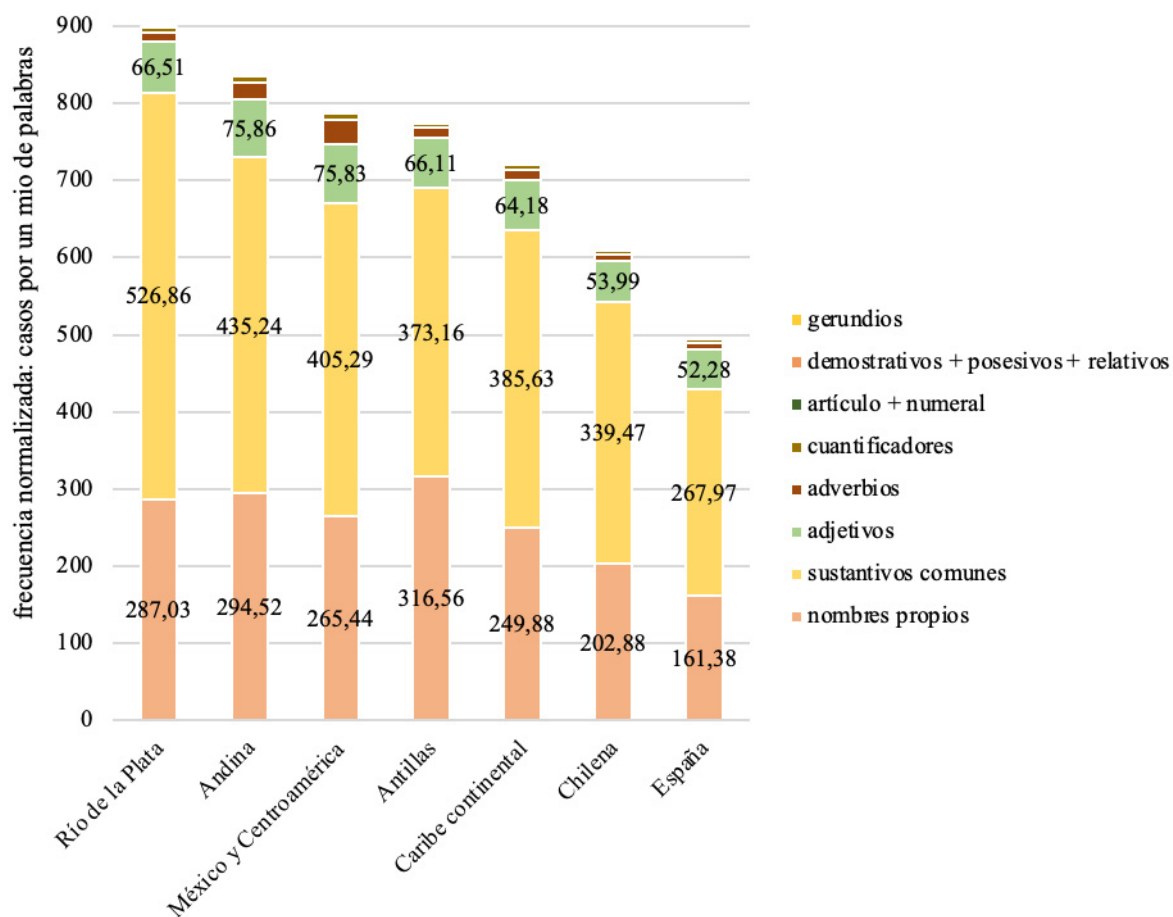
1.3. Frecuencia de diminutivos según clases de palabra por zonas

La tabla 6 y su visualización en la Figura 4 muestran la frecuencia normalizada de diminutivos en cada una de las categorías gramaticales. Para cada categoría, se acentúan mediante colores las tres zonas donde se encuentran las frecuencias más altas.

Tabla 6. Frecuencia normalizada (por un millón de palabras) de las categorías gramaticales con diminutivo según zona

	nombres propios	sustantivos comunes	adjetivos (incl. participios)	adverbios	cuantificadores	artículo + numeral	demonstrativos + posesivos + relativos	gerundios	suma
Río de la Plata	287.03	526.86	66.51	11.46	5.89	0.13	0.01		897.89
Andina	294.52	435.24	75.86	20.33	8.36	0.22	0.12		834.65
México y Centroamérica	265.44	405.29	75.83	32.61	7.67	0.17	0.14	0.20	787.35
Antillas	316.56	373.16	66.11	13.05	4.59	0.12			773.59
Caribe continental	249.88	385.63	64.18	14.11	5.76	0.20		0.18	719.94
Chilena	202.88	339.47	53.99	8.29	4.21	0.12			608.96
España	161.38	267.97	52.28	6.98	5.08	0.05		0.02	493.76

Figura 4. Visualización de frecuencia normalizada de las categorías gramaticales con diminutivo según zonas



Los datos confirman para cada una de las clases de palabras que España es la zona de uso más bajo de diminutivos, con la única excepción de los cuantificadores: en esta categoría, España no ocupa el último puesto sino el antepenúltimo.

Entre las categorías más prototípicas, es decir, las que más frecuentemente llevan sufijos diminutivos, resalta el uso altísimo de nombres propios con diminutivo en las Antillas (316.56). En esta zona, el nombre propio *Chiquita* es con diferencia el más frecuente y distorsiona los resultados: aparece en tan solo 5 obras, pero con mucha frecuencia en una novela cubana en la que es el nombre de la protagonista. Como muestra este ejemplo, ciertos nombres propios de protagonistas pueden distorsionar las frecuencias globales del CORPES XXI. Por lo tanto, las conclusiones sobre los nombres propios son muy provisionales y requieren un filtrado manual más detallado. Volviendo a los datos de los países por separado que acabamos de presentar (sección 6.1), hay, además, dos países centroamericanos que destacan por su alta frecuencia de nombres

propios con diminutivo: son Honduras (406.70) y Panamá (350.63), con *Anita* y *Miguelito* como los *types* más frecuentes.

En el caso de los sustantivos comunes, el valor más alto se encuentra en el Río de la Plata, con una frecuencia normalizada de 526.86. Se trata del valor más elevado de toda la tabla 6. La diferencia respecto a las otras zonas es enorme, de casi 100 casos por un millón de palabras: en segundo lugar, la zona Andina documenta una frecuencia normalizada de 435.24 y México y Centroamérica de 405.29. El valor tan alto de sustantivos comunes con diminutivos en el Río de la Plata eleva considerablemente la suma para esta zona: es por eso por lo que el Río de la Plata queda en primer lugar.

Una posible explicación podría ser una frecuencia más elevada de sustantivos con el sufijo *-ito* lexicalizados en el Río de la Plata, indicios de la idiosincrasia lingüística rioplatense: por ejemplo, *mañanita* 'prenda de vestir', *culito* 'resto' o *besito* 'galleta dulce' en Argentina (Resnik 2019: 240, 242, 244). Si tales sustantivos no están lematizados específicamente para esta área en el CORPES XXI, son 'invisibilizados' y aumentan la frecuencia de sustantivos con diminutivo en el Río de la Plata. Habría que comprobarlo atendiendo al número de *types* y repasando cada uno de los *types* en diccionarios, una tarea que quedará para otro trabajo futuro. Además, esta tarea se dificulta porque el CORPES XXI no nos permite sacar conclusiones limpias sobre el número de *types* por regiones. Aunque es posible contar el número de *types* distintos por región, estos valores no se pueden comparar directamente, ya que los subcorpus de las zonas geográficas no son del mismo tamaño. A diferencia del número relativo-normalizado de *tokens*, la frecuencia de *types* no se puede normalizar sobre el tamaño de los subcorpus, porque el aumento del número de palabras en un corpus no es directamente proporcional al aumento de *types* (Rojo 2017). A modo de ilustración, la tabla 7 muestra el número total de palabras de cada subcorpus (los 'universos de palabras'), el número de *types* distintos de sustantivos comunes con sufijo diminutivo y su relación con el total de palabras. Este valor indica que cada cuántas palabras ocurre un nuevo *type* (para ello, se divide el total de palabras por el número de *types*). Además, se agrega la frecuencia normalizada de *tokens* de la tabla 6.

Tabla 7. Comparación de *types* y *tokens* de sustantivos comunes con sufijo diminutivo *-ito/a*

Zona	Total de palabras del subcorpus	número de <i>types</i> distintos (sust. dim.)	un nuevo <i>type</i> cada ... palabras	F.norm. <i>tokens</i> (sust. dim.)
España	156 161 726	1 924	81 165	267.97
México y Centroamérica	84 563 139	2 353	35 938	405.29
Río de la Plata	59 970 039	1 903	31 513	526.86
Caribe continental	54 732 737	1 757	31 151	385.63
Andina	35 657 254	1 427	24 987	435.24
Antillas	28 448 661	1 273	22 347	373.16
Chilena	27 131 839	1 054	25 741	339.47

Como muestra la tabla 7, a mayor tamaño del corpus, mayor diversidad de *types* (con la excepción de España). Al mismo tiempo, a medida que va aumentando el tamaño de un corpus, se van necesitando cada vez más palabras para documentar un nuevo *type*. Destaca España, que muestra un valor relativamente bajo de *types* (es decir, una productividad más baja) teniendo en cuenta el tamaño del corpus y en comparación con los valores de las zonas americanas. El caso del Río de la Plata no destaca especialmente dentro del panorama americano; hay menos *types* que en México y Centroamérica y más que en el Caribe Continental. Además, en el corpus rioplatense, cada 31 513 palabras se documenta un nuevo *type* de un sustantivo común con diminutivo. Este valor es prácticamente igual al del Caribe continental, que cuenta con un universo de palabras casi equiparable. Es decir, el número de distintos *types* de sustantivos comunes con diminutivo en la variedad rioplatense no destaca especialmente dentro del panorama global latinoamericano y correlaciona con el tamaño del subcorpus. En cambio, sí destaca la alta frecuencia relativa de *tokens* en el subcorpus rioplatense (por ejemplo, en comparación con el Caribe continental).

Por lo tanto, los sustantivos comunes con sufijo diminutivo no parecen ser más productivos en la variedad rioplatense comparado con otras variedades americanas, sino que se documentan con una mayor frecuencia *token*. Esto puede interpretarse, por un lado, como un indicio de un mayor grado de lexicalización de ciertos *types* en el Río de la Plata, suponiendo que la lexicalización causaría una mayor frecuencia *token* del diminutivo porque la forma en cuestión ya no alterna con la forma base sin sufijo. Sin embargo, los indicios cuantitativos son meramente especulativos y habría que comprobarlo en detalle para cada uno de los *types* en cuestión.⁴ Por otro lado, una segunda explicación, sobre la que volveremos con más detalle en

⁴ Según una/a revisor/a anónimo/a, la forma *cachito* (usado como cuantificador en el Río de la Plata) podría empujar hacia arriba el número de sustantivos comunes en el corpus rioplatense. Sin embargo, la búsqueda en el CORPES XXI revela que *cachito* incluso es más frecuente en otros países (Venezuela, Uruguay, Guatemala, entre otros). En cualquier caso, es un ejemplo perfecto de un diminutivo lexicalizado con diversos significados en distintas variedades americanas (véanse las ocho definiciones en el *Diccionario de Americanismos* 2010). Tal y como critica el/la revisor/a, no hay estudios que prueben que en Argentina el número de

el apartado 6.4.3, es la mayor permeabilidad de diminutivos en la lengua escrita argentina, lo cual eleva el número de diminutivos (*tokens*) en el resultado global.

En cuanto a las demás clases de palabras, el panorama latinoamericano cambia. La frecuencia más alta de adjetivos con sufijo diminutivo (véase figura 4) se documenta en la zona andina (75.86) prácticamente a la par con México y Centroamérica (75.83), seguido por el Río de la Plata (66.51). En los datos de los adverbios también resalta su alta frecuencia en México y Centroamérica. Se debe a la elevada frecuencia de una sola forma, *ahorita* (19.81 en México y Centroamérica, seguido con mucha diferencia por 8.56 en la zona Andina).

En el caso de los cuantificadores, asimismo se documenta el valor más alto en la zona andina, seguida por México y Centroamérica y el Río de la Plata. Los cuantificadores son la única clase de palabra en la que España no está en último lugar sino en antepenúltimo. Esto se debe a una sola forma: *poquito*, bastante habitual en España. De hecho, el cuantificador *poquito* muestra su frecuencia de uso (normalizada) más alta en España y aproximadamente la mitad de los casos de cuantificadores en España se deben a la forma *poquito*. En cambio, en las variedades americanas hay una mayor diversidad de *types* distintos (*todito*, *tantito*, *muchito*, *alguito*, *bastantito* y *nadita*). Estos seis cuantificadores juntos predominan en las variedades americanas. La frecuencia más alta de estas seis formas se da en México y Centroamérica y, en segundo lugar, en la zona Andina.

El corpus de España es tan amplio que favorece también la documentación de diminutivos en artículos y numerales (0.05) y en gerundios (0.02), por muy infrecuentes que sean: son valores bajísimos en comparación con los de países americanos (por ejemplo: 0.17 y 0.20, respectivamente, en México y Centroamérica). Advertimos que la categorización del origen geográfico de algunas ocurrencias puede no ser precisa: en el CORPES XXI, hay textos que contienen hablantes de orígenes variados (Sánchez Sánchez 2019: 25). Puede haber, por lo tanto, muestras americanas de diminutivos en textos clasificados como españoles. En todo caso, los gerundios con diminutivo no son imposibles en España: los datos de Ruiz Pareja (2020: 64) confirman que el gerundio con diminutivo se documenta también en España; no es exclusivo del español americano (sin embargo, los datos del *Corpus Davies* que presenta la autora son números absolutos; no creemos que sean datos representativos si no se han normalizado).

La categoría “demostrativos, posesivos y relativos” es la única que no se documenta en España: esta categoría se encuentra exclusivamente en México y Centroamérica (0.14) y en la zona Andina (0.14), así como muy marginalmente en la rioplatense (0.01).

En el conjunto de las categorías gramaticales menos prototípicas para el uso de sufijos diminutivos destaca que tanto México y Centroamérica como la zona andina estén siempre en uno de los primeros puestos (véanse las marcas de color en la tabla 6). La siguiente lista ordena las zonas según la suma de las frecuencias normalizadas de adverbios, cuantificadores, artículos, numerales, demostrativos, posesivos, relativos y gerundios:

– México y Centroamérica:	40.79
– Andina:	29.03
– Caribe continental:	20.25
– Antillas:	17.76
– Río de la Plata:	17.49
– Chilena:	12.62
– España:	12.13

Estos datos confirman que son las variedades mexicano-centroamericana y andina las más productivas en la creación de diminutivos con categorías gramaticales menos prototípicas. Por detrás va el Caribe continental, que muestra una frecuencia relativamente elevada en los adverbios, artículos, numerales y gerundios.

A modo de resumen, destacamos que el Río de la Plata muestra una altísima frecuencia de *tokens* (pero no tanto de *types*) de diminutivos en sustantivos comunes, mientras que las categorías gramaticales menos prototípicas (adverbios, cuantificadores, artículos, numerales, demostrativos, posesivos, relativos y gerundios) son más productivas en cuanto al uso de diminutivos en México y Centroamérica y en la zona Andina.

6.1. Distribución geográfica de los resultados en función del medio (escrito vs. oral)

El uso de sufijos diminutivos se favorece en registros coloquiales y familiares. Como ya hemos visto en la sección 6.2., el tema y la tipología textual influyen significativamente en la frecuencia de los diminutivos. A continuación, nos basamos en el parámetro lengua escrita vs. lengua hablada. Somos conscientes de que la lengua oral no representa exclusivamente el registro coloquial-familiar ni la lengua escrita exclusivamente registros formales, pero para no parcelar demasiado los datos (en zonas, en tipologías y en clases de palabras) y poner en peligro la representatividad del corpus preferimos introducir sólo la división oral-escrito.

6.4.1. Variación diatópica en el subcorpus oral

Los resultados del subcorpus oral del CORPES XXI cambian bastante el panorama presentado hasta ahora. En el subcorpus oral, la zona andina es la que más diminutivos usa, seguida de las Antillas y de México y

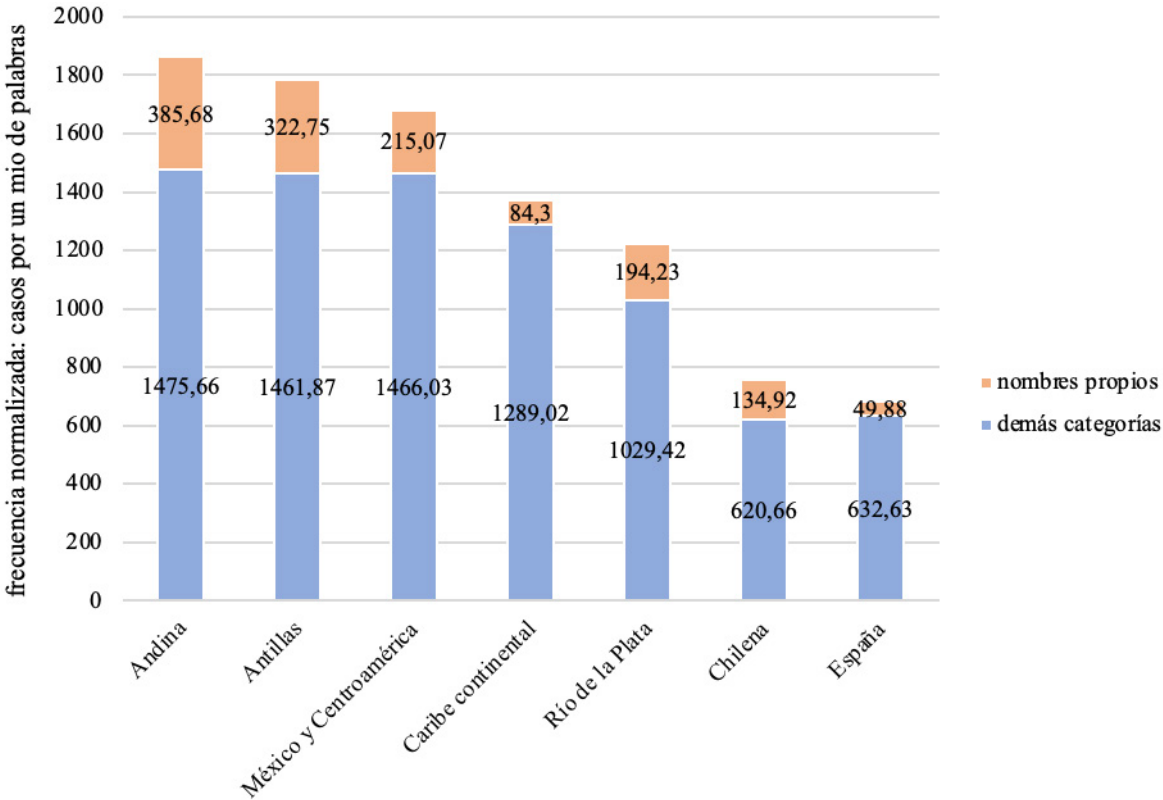
diminutivos lexicalizados sea superior al de otras regiones. Es un fenómeno difundido en toda América, que precisa de estudios cuantitativos y lexicográficos rigurosos, sopesando usos lexicalizados y no lexicalizados de determinados diminutivos en todas las áreas americanas.

Centroamérica. Si excluimos los nombres propios (por presentar problemas metodológicos, véase el apartado 5.1), estas tres zonas muestran frecuencias normalizadas muy similares: 1 475.66 en la zona andina, 1 466.03 en México y Centroamérica y 1 461.87 en las Antillas. En cambio, el Río de la Plata queda por debajo de la media americana (véase la tabla 7 y su visualización en la figura 5). Tal y como se apuntó ya en los datos anteriores, Chile y España ocupan los últimos puestos.

Tabla 7. Frecuencia normalizada (por un millón de palabras) de diminutivos en la lengua oral según zonas

	nombres propios	demás categorías	suma
Andina	385.68	1 475.66	1 861.34
Antillas	322.75	1 461.87	1 784.62
México y Centroamérica	215.07	1 466.03	1 681.10
Caribe continental	84.30	1 289.02	1 373.32
Río de la Plata	194.23	1 029.42	1 223.65
Chilena	134.92	620.66	755.58
España	49.88	632.63	682.51

Figura 5. Visualización de la dispersión geográfica de diminutivos en el subcorpus oral



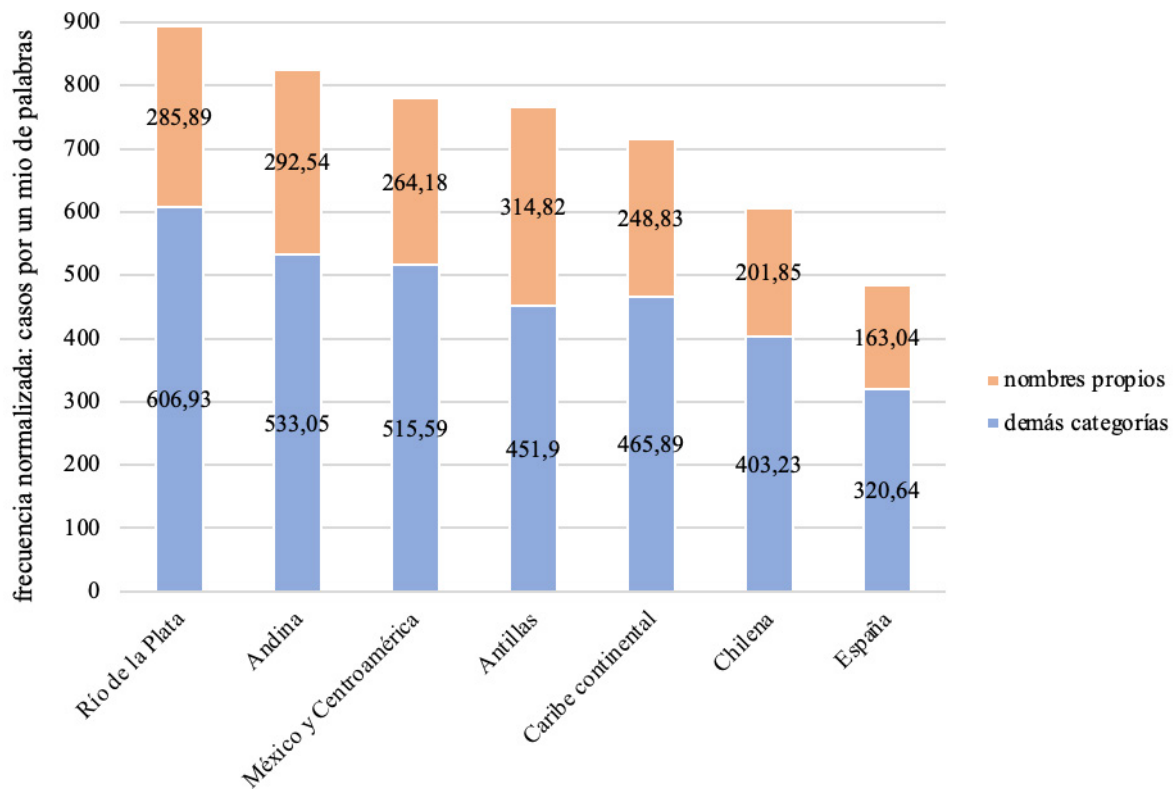
6.4.2. Variación diatópica en el subcorpus escrito

Si comparamos los datos del subcorpus oral con los del subcorpus escrito, hay un cambio drástico para la zona del Río de la Plata, que pasa al primer puesto. Las demás zonas mantienen el orden que ya se apuntó para el subcorpus oral, solo con un intercambio de puestos entre México-Centroamérica y las Antillas (véanse la tabla 8 y su visualización en la figura 6).

Tabla 8. Frecuencia relativa normalizada (por un millón de palabras) de diminutivos en la lengua escrita según zonas

	nombres propios	demás categorías	suma
Río de la Plata	285.89	606.93	892.82
Andina	292.54	533.05	825.59
México y Centroamérica	264.18	515.59	779.77
Antillas	314.82	451.90	766.72
Caribe continental	248.83	465.89	714.72
Chilena	201.85	403.23	605.08
España	163.04	320.64	483.68

Figura 6. Visualización de la dispersión geográfica de diminutivos en el subcorpus escrito

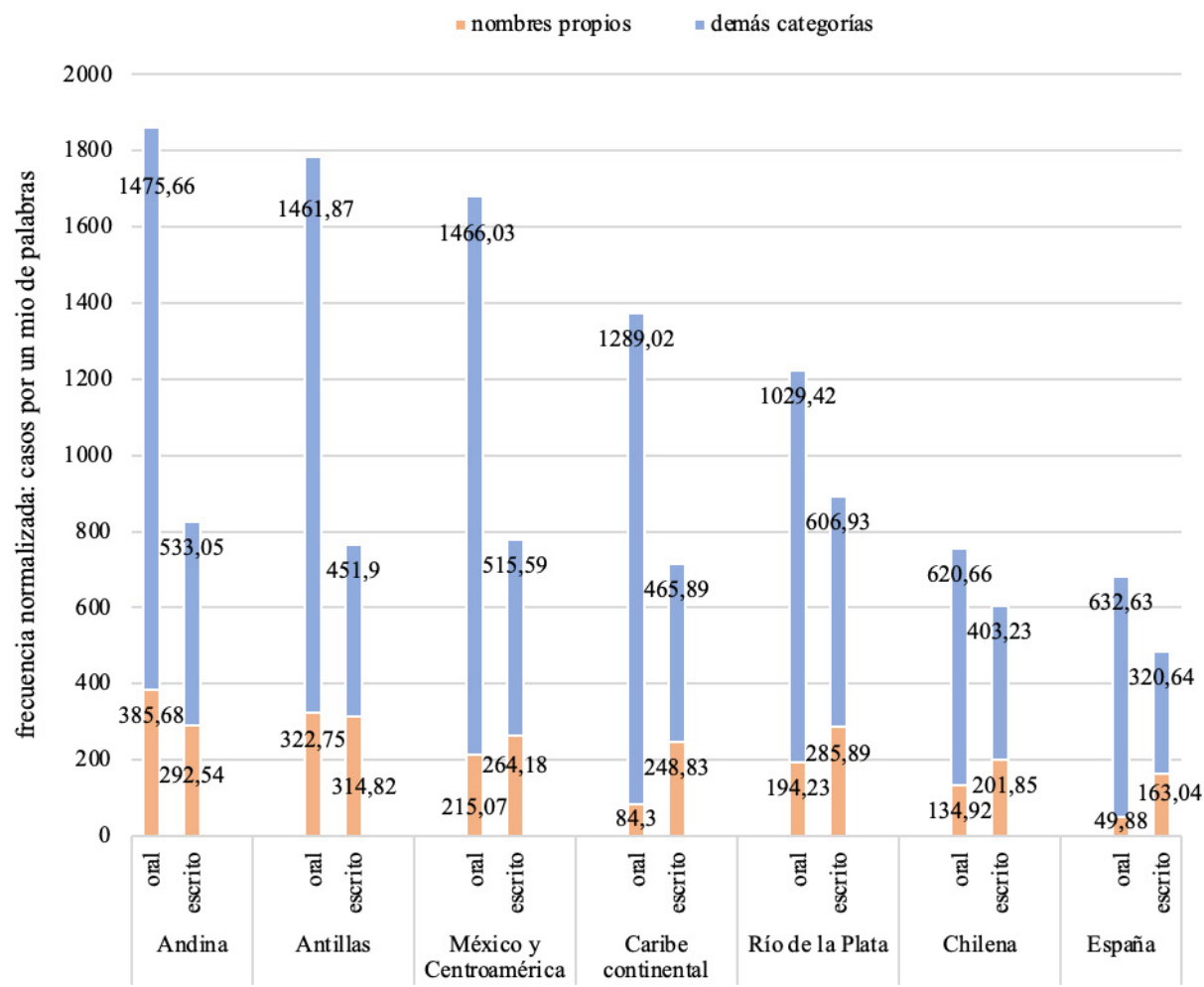


El hecho de que los datos escritos son prácticamente idénticos a los resultados generales presentados anteriormente (sección 6.1) se debe al diseño del CORPES XXI: el 90 % de las fuentes son escritas y sólo el 10 % orales, por lo cual pesan mucho más las primeras en los resultados globales.

6.4.3. Divergencias entre datos escritos y orales

Comparando los dos subcorpus se confirma que, efectivamente, en cada zona los diminutivos son más frecuentes en la lengua oral que en la lengua escrita. Esta diferencia se ve también en la figura 7 que contrasta los dos subcorpus.

Figura 7. Comparación entre los subcorpus oral y escrito por regiones



Las diferencias más pronunciadas entre zonas distintas se encuentran en el subcorpus oral, donde los diminutivos son 2.73 veces más frecuentes en la zona Andina que en España. En comparación, las diferencias diatópicas son menos pronunciadas en el subcorpus escrito, donde los diminutivos 'solo' son 1.85 veces más frecuentes en el Río de la Plata que en España. Es decir: en la lengua escrita hay menos variación diatópica que en la lengua oral. Los géneros escritos muestran una mayor homogeneidad entre regiones, ya que probablemente se acerquen más a un estándar escrito compartido (con la excepción de obras de ficción que busquen reproducir la lengua oral coloquial). En cambio, las diferencias diatópicas en cuanto al uso de diminutivos se disparan en el subcorpus oral.

Al mismo tiempo, las zonas de frecuencia más alta de diminutivos en la lengua hablada son las que muestran una mayor divergencia entre lengua hablada y escrita. En otras palabras, el gráfico parece mostrar dos bloques, el primero formado por la zona andina, antillana, mexicano-centroamericana y caribeña, con una discrepancia pronunciada entre el corpus oral y el escrito: la altísima frecuencia de diminutivos en la lengua hablada no se refleja en la misma proporción en la lengua escrita, en la que este fenómeno parece moderarse. Es, sobre todo, el caso de las clases de palabras que no son nombres propios. Estos últimos, en la zona andina, antillana y mexicano-centroamericana muestran pocas divergencias entre lengua hablada y escrita (véanse las barras naranjas en la figura 7). Por el contrario, las discrepancias son más pronunciadas en el conjunto de las demás clases de palabras (véanse las barras azules en la figura 7): en la zona andina, los diminutivos excluyendo los nombres propios son 2.77 veces más frecuentes en la lengua hablada que en la escrita, en las Antillas 3.23 veces, en México y Centroamérica 2.84 veces y en el Caribe continental 2.77 veces.

Curiosamente, la diferencia entre los dos medios es menos pronunciada en el segundo bloque que observamos en la figura 7, constituido por las zonas que muestran las frecuencias más bajas en la lengua hablada: Río de la Plata, Chile y España. En estas tres áreas, se observa un mayor acercamiento entre lengua

escrita y hablada. El caso del Río de la Plata es particular: está por debajo de la media en cuanto a la frecuencia de diminutivos en la lengua hablada, pero en la lengua escrita es la zona de frecuencia más elevada de diminutivos. En el Río de la Plata, los diminutivos, excluyendo los nombres propios, son 'solo' 1.70 veces más frecuentes en la lengua hablada que en la escrita (compárese con los valores del primer bloque, que llega a valores de entre 2.77 y 3.23). Esto puede significar que la lengua escrita rioplatense, en general, es especialmente permeable a fenómenos coloquiales y familiares. De hecho, en la Argentina del siglo XX se hace patente un especial prestigio de lo popular y de la cultura oral (Rosenblat 1961, apud López García 2009: 380).⁵ Como consecuencia, algunos géneros escritos, por ejemplo, en la prensa, muestran probablemente léxico y recursos gramaticales que serían menos aceptables en el medio escrito de otros países. Siendo el tono más familiar-coloquial una característica específica de algunos registros escritos rioplatenses y especialmente argentinos, éstos son más permeables a diminutivos que la lengua escrita de otras zonas.

La discrepancia entre el corpus oral y el escrito tiene efectos importantes sobre los resultados generales presentados en la sección 6.1, que se inclinan claramente hacia las tendencias observadas en la lengua escrita, dado que el 90 % del CORPES XXI es escrito. Sin embargo, los diminutivos son más naturales en la lengua hablada, como muestra la figura 7. Consecuentemente, creemos que, para evaluar la diferencia entre regiones en cuanto al uso de diminutivos, hay que darles prioridad a los resultados orales.

7. Conclusiones

En este trabajo hemos contrastado la frecuencia de los diminutivos en *-ito/a* en España y en seis variedades americanas (andina, antillana, caribeña continental, chilena, mexicano-centroamericana y rioplatense), basándonos en el recuento cuantitativo de unos 267 000 *tokens* de diminutivos no lexicalizados en el CORPES XXI. En cuanto a la variación diafásica, los datos confirman que los diminutivos se emplean con mucha más frecuencia en tipologías y temáticas más cercanas, familiares, informales, coloquiales y cotidianas. Esto se traduce en una mayor frecuencia de diminutivos en el subcorpus oral que en el subcorpus escrito.

A modo de resultado general, nuestros datos muestran que la frecuencia de los diminutivos es más baja en España que en América, sea cual sea la clase de palabra o el medio (oral o escrito), y también si añadimos los sufijos diminutivos *-ico/a* e *-illo/a*. Dentro de América, la región del uso más bajo de diminutivos es la chilena. Ahora bien, la interpretación de las demás zonas americanas varía en función del medio: en el subcorpus oral predominan la zona andina, México y Centroamérica y las Antillas, seguidos por el Caribe continental y el Río de la Plata (en este orden).

En el caso de las zonas andina y mexicana, varios autores han explicado la mayor frecuencia de diminutivos con una posible influencia del quechua, aimara y náhuatl (por ejemplo, Escobar 2000, Reynoso Noverón 2001). A pesar de que nuestros resultados confirman que los diminutivos son altamente frecuentes en estas dos zonas, habría que matizar la hipótesis: en nuestros resultados, la zona antillana presenta una frecuencia muy similar a la zona andina y mexicano-centroamericana. Siendo precisamente una zona sin ningún adstrato indígena, las Antillas contradicen la hipótesis de una influencia indígena. En todo caso, el estudio presentado es meramente cuantitativo y no permite sacar conclusiones sobre el papel del contacto lingüístico con lenguas indígenas. Nuestros resultados simplemente indican que este no puede ser el único factor que explica la proliferación de diminutivos en la lengua hablada de algunas regiones americanas.

Observamos que la variación diatópica es más pronunciada en la lengua hablada que en la lengua escrita. En las variedades más propensas al uso de sufijos diminutivos en el subcorpus oral (Andina, Antillas, México y Centroamérica, seguidos por el Caribe continental) hay una notable diferencia entre el subcorpus oral y escrito: los diminutivos no se producen con la misma facilidad en la lengua escrita; esta última es menos permeable al diminutivo como rasgo coloquial o familiar.

En cambio, en el corpus escrito, la frecuencia más elevada de diminutivos se documenta en el Río de la Plata, especialmente en Argentina. En esta región, justamente, se observan menos diferencias entre lengua hablada y escrita. Explicamos esta particularidad con la mayor permeabilidad de fenómenos populares, coloquiales y familiares en la lengua escrita rioplatense, lo cual aumenta el uso de diminutivos en el subcorpus escrito de esta zona.

Mientras que en el Río de la Plata se documenta un mayor uso de diminutivos sobre sustantivos, el uso más productivo del sufijo diminutivo en categorías gramaticales no prototípicas se da en México y Centroamérica, en la zona andina y en el Caribe continental (en este orden). Es en estas variedades en las que se documenta una frecuencia más elevada de diminutivos sobre adverbios, cuantificadores, artículos, numerales, demostrativos, posesivos, relativos y gerundios.

La variación diafásica es mucho más pronunciada que la variación diatópica; es decir, la configuración situacional (registro, medio, tema, grado de formalidad y de familiaridad entre interlocutores, etc.) es claramente el factor clave para el uso de diminutivos. Por lo tanto, para realmente poder observar variación diatópica, es imprescindible recurrir a un corpus perfectamente balanceado en cuanto a registros y tipologías textuales. En esto, el COPRES XXI todavía no ha llegado a ser perfecto, ya que la composición del corpus de cada país parece distorsionar la representación de diminutivos. Por ejemplo, mientras que Perú es el país con mayor uso de diminutivos, los demás países andinos (Bolivia y Ecuador) son los últimos países americanos en frecuencia de diminutivos. Creemos que esto se debe a que la composición de los corpus de Bolivia

⁵ Esta observación concierne, por ejemplo, la extensión total del voseo a dispensas del tuteo en todas las clases sociales y en todos los registros en Argentina (para un panorama actual del voseo, véase, por ejemplo, Bertolotti 2015: 38-39). Otro ejemplo sería la incorporación y popularización de vocabulario de origen lunfardo en el habla argentina general (Fajardo Aguirre 1998).

y Ecuador da más peso a géneros textuales más formales, mientras que la composición del corpus de Perú muestra una composición más inclinada hacia géneros textuales más informales —y además en una proporción comparable y equivalente a España y Argentina—.

Respecto a las posibilidades metodológicas del CORPES XXI y sus limitaciones hemos mostrado que, por un lado, permite obtener con relativa facilidad una cantidad altamente representativa de datos relevantes aplicando distintos filtros. Son pocos los casos que se escapan de nuestras búsquedas. Éstos, precisamente, muestran las limitaciones del CORPES XXI en lo que respecta a fallos e incoherencias en la lematización y anotación automática. En todo caso, creemos que estos fallos son numéricamente tan escasos que no distorsionan los resultados globales presentados en este trabajo, a excepción posiblemente de los nombres propios, para los cuales los resultados presentados son más provisionales.

Las conclusiones presentadas en este trabajo se basan, sobre todo, en la frecuencia normalizada de *tokens*, sin discriminar de manera sistemática el papel de determinados *types* que pueden llegar a ser altamente frecuentes en determinadas variedades y así distorsionar los resultados globales. En un trabajo futuro, pues, se debería analizar más en detalle la relación entre *tokens* y *types* (por ejemplo, hemos detectado 4 501 *types* de sustantivos comunes con diminutivo no lexicalizados, 1 191 *types* para adjetivos y 40 para adverbios). Además, en estos inventarios, habría que refinar el análisis acerca de aquellos diminutivos que se lematizan con el lema base en el CORPES XXI a pesar de presentar un significado lexicalizado en alguna variedad particular. Asimismo, habría que refinar el análisis de aquellos *types* que son polisémicos en una o varias variedades, es decir, con *tokens* que corresponden en parte a un diminutivo productivo y en parte a uno lexicalizado.

Agradecimientos

Agradezco enormemente a los/las dos revisores/as anónimos/as sus sugerencias acertadas y valiosas. Sus propuestas han mejorado considerablemente este trabajo y han aportado mucho a la discusión presentada. Por supuesto, posibles errores y malinterpretaciones siguen siendo míos.

Referencias bibliográficas

- Aleza Izquierdo, Milagros. 2010. "Morfología y Sintaxis: Observaciones gramaticales de interés en el Español de América". En Milagros Aleza Izquierdo y José M. Enguita Utrilla (eds.), *La lengua española en América: Normas y usos actuales*, pp. 95-224. Valencia: Universitat de València.
- Aleza Izquierdo, Milagros. 2017. "Sobre los americanismos léxicos con sufijos 'diminutivos' en el Diccionario de americanismos (2010)". *Bulletin of Spanish Studies* 94 (5): 711-743. <https://doi.org/10.1080/14753820.2017.1296247>
- Alonso, Amado. 1930. "Para la lingüística de nuestro diminutivo". *Humanidades* 21: 35-41.
- Alonso, Amado. 1935. *Noción, emoción, acción y fantasía en los diminutivos*. Madrid: Gredos.
- Asociación de Academias de la Lengua Española. (2010). *Diccionario de americanismos* [en línea]. Madrid: Santillana.
- Bertolotti, Virginia. 2015. *A mí de vos no me trata ni usted ni nadie. Sistemas e historia de las formas de tratamiento en la lengua española en América*. México/Montevidео: Universidad Nacional Autónoma de México/Universidad de la República.
- Calderón Campos, Miguel. 2024. "Corpus Size and Tagging: Methodological Strategies for Research on the History of Diminutives -ito, -illo, and -ico". En Miguel Calderón Campos y Gael Vaamonde (eds.), *Linguistic Corpora and Big Data in Spanish and Portuguese*, pp. 59-84. Berlin, Boston: De Gruyter. <https://doi.org/10.1515/9783110781465-004>
- Company Company, Concepción. 2002. "Gramaticalización y dialectología comparada. Una isoglosa sintáctico-semántica del español". *Dicenda. Cuadernos de Filología Hispánica* 20: 39-71. <http://revistas.ucm.es/index.php/DICE/article/view/DICE0202110039A/12293>
- CORPES XXI = Real Academia Española. *Banco de datos (CORPES XXI) [en línea]. Corpus del Español del Siglo XXI. Versión 1.0*. <http://www.rae.es>, consultado en noviembre de 2023.
- COSER = Inés Fernández Ordóñez (dir.). 2005-. *Corpus Oral y Sonoro del Español Rural*. <http://www.corpus-rural.es>
- Criado de Diego, Cecilia & M. A. Antonieta Andión Herrero. 2016. "Variación y variedad del diminutivo en español a través de dos corpus originales. Apuntes para su enseñanza como lengua extranjera". *Revista Electrónica de Lingüística Aplicada* 15: 87-108. <https://rael.aesla.org.es/index.php/RAEL/article/view/283>
- Escobar, Ana M. 1998. "Relaciones hablante-enunciado y hablante oyente en el español en contacto con el quechua". En Julio Calvo y Daniel Jorques (eds.), *Estudios de Lengua y Cultura Amerindias*, pp. 122-144. Valencia: Universitat de València, Servei de Publicacions.
- Escobar, Ana M. 2000. *Contacto social y lingüístico. El español en contacto con el quechua en el Perú*. Lima: Pontificia Universidad Católica del Perú.
- Escobar, Ana M. 2016. "Dialectos: español andino". En Javier Gutiérrez-Rexach (ed.), *Enciclopedia de lingüística hispánica*, pp. 353-362. London, New York: Routledge.
- Fajardo Aguirre, Alejandro. 1998. "Aproximación al léxico argentino actual". *Revista de Filología de la Universidad de La Laguna* 16, 57-71.
- Fernández Lávaque, Ana M. 1998. "El diminutivo en el noroeste argentino como fenómeno de convergencia lingüística". *Boletín de Filología* 37 (1): 513-522. <https://boletinfilologia.uchile.cl/index.php/BDF/article/view/18774>

- Fontanella de Weinberg, María Beatriz. 1962. "Algunas observaciones sobre el diminutivo en Bogotá". *The-saurus: Boletín del Instituto Caro y Cuervo* 17: 556-573. http://bibliotecadigital.caroycuervo.gov.co/312/1/TH_17_003_064_0.pdf
- Hasselrot, Bengt. 1957. *Études sur la formation diminutive dans les langues romanes*. Uppsala Universitet Årskrift.
- Hu, Jingyuan. 2022. "Las teleseries como corpus para el estudio sociopragmático del diminutivo en español". *Lenguaje* 50 (1): 115-145. <https://doi.org/10.25100/lenguaje.v50i1.11470>
- Hummel, Martin. 1997. "Para la lingüística de vuestro diminutivo: los diminutivos como apreciativos". *Anuario de Estudios Filológicos* XX: 191-210.
- Jara, Margarita. 2021. "Subjetividad e intersubjetividad de los diminutivos en -it en el español amazónico peruano". *Revista Española de Lingüística* 51 (1): 23-58. <https://doi.org/10.31810/rse.51.1.2>
- Koch, Peter & Wulf Oesterreicher. 2011. *Gesprochene Sprache in der Romania: Französisch, Italienisch, Spanisch*. Berlin: De Gruyter.
- León-Castro Gómez, Marta. 2020. "El empleo del diminutivo en la ciudad de Sevilla: perspectivas sociolingüística y pragmática". *Lengua y Habla* 24: 112-131.
- López García, María. 2009. "Discusión sobre la lengua nacional en Argentina: Posiciones en el debate y repercusiones en la actualidad". *Revista de Investigación Lingüística* 12: 375-397.
- Maíz-Arévalo, Carmen. 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. <http://dx.doi.org/10.5209/CLAC.59058>
- Malaver, Irania. 2018. "Funciones del diminutivo en el español venezolano". *Cuadernos de Lingüística de El Colegio de México* 5 (2): 5-44. <https://doi.org/10.24201/clecm.v5i2.113>
- Malaver, Irania. 2021. "Estudio sociopragmático del diminutivo en Granada". En María de las Mercedes Soto Melgar et Anna Zholobova (coords.): *El español de Granada. Estudio sociolingüístico*, pp. 373-399. Bern, Berlin, Bruxelles, New York, Oxford: Peter Lang.
- Malaver, Irania & Florentino Paredes García. 2020. "Sociolinguistic patterns and processes of convergence and divergence in Spanish". *Spanish in Context* 17 (2): 317-340.
- Moreno Fernández, Francisco. 2021. "Notas epistemológicas sobre variación para una lingüística de corpus". *Revista signos* 54 (107): 919-94. <https://doi.org/10.4067/S0718-09342021000300919>
- Moreno Fernández, Francisco. 2022. "La variación geográfica y social en los corpus hispánicos". En Giovanni Parodi, Pascual Cantos-Gómez y Chad Howe (eds.), *Lingüística de corpus en español: The Routledge handbook of Spanish corpus linguistics*, pp. 296-309. London, New York: Routledge.
- Moreno-Sandoval, Antonio. 2022. "Etiquetadores morfosintácticos para corpus en español". En Giovanni Parodi, Pascual Cantos-Gómez y Chad Howe (eds.), *Lingüística de corpus en español: The Routledge handbook of Spanish corpus linguistics*, pp. 404-418. London, New York: Routledge.
- Náñez Fernández, Emilio. 1973. *El diminutivo: Historia y funciones en el español clásico y moderno*. Madrid: Gredos.
- Paredes García, Florentino. 2015. "Funciones subjetivadoras del diminutivo en el habla de Madrid". En Ana M. Cestero Mancera, Florentino Paredes García y Isabel Molina Martos (eds.), *Patrones sociolingüísticos de Madrid*, pp. 117-153. Bern: Peter Lang.
- Quesada Pacheco, Miguel Á. 2014. "División dialectal del español de América según sus hablantes Análisis dialectológico perceptual". *Boletín de Filología* 49 (2): 257-309. <https://doi.org/10.4067/S0718-93032014000200012>
- Ramírez Luengo, José L. 2016. "El diminutivo en la Bolivia andina de la primera mitad del siglo XIX: valores y funciones en el "Diario" de J. S. Vargas". *Cuadernos de investigación filológica* 42: 111-127. <https://doi.org/10.18172/cif.2969>
- Ramírez Luengo, José L. 2019. "De nuevo sobre el diminutivo en la Bolivia andina de la primera mitad del siglo XIX: Una aproximación sociolingüística". *Boletín de la Academia Peruana de la Lengua* 65_ 169-189. <https://doi.org/10.46744/bapl.201901.009>
- Real Academia Española. 2009. *Nueva gramática de la lengua española. Morfología y sintaxis*. Madrid: Espasa.
- Real Academia Española. 2014. *Diccionario de la lengua española*. Madrid: Espasa.
- Resnik, Gabriela. 2019. "Los diminutivos lexicalizados en el español rioplatense". En María L. Perassi y Martín Tapia Kwiecien (eds.), *Palabras como puentes: Estudios lexicológicos, lexicográficos y terminológicos desde el Cono Sur*, pp. 231-251. Córdoba, Argentina: Buena Vista Editora.
- Reynoso Noverón, Jeanett. 2001. *Los diminutivos en español: un estudio de diatología comparada*. Tesis doctoral, Universidad Nacional Autónoma de México, México.
- Reynoso Noverón, Jeanett. 2005. "Procesos de gramaticalización por subjetivación: el uso del diminutivo en español". En David Eddington (ed.), *Selected Proceedings of the 7th Hispanic Linguistic Symposium*, pp. 79-86. Somerville, MA: Cascadilla Proceedings Project.
- Rojo, Guillermo. 2016. "*Citius, maius, melius*: del CREA al CORPES XXI". En Johannes Kabatek (ed.), *Lingüística de corpus y lingüística histórica iberorrománica*, pp. 197-212. Berlin, Boston: De Gruyter. <https://doi.org/10.1515/9783110462357-010>
- Rojo, Guillermo. 2017. "Sobre la configuración estadística de los corpus textuales". *Lingüística* 33 (1): 121-138. <https://doi.org/10.5935/2079-312X.20170008>
- Rosenblat, Ángel. 1961. *Las generaciones argentinas del siglo XIX ante el problema de la lengua*. Buenos Aires: Instituto de Filología, FFyL-UBA.
- Ruiz Pareja, Raquel. 2020. "Andandito: sobre la construcción gerundio + diminutivo en el español actual". *Normas* 10 (1): 51-68. <http://dx.doi.org/10.7203/Normas.v10i1.18270>

- Sánchez Sánchez, Mercedes. 2019. "Corpus de referencia: CORPES XXI. Características del CORPES. Análisis de fenómenos léxicos. Análisis de fenómenos gramaticales. Búsquedas complejas." Presentación, 2 de julio de 2019. [http://gramatica.usc.es/~raquel.rivas/corpus/textuais/2019_USC_CORPES_Character%C3%ADsticas%20y%20fen%C3%B3menos\[2787\].pdf](http://gramatica.usc.es/~raquel.rivas/corpus/textuais/2019_USC_CORPES_Character%C3%ADsticas%20y%20fen%C3%B3menos[2787].pdf)
- Zuluaga Ospina, Alberto. 1970. "La función del diminutivo en español". *Thesaurus* 25: 23-48.