The Spanish Journal of Psychology (2016), **19**, e75, 1–12.

doi:10.1017/sjp.2016.75

CAMBRIDGE UNIVERSITY PRESS

Written Type and Token Frequency Measures of Fifty Spanish Derivational Morphemes

Miguel Lázaro[1], Joana Acha[2], Víctor Illera[1] and Javier S. Sainz[1]

[1] *Universidad Complutense (Spain)*
[2] *Universidad del País Vasco (Spain)*

Abstract. Several databases of written language exist in Spanish that manage important information on the lexical and sublexical characteristics of words. However, there is no database with information on the productivity and frequency of use of derivational suffixes: sublexical units with an essential role in the formation of orthographic representations and lexical access. This work examines these two measures, known as type and token frequencies, for a series of 50 derivational suffixes and their corresponding orthographic endings. Derivational suffixes are differentiated from orthographic endings by eliminating pseudoaffixed words from the list of orthographic endings (*cerveza* [beer] is a simple word despite its ending in *-eza*). We provide separate data for child and adult populations, using two databases commonly accessed by psycholinguists conducting research in Spanish. We describe the filtering process used to obtain descriptive data that will provide information for future research on token and type frequencies of morphemes. This database is an important development for researchers focusing on the role of morphology in lexical acquisition and access.

Received 9 December 2015; Revised 27 September 2016; Accepted 28 September 2016

Keywords: database, derivational morpheme, lexical access, sublexical units.

Morphological processing in visual word recognition is one of the aspects of lexical access that has generated most research over recent years. Much progress has been made since the seminal experimental and theoretical works of Butterworth (1983), Caramazza, Laudanna, and Romani (1988) or Taft and Forster (1975), which focused on whether complex words are decomposed into morphological units before lexical recognition. Progress has been made thanks to the examination of a series of variables which are crucial to the understanding of the role of morphemes in the formation of orthographic representations and lexical access to morphologically complex words. Variables which appear to have a major influence on lexical access include: i) the semantic transparency or opacity of words, that is, for example, if an opaque word such as *gatillo* (meaning "trigger" but composed of *gato* - > "cat" and *–illo* diminutive and appreciative suffix, therefore "small cat") in which the meaning of the morphemes is not related to the final meaning of the word, is morphologically segmented in the same way as the transparent word *jornalera* (female agricultural laborer), where the relationship between the components and the final meaning is clear and direct (e.g., Rastle, Davis, Marslen-Wilson, & Tyler, 2000; Schreuder & Baayen, 1995; ii) the frequency of use of both the word and the morphemes of which it is composed (e.g., Niswander, Pollatsek, & Rayner, 2000; Taft, 1994; Taft & Zhu, 1995) or the relationship between them (Hay, 2006; Kuperman & van Dyke, 2011; iii) morphological family size, that is, whether words that generate more family members are more rapidly and effectively recognized than words with smaller lexical families, (e.g., Bertram, Baayen, & Schreuder, 2000; de Jong, Schreuder, & Baayen, 2003; Keller & Schultz, 2014; Lázaro & Sainz, 2012; and, iv) morphological productivity, whether the more productive and more frequently combined suffixes in a language facilitate lexical access more than the less productive ones (e.g., Bauer, 2005; Bertram, Schreuder, & Baayen, 2000; Hay & Baayen, 2002; Keller & Schultz, 2013; Lázaro, Acha, de la Rosa, García, & Sainz, 2016; Lázaro, Illera, & Sainz, 2015; Plag & Baayen, 2009).

Despite the number of studies conducted to date, there are still some significant aspects, especially regarding the specific influence of the factors previously mentioned and their interaction in lexical access, which require more research. In the case of Spanish language, the lack of evidence on these variables mainly stems from a lack of databases with the required information. Although there are databases in Spanish with varying information on lexical (frequency, neighborhood) and sublexical measures (syllable frequency, bigrams) which influence word recognition (LEXESP, Sebastián-Gallés, Cuetos, Martí, & Carreiras, 2000; ESPAL, Duchon, Perea, Sebastián-Gallés, Martí, & Carreiras, 2013; LEXIN,

Correspondence concerning this article should be addressed to Miguel Lázaro. Universidad Complutense. Campus de Somosaguas. 28223. Pozuelo de Alarcón. Madrid (Spain). Phone: +34–913943116.
E-mail: miguel.lazaro@ucm.es

Corral, Ferrero, & Goikoetxea, 2009), there is no database with information on morphological aspects of words such as the type and token frequency of derivational suffixes (see however Davies, Izura, Soca, & Dominguez, 2015 for a database with subjective measures of complex words). In the light of this information, it is clear that access to data on the amount of words containing morphemes in a language, as well as the frequency of use of such words and morphemes, is essential to clarify their role in word formation and lexical access. An extensive database would permit more sophisticated experimental designs, more conclusive evidence and more comprehensive development of the current explanatory models (see Baayen, 2014; Moscoso del Prado-Martín, Kostic, & Baayen, 2004; Schreuder & Baayen, 1995; Taft, 1994). At present, this problem is an issue for any research aiming to study the impact of the frequency of derivational suffixes (Lázaro et al., 2015). The aim of this work is to address this deficiency and provide a database with these measures of frequency.

The role of suffixes in word recognition: State of the art

A large number of works have studied the role of derivational suffixes in the formation and recognition of orthographic representations, demonstrating their relevance for lexical access and recognition. In one of the earliest studies, Caramazza, Laudanna, and Romani (1988), using a lexical decision task, found that the time needed to reject a pseudoword composed of two real morphemes (*cant-evi*) was significantly greater than that needed to reject a pseudoword formed by a real stem and a pseudomorpheme (*cant-ovi*). The fact that it was more difficult to reject pseudowords formed by real morphemes than those without this morphological formation led to the conclusion that the lexical system activates and accesses both representations, that is to say, the stem and the grammatical morpheme, during the process of recognition (see also Burani & Laudanna, 2003; Longtin & Meunier, 2005).

Later studies reinforced this approach, finding that productivity, or the number of times a morpheme is combined with a stem in a particular language, is a key factor in facilitating the recognition of orthographic representations and lexical access. For example, in their lexical decision task with low frequency words in English, Baayen, Wurm, and Aycock (2007) found that productive suffixed words were responded to more quickly than unproductive suffixed ones. Recent studies have confirmed the importance of the productivity variable in lexical access strategies for phonologically and structurally different languages such as Arabic (Boudelaa & Marslen-Wilson, 2011), English (Ford, Davis, & Marslen-Wilson, 2010; Juhasz & Berkowitz, 2011), Spanish (Lázaro, 2012; Lázaro et al., 2015), Finnish (Bertram, Laine, & Karvinen, 1999), Dutch (Bertram, Schreuder, & Baayen, 2000), Italian (Burani & Caramazza, 1987) and German (Clahsen, Sonnensthul, & Blevins, 2003). In short, these studies show that the morphological derivation of words modulates access to lexical representations. Consequently, suffixes are key access units. Their productivity is a variable that facilitates this access since it represents the number of repetitions of a particular suffix across the language. However, evidence shows that the variable of productivity seldom explains facilitation of lexical access in isolation, but in interaction with other variables such as length, suffix frequency or the frequency of use of the words containing the suffix (Baayen, Feldman, & Schreuder, 2006; Bertram et al., 2000; Ford, Davis, & Marslen-Wilson, 2010; Giraudo & Voga, 2014). The extent to which these variables influence the way we build and access representations and whether or not they are independent is still unclear. Consequently, appropriate measures of productivity and suffix frequency in different languages need to be made available.

In Spanish, specifically, there is relatively little evidence on the role of suffixes in lexical access. For example, Duñabeitia, Perea, and Carreiras (2008) found significant priming effects when target stimuli were primed by their derivational suffixes (*dad* - > *igualdad* [equality]) compared to when the prime was morphologically unrelated (*ero* - > *igualdad* [equality]). It was also found when words were used as primes (*brevedad* [brevity] - > *igualdad* vs. *hermosura* [beauty] - > *Igualdad*, exp. 3). As an example of this lack of specific materials required for research, at the time of the study conducted by Duñabeitia et al. (2008) it was not possible to analyze separately the effects of token and type frequency, or to control the exact frequencies of the suffixes used in the experiment, despite both possibilities being potentially interesting.

Extracting type frequency values by hand for the Bosque and Pérez (1987) reverse dictionary, Lázaro (2012), analyzed the role of suffix type frequency in an unprimed lexical decision task and found that morphologically complex words were recognized more quickly when formed by high-frequency suffixes than by low-frequency ones. In a later study, Lázaro et al. (2015) observed inhibitory effects of suffix frequency i.e., morphologically complex pseudowords were recognized more quickly when formed by low-frequency suffixes than by high frequency ones. In sum, listing the words in reverse order, starting with the last letter and ending with the first, allowed the words featuring a certain suffix to be identified and counted by hand, obtaining approximate values.

Despite these interesting results, the authors recognized the need to have fully updated and computerized instruments so as to study the reasons for these response

latencies in more depth and also to discriminate between variables, improve the count and facilitate comprehensive studies in Spanish not only on type frequency (measure of productivity) but also token frequency (measure of cumulative frequency). These authors made it clear that a database with this information was a necessity for any study considering the role of morphology not only in lexical access but also in the way lexical representations are constructed in the course of the development of linguistic competence. In fact, the relevance of the role of derivational suffixes is not limited only to adult readers. It has been demonstrated that children make early use of their knowledge of the structure of words and are sensitive to sublexical units, including derivational morphemes (e.g., Singson, Mahony, & Mann, 2000; Tyler & Nagy, 1989). For example, Burani, Marcolini, and Stella (2002), in a study with Italian children, observed that in both a naming task (exp. 1) and a lexical decision task (exp. 2), differences were found between stimuli made up of roots and derivational suffixes and simple stimuli. They specifically found that pseudowords made up of stems and derivational suffixes were decided more quickly than those without this morphological composition. Furthermore, they found that pseudowords with morphological constituency produced more false alarms in the lexical decision task than simple pseudowords. This effect was replicated in a naming task conducted with children of different ages, including different group conditions, typically developing readers and children with dyslexia. Both beginning readers and children with dyslexia read complex (polymorphemic) stimuli more rapidly than simple (monomorphemic) stimuli (Burani, Marcolini, de Luca, & Zoccolotti, 2008).

Furthermore, similarly to the case of adults, suffix productivity seems to encourage the extraction of regularities and direct association between morphological components and meaning (Windsor & Hwang, 1999). Consequently, it also facilitates lexical access, although once again frequency seems to be a key factor (Lázaro, 2012). The study conducted by Bertram, Laine, and Virkkala (2000) with Finnish children is an example of the interaction between surface and suffix frequency. In a definition task, the authors found that the morphologically complex words were defined better than low frequency simple words. Moreover, when presented with high frequency words, the children defined words containing a high-productive suffix better than those with a low-productive suffix.

The need for a database with two measures

We believe that to bring some coherence to studies with both children and adults is the very first step to clarify the concepts of productivity and frequency. It is important here to remember that the notion of productivity can be operationalized in more than one way (Baayen, 2009). The previously cited studies in Spanish operationalized the productivity as type frequency, that is, the total number of different words with a certain suffix. For example, the suffix *-eja* is considered low-productive since it appears in few different words while the suffix *-ero* is considered high-productive because it appears in many new words. This operationalization is only one of the many possible[1]. Token frequency can also be considered, taking the sum of the frequencies of all the different words containing a certain suffix. As many as twenty years ago, Burani, Thornton, Iacobini, and Laudanna (1995), and Burani, Dovetto, Thornton, and Laudanna (1997) discussed the relevance of this divergence between type and token regarding derivational suffixes. Burani et al. (1995) and Burani et al. (1997) reached the theoretical conclusion that, although both variables significantly correlated, calculations based on productivity (type frequency) would be better than those based on cumulative frequency (token frequency) in order to evaluate the influence of suffixes on word recognition. This proposal is still to be studied experimentally. Other authors use different methods to explore productivity. For example, Laudanna, Burani, and Cermele (1994) analyzed this variable considering the proportion of complex words with a certain prefix in relation to the total number of words with this particular beginning. To take an example, in the case of the Spanish prefix *-sub* (*submarino*, (submarine) *subconsciente* (subconscious) there are many simple words such as *subir,* which have no derivational morpheme although orthographically they appear to have one. These words are known as pseudoaffixed words (words that are not morphologically derived, despite sharing an orthographic ending). By finding the proportion between the number of prefixed and pseudoprefixed words, Laudanna et al. (1994) obtained more comprehensive data on the productivity of the affixes analyzed in their experiment. Further possibilities exist, such as the previously cited study by Bertram et al. (2000). In order to determine which morphemes were productive and which were not, the authors asked a group of 37 adult participants to say as many words as possible including a particular suffix in a specific time limit. The three suffixes which generated the largest

[1]It must be noted that different authors use the term productivity in different ways. In the definition used by Laudanna et al. (1994) the term productivity reflects the proportion between the number of words in which a given word ending works as a suffix and the number of words in which the same ending does not play a morphological role. Baayen, however, uses the term to refer to a more general concept that would include all words subject to a certain parsing rule. We use the term to refer to the type count, that is, the number of words containing a certain suffix ("numerosity" in terms of Burani et al., 1995).

mean number of words were considered productive and the three which generated the smallest mean number of responses were considered unproductive. Although the operationalization of certain psycholinguistic variables based on subjective judgments has an empirical basis and should not be dismissed (Kuperman & van Dyke, 2013), we understand that the operationalization of suffix productivity highlights the need for appropriate material; a need which continues to exist in studies conducted in Spanish. In this work, we will focus on both type and token frequencies as representative measures of suffix productivity.

Apart from the experimental evidence previously presented, from any methodological perspective, there are also theoretical reasons that justify the need for a derivational suffix database in the study of lexical composition and word formation. The psychological reality of morphological decomposition, that is, whether the decomposition of a word into lexical subunits in word recognition can be explained in terms of orthographical and/or phonological overlapping, is a subject of constant debate. The hypothesis that morphological decomposition is a probabilistic epiphenomenon of orthographic decomposition is a theory proposed in connectionist models (Seidenberg, 2007). This is an approach, which, by necessity, requires a suffix database. In accordance with this hypothesis, word lexical subunits are a consequence of the distribution of the probability of certain combinations appearing in the lexicon. This proposal requires a database describing word properties and the way they can be decomposed into simpler sub-units.

Given this lack of material on derivational suffixes in Spanish, we: i) created a database with extensive token and type measures of derivational suffixes, taking into account and also removing pseudoaffixed words; ii) analyzed the correlations between suffix frequency and productivity in Spanish, considering also data extracted from the book by Bosque and Pérez (1987); iii) performed descriptive statistics on information drawn from the morphological database.

Method

Materials

A total of fifty Spanish derivational suffixes were selected and their token and type frequencies were computed. The masculine and feminine forms of a suffix (e.g., *-ero*, *-era*) were treated as separate suffixes. Although it could be argued that these forms should be considered in conjunction, there are no empirical data to support this idea and so we consider it more appropriate to present the data separately, so as to facilitate experimental use by researchers. This allows researchers to work with the masculine and feminine forms either in conjunction or separately. In the same vein, there are derivational forms which vary according to the thematic vowel connecting the morpheme to the word stem (e.g., *-ible, -able*). These derivational forms are treated independently. This again gives researchers the opportunity to treat these forms in conjunction or separately. The aim is to identify the suffixes at a stringent operational level, so they can be managed independently and be compared with orthographic patterns appearing to be suffixes. The suffixes selected were always treated independently if the criteria of type and token frequency indicated they were independent.

The data are taken from the ESPAL database (Duchon et al., 2013) for adults and from the Lexin database (Corral et al., 2009) for children. ESPAL is a free access electronic database containing frequency values calculated using a 300-million token database obtained from various written sources (e.g., literary texts, 23.5%; Wikipedia, 43.9%, political texts, 16.0%, news items, 8.7%). LEXIN is an early reader database offering different measures of frequency computed from 178,839 words contained in textbooks and storybooks for nursery and primary school children. For this study we have used values of frequency per million words.

Procedure

The procedure consisted in drawing the information from the reference databases and uploading them to a separate spreadsheet for each database. The words were then put in order, together with their frequencies per million. Reverse order was used so as to be able to immediately identify the words containing the target derivational suffixes. The words were ordered according to their range of use. The words selected for each suffix were ordered in columns, so the number of rows identified the number of words containing a certain suffix (type frequency). The sum of the frequency of each word gave a number identifying the token frequency. This first calculation contained both suffixed and pseudosuffixed words since at this point the words had not been filtered. It was, therefore, an orthographic count. In order to eliminate all the pseudosuffixed words and generate a morphological count, the words were examined one by one and the simple words eliminated. This elimination process was conducted after consulting the dictionary of the Royal Academy of the Spanish Language, following the criteria explained below. The first criterion is basic and the others are clarifications or amendments to this criterion from a practical point of view. The criteria are:

a) Given that derivational morphology involves the formal addition of a morpheme to a stem, only words in which the stem was identified beyond

http://dx.doi.org/10.1017/sjp.2016.75

any doubt were considered to be suffixed words. In order to confirm the existence of a stem we used the corresponding etymological or descriptive explanation of the word provided by the Royal Academy of the Spanish Language in their dictionary. If the etymological description stated that a certain word derived from another stem, it was considered complex (e.g., *caseta* [little house] etymologically described as *de casa* [from or of a house]). In the same way, if it was apparent from the definition that the word was related beyond all doubt to its stem, it was considered to be complex (e.g., *recepcionista* [receptionist], defined as *persona encargada de atender al público en una oficina de recepción* [person responsible for attending to the public in a reception office]).

b) Words were considered as complex when they had a recognized stem, regardless of the level of opacity in the relationship between stem word and derived word. In this case, the etymological description determines securely the morphological relationship. It is important to note that this criterion considers opaque words with an etymological relation with their stems to be complex, but those with no etymological relation are considered to be pseudoaffixed words, regardless of their orthographic composition (in accordance with this criterion words such as *corner* would be considered pseudoaffixed word).

c) The stem word that is used to generate the complex word must exist in Spanish. For example, the word *gaditano* refers to an inhabitant of Cadiz, which was called *Gádir* by the Phoenician settlers. Today's word for a person from Cadiz is not *cadizano* but *gaditano*, which maintains the primitive stem, although somewhat modified. Therefore, as the stem does not exist in Spanish, the word is considered to be pseudoaffixed.

d) A word stem may, at the same time, be a complex word. For example, for the count of the agentive suffix *-ista*, *constitucionalista* (*constitución-al-ista*) (constitutionalist) was included, despite the stem (*constitucional*) being complex.

e) The derived word cannot be obtained by an inflectional process. For example the second person singular of the imperative of the verb *atizar* (poke, stoke), and the third person singular of the present tense are both *atiza.* In this case we have the pseudosuffix *–iza*, but it is rejected due to its being an inflectional suffix, rather than a derivational suffix, as in *olvidadiza* (forgetful) or *pegadiza* (catchy). In other words, it is not strictly a case of the suffix *-iza.*

In short, using this procedure, we obtained four indices for each suffix in Spanish: two for type frequency and two for token frequency. In each case one index included pseudoaffixed words (orthographic data) and the other eliminated these words from the count (morphological data). This was done for both the adult and child databases.

Results

Annex 1 shows the results for each of the suffixes studied in the adult database. As a measure of validity, it was found that the correlation between the type frequency data extracted from the reverse dictionary (Bosque & Pérez, 1987) and the data obtained from the ESPAL database was significant, $r = .77$, $p < .001$, despite the reverse dictionary presenting higher values, which is a direct consequence of the nature of a dictionary. A dictionary involves an exhaustive lexical range, which is not true to the same extent in the case of frequency dictionaries based on written data collection. Annex 2 includes the data obtained from the children's frequency database, although validity measures could not be obtained due to the lack of similar dictionaries or databases freely available for children in Spanish.

A first analysis of the relationships between measures shows there is a significant positive relationship between the orthographic and morphological token frequency measures in the adult database, $r = .89$, $p < .001$, and in the children's, $r = .77$, $p < .001$. The same is true for the orthographic and morphological type frequencies for adults, $r = .74$, $p < .001$, and children, $r = .91$, $p < .001$. There is also a significant correlation between the measures of frequency of use in the two databases, for both orthographic token frequencies, $r = .79$, $p < .001$, and morphological ones, $r = .36$, $p < .05$. The same results are found in the case of productivity measures when orthographic type frequencies, $r = .68$, $p < .001$, and morphological type frequencies, $r = .78$, $p < .001$, are correlated.

There is no significant correlation between morpheme length and the two measures of morpheme type and token frequency in either the adult database ($r = .07$, $p = .62$; $r = –.13$, $p = .33$, respectively) or in the child database ($r = .20$, $p = .15$; $r = .17$, $p = .21$, respectively).

If the frequency distribution is analyzed, the distribution of token frequencies is found to be similar to that of the type frequencies in both databases. Specifically, if the type frequency is analyzed by quartiles, that is, by finding four ranges of type frequency, most examples in both databases are found to be located in the first quartile. In the case of the adult database, the first quartile comprises the suffixes up to type frequency 355, the second quartile up to 709, the third up to 1063 and the fourth up to 1418. According to these ranges the first quartile comprises 35 of the total number of suffixes (*M token* = 188), the second quartile comprises nine suffixes (*M token* = 497), the third comprises

http://dx.doi.org/10.1017/sjp.2016.75

four suffixes (*M token* = 822) and the fourth only two (*M token* = 4469). In the case of the children's database, the first quartile comprises suffixes up to type frequency 27, the second quartile up to 54, the third up to 80 and the fourth up to 107. Here, the first quartile comprises 38 suffixes (*M token* = 97), the second nine suffixes (*M token* = 981), the third two suffixes (*M token* = 2063) and the fourth only one (*M token* = 3538). Consequently, as productivity increases, the surface frequency of the whole word also increases. However, very few suffixes are found in the medium-high productivity range, especially in the case of the children's database where the linear relation between type and token frequency is noticeably high in the first quartile in comparison with the other ranges and with the adult group. Furthermore, in many of the suffixes in the children's database, especially in the first quartile, the frequency values are comparatively lower and the value of five suffixes is null. Specifically, in the adult database, the mean values of the type and token morphological frequencies per million are 270.9 (*SD* = 32.8) and 583.0 (*SD* = 1140.6) respectively, while in the child database they are 15.3 (*SD* = 21.7) and 403.6 (*SD* = 695.7). The reason for this numerical difference is the smaller numbers of words used to calculate the frequencies in the children's database. The adult database consists of the repetition of 277,771 type examples from a total of 300 million words, while the child database consists of the repetition of 13,184 examples from a total number of 178,839. This means that an example (a word) repeated ten times in the adult base, would have a frequency per million of 0.03, while in the child base the frequency would be 55.9. In order to ascertain to what extent these type frequency values are a measure of representativeness with respect to the total number of words in each database, it is important to calculate the proportion of suffixed words in relation to the total number of type examples. This proportion is higher in the child database than in the adult one (0.13 % vs. 0.09%), which shows that, despite containing fewer words, the suffixed words form a larger percentage of the total number of words in the child database than in that of the adults.

In fact, most of the sources of the LEXIN database are teaching materials for reading and writing, so that morphological words might be overdimensioned in the child database, in comparison with the adult database. This explains why the values for the token but not type frequencies in the two databases are similar. To provide a relative value that facilitates comparison of the position of morphemes across the two databases we calculated the index of representativeness (IR) (number of suffixed words/total number of type examples in the base *million words). This value allows the researcher to more easily calculate the distance between two morphemes in the two databases, once a frequency has been selected. As can be seen in figures 1 and 2, the index of representativeness shows a very low relationship with token frequency, but high correlation with morphological type frequencies in both databases, $r = 78$, $p < .001$. The dependence or independence of the measures of productivity (type) and frequency of use (token) is a key question in this study. In this respect, our data show a positive significant correlation between both measures.

Furthermore, in the adult database the IR measure correlates to the measures of both orthographic token frequency, $r = .64$, $p < .001$ and morphological token frequency, $r = .66$, $p < .001$. The same is true for the child database, which indicates that suffix productivity is related to the token frequency of both the orthographical endings, $r = .70$, $p < .001$, and the morphological ones, $r = .96$, $p < .001$. In the latter, the relationship between productivity and frequency of use is especially high, as can be observed in Figure 2. This indicates, on the one hand, that the IR is a comprehensive index of representativeness of a suffix in the lexicon, and, on the other, that the number and frequency of pseudoaffixed words (orthographic measures) should be used in research as controlled and discriminated variables.

In order to see whether these measures for orthographic endings could be a factor which dilutes the possible effects of measures of morphological frequency, we calculated the index of discrimination for the suffixes in each database (ID = number of suffixed words/total number of words sharing this orthographic ending *100). This measure is the same as the one proposed for prefixes by Laudanna et al. (1994).

In fact, in the adult database, the words with morphemes account for 54% of the words sharing this ending, while in the child database they account for 41%. This appears to mean that there is more confusion or more competition when discriminating between a suffix and a purely orthographic ending for a child than for an adult. Interestingly, when the relationship between this index and the token frequencies is analyzed, we find that there is a significant relation with morphological token frequency, $r = .43$, $p < .01$. The same result is found in the adult database, $r = .38$, $p < .05$. In other words, suffixes with a high ID correspond mainly to orthographic endings that are suffixes (for example, the ending *–oso* acts as a suffix in 89% of the words with this ending). Furthermore, in the adult database, they present higher token frequencies, that is, words with these suffixes are used more frequently. Consequently, the index of discrimination may be an important indicator, together with the measures of type and token frequency, of lexical facilitation in both children and adults.

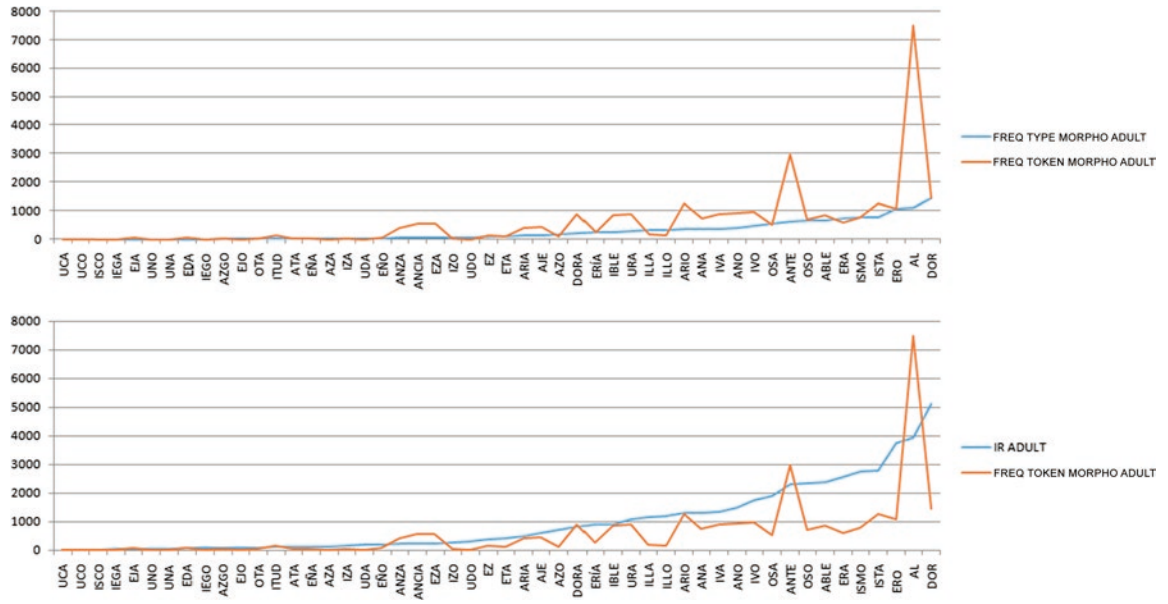

Figure 1. Relation between token and type frequency measures and IR index by suffix in the adult database.

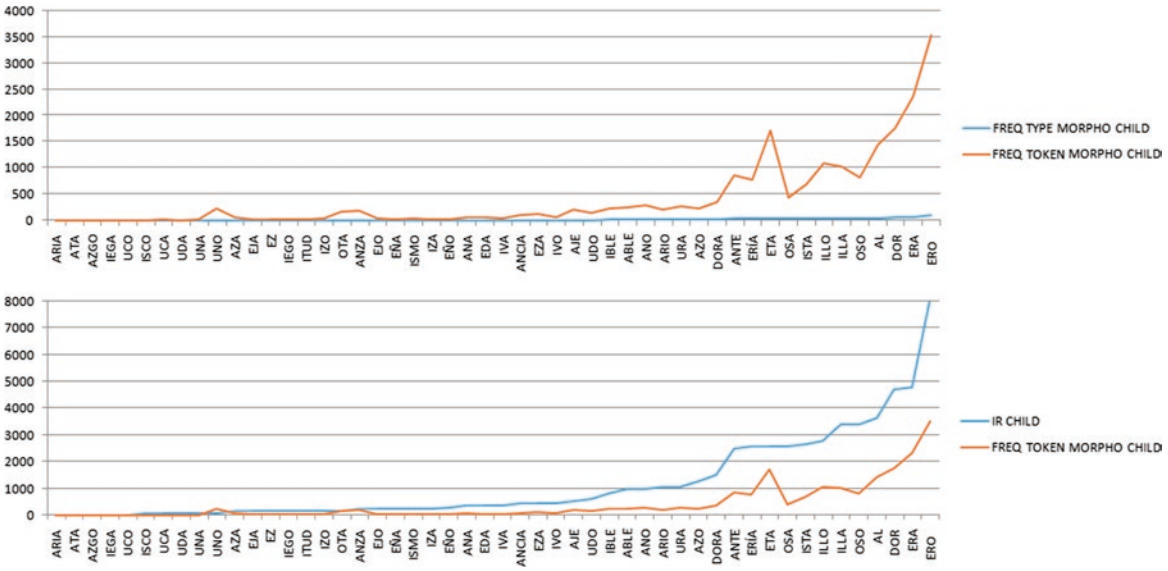

Figure 2. Relation between token and type frequency measures and IR index by suffix in the children database.

Discussion

The aim of this study was to generate a database providing reliable measures of frequency of use of the most important suffixes used in Spanish. To this end, we provided the measures of type and token frequency of 50 Spanish derivational suffixes for child and adult populations, as well as various indices that indicate to what extent these suffixes might be representative in adult and child lexicons, according to the relation between suffixes and frequencies in each database. This study is a response to the lack of morphological databases providing researchers with reliable information to use in research into Spanish language. The data we present open the way for future research into aspects related to the productivity of derivational morphemes in Spanish, using either type or token frequency.

Specifically, our analyses show a significant positive correlation between the data provided in the book by Bosque and Pérez (1987) and our orthographic data obtained using the ESPAL database. This very high significant correlation supports the use of a reverse dictionary in the studies by Lázaro (2012) or by Lázaro et al. (2015). Similarly, this significant correlation provides further guarantees regarding type frequency data based

on the ESPAL database, since they are consistent with other data based on the use of dictionaries. In this sense, it is important to note that the high correlation between our database and the other tools does not imply the information we provide is already provided by other sources. The reverse dictionary, for example, does not provide frequencies or discriminate between suffixed and pseudo-suffixed words, which is a critical measure for researchers on psycholinguistics. The Espal database, however, does provide frequencies, and many other word measures, but does not offer type and token frequencies of suffixes as morphological units. Hence, this database satisfies the need for researchers to have specific frequency values of such units.

The analyses carried out confirm the significant positive relationship between type and token frequencies in adults observed by Burani et al. (1997) or Hay and Baayen (2002). The higher the type frequency of a suffix, the higher is its token frequency. This relationship also has a significant value in the case of the children, since in this population the suffixes with higher type frequencies also presented higher token frequencies. These results were not so predictable in the case of the children since they have a much smaller vocabulary than adults and, therefore, particular high-frequency words may have had a more decisive influence. The high frequency of certain words, along with children's still limited vocabulary could well have led us not to expect the resulting correlation, since the high token frequencies did not necessarily imply the high type frequencies. However, the data invalidate this conclusion and present results similar to those for the adults. This similarity in the results between the child and adult databases is sustained not only by the high positive correlation found between the type and token frequencies between children and adults, but also by the distribution of the relationship between both frequencies and the fit of the index of representativeness. Regarding the distribution of the frequencies, in both adults and children the majority of the suffixes are located in a low range of distribution (first quartile), and in both cases few suffixes have high type and token frequencies. It is true that the most productive suffixes are those with the highest frequencies of use, but in relation to the total number of suffixes, very few present high values, and this is more evident in the child database. In this database most of the suffixes are located in a range of type frequencies between 0 and 27, and it is here where the relationship between type and token frequencies is more clearly observable, while in the adult database, the distribution of token frequencies is more heterogeneous. Due to the fact that type measures give absolute values regarding the total of lexical items on each database, we provide the index of representativeness to facilitate comparisons between type frequency values. This index is only complementary and provides a relative value, representing the location of a suffix with respect to the number of suffixed words in each database, once frequency values have been taken into account.

The results have also shown that the data on orthographic endings and morphological endings correlate positively and significantly. This correlation indicates the existence of a similar proportion of pseudosuffixes in each one of the derivational morphemes; more pseudoaffixed words in the case of the most productive suffixes and fewer pseudoaffixed words in the least productive suffixes. Initially, this seems reasonable, although it should be remembered that some suffixes have an inflectional form that notably increases orthographic frequency but considerably decreases morphological frequency. In order to control for this factor in later experimental studies, we provide the index of discrimination of each suffix, which shows the percentage of the total number of words that are suffixed words sharing the same orthographic ending. This could be an important variable when analyzing the facilitative potential of a suffix once the other variables have been controlled for.

In conclusion, our study provides interesting data on measures of productivity and frequency in Spanish suffixes, presenting information on the way these measures are distributed in the lexicons of children and adults, on the representativeness of each suffix in relation to the total number of words in both populations and on the relative frequency of a suffix in comparison to the number of words in the lexicon with the same ending. Taking into account the relevance of morpheme analysis strategies in word learning and reading in normal developing children (Casalis, Dusautoir, Colé, & Ducrot, 2009) and especially in children with learning difficulties (Casalis, Colé, & Sopo, 2004; Duranovik, Tinjak, & Turbic-Hadzagic, 2014; Siegel, 2008; Traficante, 2012), the information provided in this database could be helpful to design scaffolding programs devoted to the progressive internalization of morphological units according to their frequency and their morphological or orthographic representativeness in the lexicon.

Our results show that, despite the correlations between children and adults, the same criteria cannot be applied nor can the same suffixes be used to evaluate children and adults. Taking into account the variability by range, one possibility for further research could be to select morphemes that, within each range of type frequency, present different token frequency values. At the same time, once the range and frequencies are selected, it might be useful to compare and even work with the measure of IR, especially in transversal and comparative studies. The measures we present form a database of suffixes in Spanish that will be key for selecting and managing material in future experimental studies aimed

at analyzing the role of suffixes in the process of the formation of orthographic representations and lexical access.

References

Baayen R. H. (2009). Corpus linguistics in morphology: Morphological productivity. In A. Luedeling & M. Kyto (Eds.), *Corpus linguistics* (pp. 900–919). Berlin, Germany: Mouton De Gruyter.

Baayen R. H. (2014). Experimental and psycholinguistic approaches to studying derivation. In R. Lieber & P. Stekauer (Eds.), *Handbook of derivational morphology*. Oxford, UK: Oxford University Press.

Baayen R. H., Feldman L. B., & Schreuder R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*, 290–313. http://dx.doi.org/10.1016/j.jml.2006.03.008

Baayen R. H., Wurm H. L., & Aycock J. (2007). Lexical dynamics for low-frequency complex words. A regression study across tasks and modalities. *The Mental Lexicon*, *2*, 419–463. http://dx.doi.org/10.1075/ml.2.3.06baa

Bauer L. (2005). Productivity: Theories. In P. Stekauer & R. Lieber (Eds.), *Handbook of word-formation* (pp. 315–334). Dordrecht, The Netherlands: Springer.

Bertram R., Baayen R. H., & Schreuder R. (2000). Effects of family size for complex words. *Journal of Memory and Language*, *42*, 390–405. http://dx.doi.org/10.1006/jmla.1999.2681

Bertram S., Laine M., & Karvinen K. (1999). The interplay of words formation type, affixal homonymy, and productivity in lexical processing: Evidence from a morphologically rich language. *Journal of Psycholinguistic Research*, *28*, 213–226.

Bertram R., Laine M., & Virkkala M. M. (2000). The role of derivational morphology in vocabulary acquisition: Get by with a little help from my morpheme friends. *Scandinavian Journal of Psychology*, *41*, 287–296. http://dx.doi.org/10.1111/1467-9450.00201

Bertram R., Schreuder R., & Baayen R. H. (2000). The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 489–511. http://dx.doi.org/10.1037/0278-7393.26.2.489

Bosque I., & Pérez M. (1987). *Diccionario inverso de la lengua española* [Inverse dictionary of the Spanish Language]. Madrid, Spain: Gredos.

Boudelaa S., & Marslen-Wilson W. D. (2011). Productivity and priming: Morphemic decomposition in Arabic. *Language and Cognitive Processes*, *26*, 624–652. http://dx.doi.org/10.1080/01690965.2010.521022

Burani C., & Caramazza A. (1987). Representation and processing of derived words. *Language and Cognitive Processes*, *2*, 217–227. http://dx.doi.org/10.1080/01690968708406932

Burani C., & Laudanna A. (2003). Morpheme-based lexical reading: Evidence from pseudoword naming. In E. Assink & D. Sandra (Eds.), *Reading complex words* (pp. 241–264). New York, NY: Springer US.

Burani C., Dovetto F. M., Thornton A. M., & Laudanna A. (1997). Accessing and naming suffixed pseudo-words. In G. E. Booij & J. van Marle (Eds.), *Yearbook of morphology* (pp. 55–72). Dordrecht, the Netherlands: Kluwer.

Burani C., Marcolini S., De Luca M., & Zoccolotti P. (2008). Morpheme-based reading aloud: Evidence from dyslexic and skilled Italian readers. *Cognition*, *108*(1), 243–262. http://dx.doi.org/10.1016/j.cognition.2007.12.010

Burani C., Marcolini S., & Stella G. (2002). How early does morpho-lexical reading develop in readers of a shallow orthography? *Brain and Language*, *81*, 568–586. http://dx.doi.org/10.1006/brln.2001.2548

Burani C., Thornton A. M., Iacobini C., & Laudanna A. (1995). Investigating morphological non-words. In W. U. Dressler & C. Burani (Eds.), *Crossdisciplinary approaches to morphology* (pp. 37–53). Wien, Austria: Verlag der Österreichischen Akademie der Wissenschaften.

Butterworth B. (1983). *Lexical representation*. In B. Butterworth (Ed.), *Language production* (pp. 257–294). London, UK: Academic Press.

Caramazza A., Laudanna A., & Romani C. (1988). Lexical access and inflectional morphology. *Cognition*, *28*, 297–332. http://dx.doi.org/10.1016/0010-0277(88)90017-0

Casalis S., Colé P., & Sopo D. (2004). Morphological awareness in developmental dyslexia. *Annals of Dyslexia*, *54*, 114–138. http://dx.doi.org/10.1007/s11881-004-0006-z

Casalis S., Dusautoir M., Colé P., & Ducrot S. (2009). Morphological effects in children word reading: A priming study in fourth graders. *British Journal of Developmental Psychology*, *27*, 761–766. http://dx.doi.org/10.1348/026151008X389575

Clahsen H., Sonnensthul I., & Blevins J. (2003). Derivational morphology in the German mental lexicon: A dual mechanism account. In H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (pp. 151–125). Berlin, Germany: Mouton de Gruyter.

Corral S., Ferrero M., & Goikoetxea E. (2009). LEXIN: A lexical database from Spanish kindergarten and first-grade readers. *Behavior Research Methods*, *41*, 1009–1017. http://dx.doi.org/10.3758/BRM.41.4.1009

Davies S. K., Izura C., Soca R., & Dominguez A. (2015). Age of acquisition and imageablity norms for base and morphologically complex words in English and in Spanish. *Behavioral Research Methods*, *48*, 349–365. http://dx.doi.org/10.3758/s13428-015-0579-y

De Jong N. H., Schreuder R., & Baayen R. H. (2003). Morphological resonance in the mental lexicon. In R. H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (pp. 65–88). Berlin, Germany: Mouton de Gruyter.

Duchon A., Perea M., Sebastián-Gallés N., Martí A., & Carreiras M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods*, *45*, 1246–1258. http://dx.doi.org/10.3758/s13428-013-0326-1

Duranovic M., Tinjak S., & Turbic-Hadzagic A. (2014). Morphological knowledge in children with dyslexia. *Journal of Psycholinguistic Research*, *43*, 699–713. http://dx.doi.org/10.1007/s10936-013-9274-2

Duñabeitia J. A., Perea M., & Carreiras M. (2008). Does darkness lead to happiness? Masked suffix priming effects. *Language and Cognitive Processes*, *23*, 1002–1020. http://dx.doi.org/10.1080/01690960802164242

Ford M. A., Davis M. H., & Marslen-Wilson W. D. (2010). Derivational morphology and base morpheme frequency. *Journal of Memory and Language*, *63*, 117–130. http://dx.doi.org/10.1016/j.jml.2009.01.003

Giraudo H., & Voga M. (2014). Measuring morphology: The tip of the iceberg? A retrospective on 10 years of morphological processing. *Carnets de Grammaire*, *22*, 136–167.

Hay J. (2006). Lexical frequency in morphology: Is everything relative? *Linguistics*, *39*, 1041–1070. http://dx.doi.org/10.1515/ling.2001.041

Hay J., & Baayen R. H. (2002). Parsing and productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of morphology* (pp. 203–235). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Juhasz B. J., & Berkowitz R. N. (2011). Effects of morphological families on English compound word recognition: A multitask investigation. *Language and Cognitive Processes*, *26*, 653–682. http://dx.doi.org/10.1080/01690965.2010.498668

Keller D. B., & Schultz J. (2013). Connectivity, not frequency, determines the fate of a morpheme. *PloS One*, *8*, e69945. http://dx.doi.org/10.1371/journal.pone.0069945

Keller D. B., & Schultz J. (2014). Word formation is aware of morpheme family size. *PloS One*, *9*, e93978. http://dx.doi.org/10.1371/journal.pone.0093978

Kuperman V., & van Dyke J. A. (2011). Individual differences in visual comprehension of morphological complexity. In L. Carlson, C. Hoelscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 1643–1648). Austin, TX: Cognitive Science Society.

Kuperman V., & van Dyke J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 802–823. http://dx.doi.org/10.1037/a0030859

Laudanna A., Burani C., & Cermele A. (1994). Prefixes as processing units. *Language and Cognitive Processes*, *9*, 295–316. http://dx.doi.org/10.1080/01690969408402121

Lázaro M. (2012). The effects of base frequency and affix productivity in Spanish. *The Spanish Journal of Psychology*, *15*, 505–512. http://dx.doi.org/10.5209/rev_SJOP.2012.v15.n2.38861

Lázaro M., Acha J., de la Rosa S., García S., & Sainz J. (2016). Exploring the derivative suffix frequency in Spanish speaking children. *Reading and Writing*, 1–23. http://dx.doi.org/10.1007/s11145-016-9668-2

Lázaro M., Illera V., & Sainz J. (2015). The role of derivative suffix productivity in the visual word recognition of complex words. *Psicológica*, *36*, 165–184.

Lázaro M., & Sainz J. S. (2012). The effect of family size on Spanish simple and complex words. *Journal of Psycholinguistic Research*, *41*, 181–193. http://dx.doi.org/10.1007/s10936-011-9186-y

Longtin C. M., & Meunier F. (2005). Morphological decomposition in early visual word processing. *Journal of Memory and Language*, *53*, 26–41. http://dx.doi.org/10.1016/j.jml.2005.02.008

Moscoso del Prado Martin F., Kostic A., & Baayen R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, *94*, 1–18. http://dx.doi.org/10.1016/j.cognition.2003.10.015

Niswander E., Pollatsek A., & Rayner K. (2000). The processing of derived and inflected suffixed words during reading. *Language and Cognitive Processes*, *15*, 389–420. http://dx.doi.org/10.1080/01690960050119643

Plag I., & Baayen R. H. (2009). Suffix ordering and morphological processing. *Language*, *85*, 106–149.

Rastle K., Davis M. H., Marslen-Wilson W. D., & Tyler L. K. (2000). Morphological and semantic effects in visual word recognition: A time-course study. *Language and Cognitive Processes*, *15*, 507–537. http://dx.doi.org/10.1080/01690960050119689

Schreuder R., & Baayen R. H. (1995). Modelling morphological processing. In B. Feldman (Ed.), *Morphological aspects of language processing* (pp. 131–154). Hillsdale, NJ: Erlbaum.

Sebastián N., Cuetos F., Martí M. A., & Carreiras M. F. (2000). *LEXESP: Léxico informatizado del español*. [Lexesp: A Spanish Computerized Lexical DataBase]. Barcelona, Spain: Ediciones de la Universitat de Barcelona.

Seidenberg M. S. (2007). Connectionist models of reading. In G. Gaskell (Ed.), *Oxford handbook of psycholinguistics* (pp. 235–250). Oxford, UK: Oxford University Press.

Siegel L. S. (2008). Morphological awareness skills of English language learners and children with dyslexia. *Topics in Language Disorders*, *28*(1), 15–27. http://dx.doi.org/10.1097/01.adt.0000311413.75804.60

Singson M., Mahony D., & Mann V. (2000). The relation between reading ability and morphological skills: Evidence from derivational suffixes. *Reading and Writing*, *12*, 219–252. http://dx.doi.org/10.1023/A:1008196330239

Taft M., & Forster K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, *14*, 638–647. http://dx.doi.org/10.1016/S0022-5371(75)80051-X

Taft M. (1994). Interactive-activation as a framework for understanding morphological processing. *Language and Cognitive Processes*, *9*, 271–294. http://dx.doi.org/10.1080/01690969408402120

Taft M., & Zhu X. (1995). The representation of bound morphemes in the lexicon: A Chinese study. In L. Feldman (Ed.), *Morphological aspects of language processing* (pp. 293–316). Hillsdale, NJ: Lawrence Erlbaum Associates.

Taft M. (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology*, *57*, 745–765. http://dx.doi.org/10.1080/02724980343000477

Traficante D. (2012). From graphemes to morphemes: An alternative way to improve reading skills in children with dyslexia. *Revista de Investigación en Logopedia*, *2*, 163–185.

Tyler A., & Nagy W. (1989). The acquisition of English derivational morphology. *Journal of Memory and Language*, *28*, 649–667. http://dx.doi.org/10.1016/0749-596X(89)90002-8

Windsor J., & Hwang M. (1999). Derivational suffix productivity for students with and without language-learning disabilities. *Journal of Speech, Language, and Hearing Research*, *42*, 220–230. http://dx.doi.org/10.1044/jslhr.4201.220

ANNEX 1. Productivity and frequency measures, and suffix indexes in the adult database

SUFFIX	LENGTH	FREQ TYPE ORTHO	FREQ TOKEN ORTHO	FREQ TYPE MORPHO	FREQ TOKEN MORPHO	IR	ID
ABLE	4	676	911	661	842	2383	98
AJE	3	228	778	168	451	605	74
AL	2	1387	11387	1085	7305	3935	79
ANA	3	668	1788	366	749	1318	55
ANCIA	5	96	886	65	475	234	69
ANO	3	722	2107	416	932	1498	58
ANTE	4	695	3796	635	2962	2286	91
ANZA	4	80	522	64	415	230	80
ARIA	4	2158	1428	131	408	472	6
ARIO	4	473	1795	346	1197	1246	76
ATA	3	192	631	37	24	133	19
AZA	3	101	370	39	12	140	39
AZGO	4	28	39	26	39	94	93
AZO	3	277	339	201	93	724	73
DOR	3	1418	1443	1208	1252	4349	85
DORA	4	900	241	888	225	3197	99
EDA	3	108	525	22	76	79	20
EJA	3	86	351	14	63	50	16
EJO	3	126	906	27	12	97	21
EÑA	3	78	235	34	42	122	44
EÑO	3	106	433	57	81	205	54
ERA	3	1824	4418	716	611	2578	39
ERÍA	4	551	1079	242	277	871	44
ERO	3	1318	5743	1039	1086	3737	79
ETA	3	368	663	112	90	403	30
EZ	2	248	1299	110	155	396	44
EZA	3	101	901	68	557	245	67
IBLE	4	270	880	252	845	904	93
IEGA	4	31	72	10	3	36	32
IEGO	4	43	151	22	5	79	51
ILLA	4	443	662	319	171	1148	72
ILLO	4	447	419	323	145	1163	72
ISCO	4	42	299	7	4	25	17
ISMO	4	793	1876	768	789	2761	97
ISTA	4	809	1699	773	1264	2779	96
ITUD	4	40	348	31	144	112	78
IVA	3	471	1252	359	875	1292	76
IVO	3	556	1405	488	959	1757	88
IZA	3	400	492	48	23	173	12
IZO	3	458	593	80	41	288	17
OSA	3	655	1036	523	533	1883	80
OSO	3	722	803	648	725	2336	90
OTA	3	124	350	28	20	101	23
UCA	3	45	20	3	2	11	7
UCO	3	55	21	5	0,7	18	9
UDA	3	120	507	52	4,8	187	43
UDO	3	170	410	88	19	317	52
UNA	3	73	709	17	5	61	23
UNO	3	59	197	15	9	54	25
URA	3	446	2022	302	846	1084	68

http://dx.doi.org/10.1017/sjp.2016.75

ANNEX 2. Productivity and frequency measures, and suffix indexes in the children database

SUFFIX	LENGHT	FREQ TYPE ORTHO	FREQ TOKEN ORTHO	FREQ TYPE MORPHO	FREQ TOKEN MORPHO	IR	ID
ABLE	4	16	402	13	245	986	81
AJE	3	15	1157	7	200	531	47
AL	2	90	10041	48	1496	3641	53
ANA	3	46	4785	5	67	379	11
ANCIA	5	10	134	6	95	455	60
ANO	3	45	2946	13	246	986	29
ANTE	4	41	1612	32	856	2427	78
ANZA	4	10	389	3	195	228	30
ARIA	4	32	1090	0	0	0	0
ARIO	4	31	945	14	278	1062	45
ATA	3	30	2585	0	0	0	0
AZA	3	11	689	2	72	152	18
AZGO	4	0	0	0	0	0	0
AZO	3	27	851	17	211	1289	63
DOR	3	66	1906	62	1757	4703	94
DORA	4	22	380	22	378	1669	100
EDA	3	13	1124	4	56	303	31
EJA	3	19	1336	2	28	152	11
EJO	3	14	1156	2	28	152	14
EÑA	3	10	1330	3	22	228	30
EÑO	3	13	1034	4	28	303	31
ERA	3	141	7770	63	2359	4779	45
ERÍA	4	49	1346	34	773	2579	69
ERO	3	143	7798	107	3538	8116	75
ETA	3	80	5338	34	1707	2579	43
EZ	2	19	1604	2	22	152	11
EZA	3	17	822	6	122	455	35
IBLE	4	11	212	10	200	758	91
IEGA	4	4	173	0	0	0	0
IEGO	4	4	73	2	28	152	50
ILLA	4	75	2424	45	1012	3413	60
ILLO	4	60	2493	37	1084	2806	62
ISCO	4	2	24	1	6	76	50
ISMO	4	4	252	3	33	228	75
ISTA	4	41	978	35	739	2655	85
ITUD	4	2	28	2	28	152	100
IVA	3	11	189	5	33	379	45
IVO	3	13	302	6	73	455	46
IZA	3	17	261	3	22	228	18
IZO	3	16	699	3	61	228	18
OSA	3	52	2175	34	430	2579	65
OSO	3	58	2024	45	822	3413	78
OTA	3	18	1984	2	167	152	11
UCA	3	5	56	1	17	76	20
UCO	3	3	145	0	0	0	0
UDA	3	15	615	4	39	303	26
UDO	3	24	961	6	129	455	25
UNA	3	14	2185	1	17	76	7
UNO	3	7	1173	1	239	76	14
URA	3	38	805	14	195	1062	37

http://dx.doi.org/10.1017/sjp.2016.75