Zipf's Laws of Meaning and Semanticity in Catalan Language Acquisition

Maria Tubella Salinas[1] (0009-0000-3032-1399), Neus Català Roig[2] (0000-0002-6184-0367), Antoni Hernández-Fernández[1*] (0000-0002-9466-2704)

[1] Complexity and Quantitative Linguistics Laboratory, Institut de Ciències de l'Educació, Universitat Politècnica de Catalunya, 08034 Barcelona, Catalonia, Spain
[2] TALP Research Center, Intelligent Data Science and Artificial Intelligence Research Group (IDEAI-UPC), Computer Science Department, Universitat Politècnica de Catalunya, 08034 Barcelona, Catalonia, Spain
[*] Corresponding author's email: antonio.hernandez@upc.edu

DOI: https://doi.org/10.53482/2025_59_428

ABSTRACT

This study explores Zipf's laws of meaning and semanticity in Catalan child language acquisition, focusing on the interaction between syntactic regularities and semantic relationships. Building on previous research on semantic organization and Zipfian distributions in adult speech, using the CHILDES database, we analyse longitudinal corpora of Catalan-speaking children to test whether these statistical laws also emerge early in language development. Statistical and computational analyses show that rank–frequency distributions and related linguistic laws (Zipf's law, the Brevity law, and Heaps–Herdan's law) hold across different age groups and interaction contexts, whereas semantic regularities exhibit a weaker frequency–meaning correlation among younger speakers. However, the measure of semanticity captures the joint evolutionary changes in meaning and structural organization during early language acquisition.

Keywords: Acquisition of Catalan, semanticity, Zipf's laws of meaning, CHILDES Database, linguistic laws

1 Introduction

From Zipf's pioneering works (Zipf, 1932, 1935, 1945, 1949), previous studies have demonstrated the applicability of Zipf's laws and other well-known linguistic laws (Torre et al., 2019) to analyse communication efficiency in adult linguistic corpora (Bentz et al., 2017; Piantadosi et al., 2011). More recently, the study of Zipf's laws of meaning (Català et al., 2021; Ferrer-i-Cancho and Vitevitch, 2018) and a novel quantitative measure, the so-called *semanticity* of words, have been studied in Catalan-speaking adults (Català et al., 2023; Català et al., 2024), but its role in the acquisition of language in children remains underexplored.

Charles F. Hockett introduced semanticity as one of the key design features of human language (Hockett, 1960). This original qualitative concept refers to the capacity of linguistic signs—such as words or symbols—to convey specific meanings related to entities, actions, or features of the external world. According to Hockett, semanticity implies that stable associations exist between linguistic elements and real-world referents, enabling communication grounded in shared understanding (Hockett, 1960), and, *de facto*, in the establishment of a network between signifiers and meanings. Furthermore, Hockett (1960) emphasized the importance of the interaction between syntax and semantics, pointing out the inherent complexity in how structural and meaning-bearing components of language work together. In contemporary linguistics, the notion of semanticity has been revisited with new quantitative models, especially those emerging from the study of language as a complex network (Català et al., 2023; Català et al., 2024). Language is no longer viewed merely as a system of rules or symbols, but as a dynamic structure where words are nodes and their co-occurrence relationships form the edges. This networked perspective has gained traction through the foundational work of Ferrer-i-Cancho and Solé (2001), who demonstrated that linguistic networks exhibit small-world properties (Ferrer i Cancho, 2005; Ferrer-i-Cancho and Solé, 2001). This small-world structure facilitates efficient communication and cognitive processing and explains the emergence of some linguistic laws and properties of syntax (Ferrer i Cancho, 2005; Ferrer-i-Cancho, 2015).

Within this framework, a new quantitative measure of semanticity has been proposed (Català et al., 2023; Català et al., 2024), integrating both semantic and syntactic aspects of language. In this model, the semanticity of a word is defined in an easy way as the ratio between the number of meanings it has—its *polysemy*—and the number of distinct words that appear within a given lexical distance d in a corpus. Formally:

$$S_d(w) \propto \frac{\mu(w)}{\lambda_d(w)} \tag{1}$$

where where $\mu(w)$ is the number of meanings of the word w, and $\lambda_d(w)$ is the number of different words at distance d from word w in a sentence (independent of direction) (Català et al., 2024). By using this semanticity definition, it is observed that very high-frequency words will have a value which will tend to 0 while words that occur infrequently will show higher semanticity values.

This definition reflects an important insight from both linguistic typology and information theory: frequent words tend to be more polysemous. But this generalization has nuances, depending on whether we are talking about functional words (usually hubs of the linguistic network) or content words. High-frequency function words, like "the", have many *syntactic* connections but convey little semantic specificity, leading to low semanticity scores. In contrast, rare or specialized words —hapax legomena and

dis legomena— typically have fewer connections in the linguistic network and higher semantic density, resulting in higher semanticity. This quantitative approach bridges traditional qualitative dichotomies such as function vs. content words by offering a continuous, data-driven way to quantify meaning (Català et al., 2024).

Importantly, this model leverages the structural properties of the linguistic network to operationalize a concept that originally was once purely qualitative. Classical work by Hockett highlights that semanticity is not an intrinsic property of isolated words, but emerges from their interaction patterns within the linguistic system (Hockett, 1960), aligning with empirical findings about the small-world nature of lexical graphs (Ferrer i Cancho, 2005; Ferrer i Cancho et al., 2004). Complex network analysis thus provides a foundation for revisiting core linguistic concepts like semanticity, and for understanding how the frequency and polysemy of words are shaped by the structure of the lexicon itself.

Based on Equation 1, to decrease the influence of very frequent words on the values of $\lambda_d(w)$, a first normalization is proposed (Català et al., 2024), which involves computing the ratio between $\lambda_d(w)$ and $\lambda_{max,d}(w)$, where $\lambda_{max,d}(w)$ represents the total number of words found at distance d from the word w, regardless of direction. By doing so, it controls the phenomenon that more frequent words naturally have more neighbours just because they appear more often. When applying this normalization, the semanticity of a word is computed as

$$S_{\lambda-norm,d}(w) \propto \frac{\mu(w)}{\lambda_{norm,d}(w)} \tag{2}$$

where

$$\lambda_{norm,d}(w) = \frac{\lambda_d(w)}{\lambda_{max,d}(w)}. \tag{3}$$

Another kind of normalization would involve μ normalization. Even if a word has many dictionary entries (a proxy of the number of meanings), it is likely that in a specific corpus it will only appear in a few actual senses. This avoids overestimating the word's semanticity based on unused meanings. We therefore assume that a word will not present more meanings in a given context than in all possible linguistic context. In this case, the numerator of the semanticity is normalized by taking the minimum between the number of meanings $\mu(w)$ and the number of links that the word has in that specific corpus

$$S_{\mu-norm,d}(w) \propto \frac{\mu_{min}(w)}{\lambda_d(w)} \tag{4}$$

where now

(5) $$\mu_{min}(w) = min(\mu(w), \lambda_d(w)).$$

Lastly, a final normalization applies the two previous ones, that is, normalizing the number of connections (λ_{norm}) and normalizing the number of meanings (μ_{min}), formulated as

(6) $$S_{norm,d}(w) \propto \frac{\mu_{min}(w)}{\lambda_{norm,d}(w)}$$

To ensure computational feasibility and to reflect cognitive plausibility, the quantitative measure of semanticity is typically bounded by a maximum distance of $d = 4$ (Català et al., 2024), aligning with empirical findings about the small-world nature of lexical graphs: At this radius, the majority of the language network becomes connected (Ferrer i Cancho, 2005; Ferrer-i-Cancho and Solé, 2001), suggesting that beyond this threshold, lexical influence becomes saturated. This perspective resonates with classic works in network theory (Watts and Strogatz, 1998) and with linguistic findings showing that human languages optimize for both expressivity, syntax and cognitive economy (Ferrer i Cancho et al., 2004; Ferrer-i-Cancho et al., 2005, 2022).

This paper examines how Zipf's laws of meaning manifest in the early lexical and semantic development of children, particularly in Catalan speakers (usually bilingual Catalan-Spanish speakers). Catalan has been studied here because it is the language for which an official dictionary is available (Institut d'Estudis Catalans, 2007) and has previously been studied in detail both in oral and written form among adults (Català et al., 2021, 2024; Hernández-Fernández et al., 2019, 2023). Catalan is a Romance language spoken by over ten million people, primarily along the Western Mediterranean coast, as well as by smaller communities worldwide. It holds official status in Andorra and is recognized as a co-official language in several autonomous communities of Spain, including Catalonia, the Balearic Islands, and the Valencian Community (Català et al., 2021; Hernández-Fernández et al., 2023). Linguistically, Catalan occupies an intermediate position between the Ibero-Romance languages—such as Spanish, Portuguese, and Galician—and the Gallo-Romance group (including French, Occitan and Provençal).

In terms of linguistic complexity, Catalan exhibits moderate levels according to information theory: its entropy rate is 5.84, compared to a cross-linguistic average of 5.97 ± 0.91, ranking 202nd out of 520 languages in terms of unigram word complexity (Bentz et al., 2016; Bentz et al., 2017). Like other Romance languages, Catalan shows considerable inflectional variation in verb conjugation, while nominal morphology is comparatively limited, as nouns are only inflected in number (singular vs. plural), and grammatical gender is lexically specified rather than expressed through a nominal declensional system.

This contrasts with languages such as Latin or German, in which nouns are declined for additional grammatical categories, including case (e.g., *dominus*, *domini*, *domino* in Latin), or with Slavic languages such as Russian, in which number, case, and sometimes animacy are encoded morphologically in the noun. It also contains some distinctive vocabulary. Furthermore, derivational processes, especially suffixation, play a key role in word formation, with documented regional variation across Catalan-speaking areas (Hernández-Fernández et al., 2023).

Table 1: Summary of the studied linguistic laws in Catalan children. From left to right, the columns display: the name of the linguistic law, its mathematical formulation, a description of its parameters, and key references associated with each law.

	Mathematical formulation	Details	References
Zipf's law	$f = \frac{A}{r^{\alpha}}$	f: frequency r: word rank α, A: parameters	(Zipf, 1932, 1935, 1949)
Brevity law	$f \sim \exp(-\lambda \ell), \quad \lambda > 0$	f: frequency l: length λ: parameter	(Torre et al., 2019) (Bentz and Ferrer-i-Cancho, 2016)
Herdan-Heaps' law	$n = cT^{\theta}$	n : word types T : word tokens c, θ :parameters	(Herdan, 1960) (Heaps, 1978)
Zipf's law of meaning distribution	$\mu = C_1 r^{\gamma}$	μ : number of meanings r : word rank C_1, γ : parameters	(Zipf, 1945) (Ferrer-i-Cancho and Vitevitch, 2018)
Zipf's meaning-frequency law	$\mu = C_2 f^{\delta}$	μ : number of meanings f : frequency C_2, δ : parameters	(Zipf, 1945) (Ferrer-i-Cancho and Vitevitch, 2018)

The remainder of the article is structured as follows. Section 2 provides a concise overview of early language acquisition stages. Section 3 details the characteristics of the corpus, the preprocessing steps undertaken, and the analytical methodologies applied. We used the CHILDES corpus (MacWhinney, 2000) to analyze some classical linguistic laws and semanticity in child speech. Section 4 presents an in-depth examination of the empirical findings related to the linguistic laws under investigation. Table 1 displays the different linguistic laws examined in addition to *semanticity*. Finally, Section 5 summarizes the principal outcomes and discusses their implications.

2 Language acquisition and development

From birth, infants progress through stages of language development, starting with cooing (vowel sounds), followed by babbling (repeated syllables with consonants and vowels). This babbling is not necessarily communicative, as they express it both when there is a caregiver around and when they are alone (Gaztambide-Fernández et al., 2011). Around 1 year of age, children typically say their first word,

entering the one-word utterance stage. During this time, children know a number of words, but they only produce one-word utterances. The child's early vocabulary is limited to familiar objects or events, often nouns. Although children in this stage only make one-word utterances, these words often carry larger meaning. For example, they can say "*water*" to express "*I want water*" (Gaztambide-Fernández et al., 2011), or other types of previous vocalizations that evolve dynamically following external stimulation (Roy et al., 2015).

Exposure (or stimulation) is pivotal in language acquisition, as Roy et al. (2015) demonstrate: abundant, varied linguistic input is far from insufficient; rather, it provides the rich groundwork essential for the language learning trajectory. The traditional "poverty of the stimulus" argument, which suggests that the environmental input of children is too weak to explain the complexity of language acquisition, is effectively challenged by previous findings, showing that exposure to real-world language contains the redundancy, structure and cues needed for learning (Roy et al., 2015).

Moreover, language acquisition unfolds dynamically: As vocabulary expands, children progress from mastering simple elements toward more complex structures, begin forming simple sentences, and show an understanding of grammar rules (Gaztambide-Fernández et al., 2011; Roy et al., 2015), often demonstrated through overgeneralization (Ambridge et al., 2013). In this context, overgeneralization refers to an extension of a language rule to an exception to the rule. This reflects their grasp of language structure, even if they haven't mastered exceptions. For instance, they are able to understand that usually, in English, an 's' must be added to words in order to form their plural. Young children will overgeneralize this rule to cases that are exceptions and say things like "those two gooses" or "three mouses". Clearly, the rules of the language are understood, even if the exceptions to the rules are still being learned (Ambridge et al., 2013; Moskowitz, 1978).

Table 2: Stages of Language and Communication Development (Stevens, 2020).

Stage	Age	Developmental Language and Communication
1	From birth	Crying
2	0–6 months	Cooing
3	5/6 months	Babbling
4	12–18 months	One word utterances
5	18–24 months	Two words utterances
6	2–3 years	Sentence phase
7	3–5 years	Complex sentences

Importantly, there is a sensitive period for language acquisition (Kuhl, 2000, 2004), peaking in early childhood and tapering off around age 12, after which learning new languages becomes more difficult (Stevens, 2020). Table 2 summarizes the different stages of language and communication development,

considered later. This is obviously a statistical simplification of what is a complex and dynamic phenomenon, which depends heavily on individual traits, which are outside the scope of quantitative work with a general perspective such as that carried out here.

3 Materials and Methods

To carry out this research, a three-step solution was proposed (Tubella Salinas, 2025), as shown in Figure 1. The first phase of the study is dedicated to data collection. It involves compiling two main sources: transcribed speech data from the CHILDES database (MacWhinney, 2000), which includes child-adult interactions, and lexical data from the DIEC2 (Institut d'Estudis Catalans, 2007), a structured Catalan dictionary. These sources provide both empirical language use and normative lexical information to support the analysis. In a second phase, these data undergo a cleaning and preprocessing phase that includes grouping by age, correcting transcription irregularities, and isolating child utterances. Using tools such as those provided by the open-source library (Honnibal et al., 2020), texts are further processed through tokenization, part-of-speech tagging, lemmatization, and syntactic parsing.

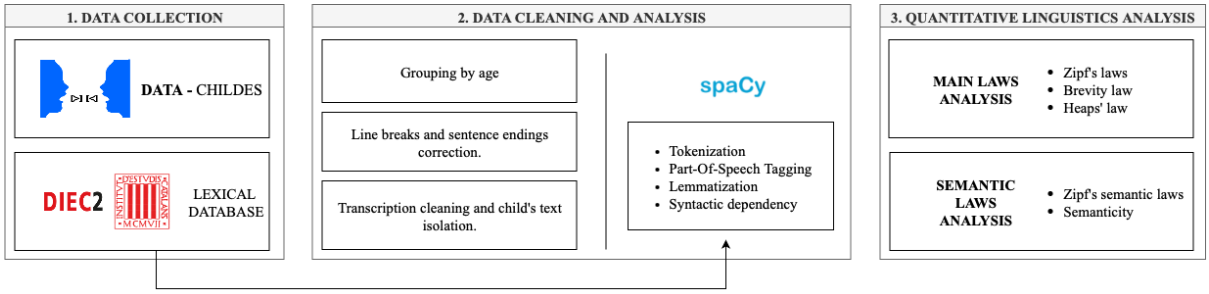

Figure 1: Schematic of the proposed analysis workflow. Source: Tubella Salinas, 2025, with permission.

Finally, the third phase consists of a quantitative linguistic analysis. This involves the study of linguistic laws to the processed data. First of all, the analysis includes Zipf's law, the Brevity law, and Herdan's law, which help characterize the structural and statistical patterns of language. Second, the focus is on Zipf's semantic laws and the concept of semanticity (Català et al., 2021, 2024), which together offer insight into how meaning is distributed and evolves in the development of children's language.

3.1 Data collection

As noted above, the study began with collecting data and obtaining the resources needed for the analyzes to be performed (Figure 1). Two main data sources were used: CHILDES, to obtain transcripts of children's conversations, and DIEC2 (Institut d'Estudis Catalans, 2007) as a lexical database for analyses requiring

word meanings. The Child Language Data Exchange System (CHILDES) is a corpus established in 1984 by Brian MacWhinney and Catherine Snow to serve as a central repository for data of first language acquisition (MacWhinney, 2000). There, researchers can find contents (transcripts, audio and video) in more than 25 languages from 230 different corpora. CHILDES has been made into a component of the larger corpus TalkBank (MacWhinney, 1999), which also includes language data from people with aphasia, second language acquisition, conversation analysis, and classroom language learning.

The data used in this study are in CHAT (Codes for the Human Analysis of Transcripts) transcription format and were obtained from five different corpora. The GRERLI corpus was compiled by Liliana Tolchinski as part of a cross-linguistic project on text construction development and includes spoken texts from 80 bilingual Catalan-Spanish participants (Llinàs-Grau, 1998). The Jordina corpus (Llinàs-Grau, 2000; Llinàs-Grau et al., 2003), directed by Mireia Llinàs-Grau, includes transcripts of three girls recorded from 1 year and 7 months of age to 2 years and 10 months as part of a study on early grammatical category acquisition. The Júlia corpus (Bel, 2001), transcribed by Aurora Bel, contains recordings of a Catalan girl in naturalistic settings from the onset of her first words until the age of 2 years and 6 months. The Mireia/Eva/Pascual corpus (Llinàs-Grau and Coll-Alfonso, 2001) includes data from three Catalan siblings recorded during daily activities at home. Lastly, the Serra and Solé (1986) longitudinal study includes recordings of ten children - monolingual and bilingual Catalan-Spanish - aged 1 to 4 years in spontaneous interactions at home. Table 3 presents a summary of the corpora used in the study.

Table 3: Summary of the corpora used in the study from CHILDES Database (MacWhinney, 2000).

Corpus Name	Range of Ages	Number of participants	Goal
GRERLI	From 9 years until 18 years old	80	Cross-linguistic project on text construction development.
Jordina	From 1 year and 7 months until 2 years and 10 months old	3	Study on early grammatical category acquisition.
Júlia	From onset of first words until 2 years and 6 months old	1	Study of language development.
Mireia/Eva/Pascual	From 1 year and 6 months until 3 years and 3 months old	3	Study of language in Catalan children during daily activities.
Serra/Solé	From 1 year until 4 years old	10	Study of language development in monolingual and bilingual Catalan-Spanish children.

On the other hand, DIEC2 is the official dictionary of Catalan (Institut d'Estudis Catalans, 2007). This resource includes information on the Part-of-Speech (PoS) of each lemma and its corresponding number of meanings, since the number of dictionary entries is taken as a proxy for the number of meanings of

each word. This, in turn, serves as an indicator of the semantic diversity of each lemma. It is important to note that the number of recorded uses is usually higher than expected. In total, the dictionary contains 70,170 entries, each consisting of a lemma with its associated PoS and number of meanings.

The initial choice for the lexical resource was WordNet (Miller, 1994), given its widespread use in computational linguistics and lexical semantics, and because it is multilingual. However, preliminary testing revealed significant limitations for the Catalan language: word coverage was notably low, and a considerable number of expected entries were missing. Due to these shortcomings, WordNet was deemed unsuitable for the task at hand, leading to the transition to the DIEC2 dictionary, following previous works (Català et al., 2021, 2024).

3.2 Data cleaning and preprocessing

This section aims to describe the main steps undertaken as part of data preprocessing and cleaning, which can be seen in the central box of Figure 1. This includes initially grouping data by age, the selection of children utterances, and finally the symbol cleaning.

Grouping by age The data collected included individuals ranging from newborns to 18-year-olds. However, data from very young children (under two years of age) were excluded due to difficulties in interpretation and legibility. Therefore, the analysis began with participants aged two and above. Based on the classifications outlined in Table 2, the first two datasets were grouped according to stages 6 and 7, resulting in two initial age groups: 2–3 years and 3–5 years. A notable gap in the data appeared between ages 5 and 9, after which the dataset expanded to include ages up to 18. To address this issue, and considering both the volume of data and the educational stages (grade school, junior high school and high school), the remaining data was organized into the following age groups: 9–10 years, 12–13 years, and 16–18 years. Overall, the dataset spans from the early stages of language development — when children are just beginning to speak — to late adolescence, when we might think that speech patterns increasingly resemble those of adults. Table 4 summarizes the aged groups that were considered along with an exploratory analysis of the data for each aged group.

Table 4: Summary of the corpora for the different age groups.

Age group	Word tokens (T)	Word types (n)	Average word length	Average sentence length
2–3 years	7335	1237	3.37 characters	2.58 words
3–5 years	18344	2054	3.43 characters	3.25 words
9–10 years	3248	643	3.56 characters	28.16 words
12–13 years	3080	683	3.66 characters	33.89 words
16–18 years	6758	1047	3.66 characters	24.49 words

Selection of children's contributions Given that the original corpus consisted of conversations between children (annotated as CHI) and adults (annotated as PAR), it was necessary to extract and retain only the utterances produced by the child speakers for subsequent analysis.

Symbol cleaning Original corpus text offered different symbology added by the investigators, such as `@i` for interjections and fillers (e.g., *bueno@i*), `@fp` for filled pauses (e.g., *ehm@fp*), `@d` for dialecticisms or `@o` for onomatopoeias. Other symbols like *, : or // that appeared in the text were also eliminated, as well as & symbols before tags such as "eh". Clitic pronouns that are conventionally spelled as one word (or that, in Catalan, are written separated by a hyphen or an apostrophe) were marked with ~ within the scope symbol `[: xxx]`, in order to allow for different types of word counts or searches. For instance, *trencar-les [: trencar~les], trenca'ls [: trenca~ls].*

When participants use a non-standard form, the correctly pronounced form follows the written/produced original word, using the scope symbol `[: xxx]` (e.g., *vem [: vam]*). However, to maintain the original spoken words and produce a more accurate study, the original words were kept, as well as some tag words used by the children. Additionally, the Catalan letter "l·l" was transcribed as "lll" and the name of the letter was indicated within the scope symbol for transcriber's comments (e.g., *colllegi [% ela geminada]*). This token was modified and the word considered was the one with the Catalan token added.

Tokenization and Sentence Splitting Sentence splitting, as it is indicated, is the process of splitting the text into different sentences. It is not trivial, as not all periods mean sentence endings. For instance, in "Dr. Smith is here." the period after "Dr" does not mean the sentence has ended. Specifically, CHILDES corpus indicate end of sentence with `(.)`. Tokenization is the process of splitting raw text into smaller, meaningful units called tokens. Tokens are usually words, but they can also include punctuation marks, numbers, or other meaningful symbols. By dividing the text into tokens, we can compute word frequencies and perform various types of linguistic or statistical analyses.

Morphological analysis and PoS-Tagging Morphological analysis breaks words down into their smallest meaningful units (morphemes) and identifies grammatical features such as tense, number, gender or case, among others, often including part-of-speech (PoS) information (e.g., noun, verb, adjective). PoS tagging then selects the most appropriate PoS based on the word's context. Traditionally, morphological analysis precedes PoS tagging, especially in morphologically rich languages like Catalan. However, more recent approaches (especially neural models), often perform both tasks jointly.

Lemmatization Lemmatization is the process of reducing a word to its base or dictionary form, called a lemma. Unlike stemming (which crudely chops off word endings), lemmatization uses linguistic knowledge to return a valid word that represents all its inflected forms. It often requires POS-tagging to be accurate, since the lemma of a word can depend on its role in a sentence. This step is highly important to obtain the meanings of a word through a dictionary, as the lemma is the form you would look up in it.

These processes of tokenization, PoS-tagging and lemmatization were done using spaCy (Honnibal et al., 2020), a library for advanced Natural Language Processing in Python and Cython. It offers pre-trained processing pipelines, which typically include a tagger, a lemmatizer, a syntactic analyzer, and an entity recognizer. It currently supports more than 70 languages, including Catalan.

Nevertheless, a manual correction was finally performed to correct some of the resulting lemmatization, specially in the infinitive forms of verbs. Because the lemma was incorrect, the dictionary search did not return any results. Despite this, words that are not correct in Catalan, such as some common contact-induced forms in Spanish like *bueno*, *vale* or *pues* were not rectified as either way they would not be found in the dictionary data.

To determine the number of meanings per word, dictionary entries were matched using the following criteria. If a single entry is found, its number of meanings is used, regardless of PoS mismatches, to allow for possible spaCy tagging errors. When multiple entries exist, the count is taken from the one matching the grammatical category; if none match, the average across all entries is used. Words not found in the dictionary are considered out-of-vocabulary and assigned zero meanings.

3.3 Binning and function fitting

To study the two Zipfian semantic laws, a binning procedure was applied to the data in order to improve the reliability and interpretability of the statistical analysis. Binning plays a crucial role in this type of analysis (Català et al., 2021). Specifically, equal-size binning was used, dividing the range of lemma frequencies or ranks into intervals containing the same number of points. Bin sizes were selected from the divisors of the total number of lemmas in each corpus to ensure that no data point was lost and that all bins had the same number of elements. In a few cases, where the total number of word types did not allow for an even division, between one and three lemmas were excluded to allow for a more balanced factorization.

In rank-frequency distributions, particularly those that follow power-law or heavy-tailed behavior, the tail contains a large number of data points (LNRE, large number of rare events). These numerous low-frequency items tend to dominate the fitting process due to their sheer quantity, while the highest-ranking items—often the most linguistically significant—are few in number and exhibit high variance, making

them susceptible to noise in traditional curve-fitting techniques (Baayen, 2008), and this imbalance can result in unstable or misleading parameter estimates (Baayen and Tweedie, 1998). To address this issue, binning was used to average over ranges of values, smoothing out statistical noise and reducing the disproportionate influence of low-frequency items when fitting models in log-log space. This approach helps balance the contribution of different parts of the distribution, rather than allowing numerous tail observations to dominate the fitting process. Binning is especially important for revealing underlying trends that would otherwise be obscured by fluctuations in both the high-rank regions (due to small sample sizes) and the long tail (due to excessive influence of numerous low-frequency observations). In this context, binning acts as a regularization technique that also helps prevent overfitting to the long tail while stabilizing parameter estimates across the entire distribution. The specific binning strategy used determines how observations are weighted, but the general effect is to create a more balanced representation of the underlying power-law relationship (Milojević, 2010; Nowak et al., 2024).

Lastly, to adjust the different functions, the function `curve_fit` from Python package `SciPy` (Jones et al., 2001) was used to find the best fit curve through the data points. This function implements a least-squares method that finds an optimal fit based on the parameterized function provided by the user.

4 Results

4.1 Zipf's, Brevity and Heaps' laws

A preliminary study of the data revealed that all different age groups comply with Zipf's law, Heaps' law and Brevity law. Figure 2 shows a linear relationship with negative slope, consistent with the idea behind Zipf's work (Zipf, 1932, 1935, 1949). The adjusted parameter α was found to be around 0.80, which falls below the established parameter $\alpha = 1$. However, this result is consistent with other studies on child data (Baixeries et al., 2013).

Figure 3 shows the fittings of Heaps' law to the dataset for each age group, while Table 5 displays the parameters found after fitting the function. Studying the different c and θ values, a primary observation is that none of the parameters show the typical values $10 \leq c \leq 100$ and $0.4 \leq \theta \leq 0.6$. However, it does achieve the expected behaviour of both them being positive and θ having a value between 0 and 1.

Analysis across the different datasets reveals that the 2–3 years age group exhibits markedly distinct parameter values compared to the other age groups. Specifically, this age group yields a c value below 1, in contrast to all other groups, which exhibit a value greater than 2. Additionally, the corresponding

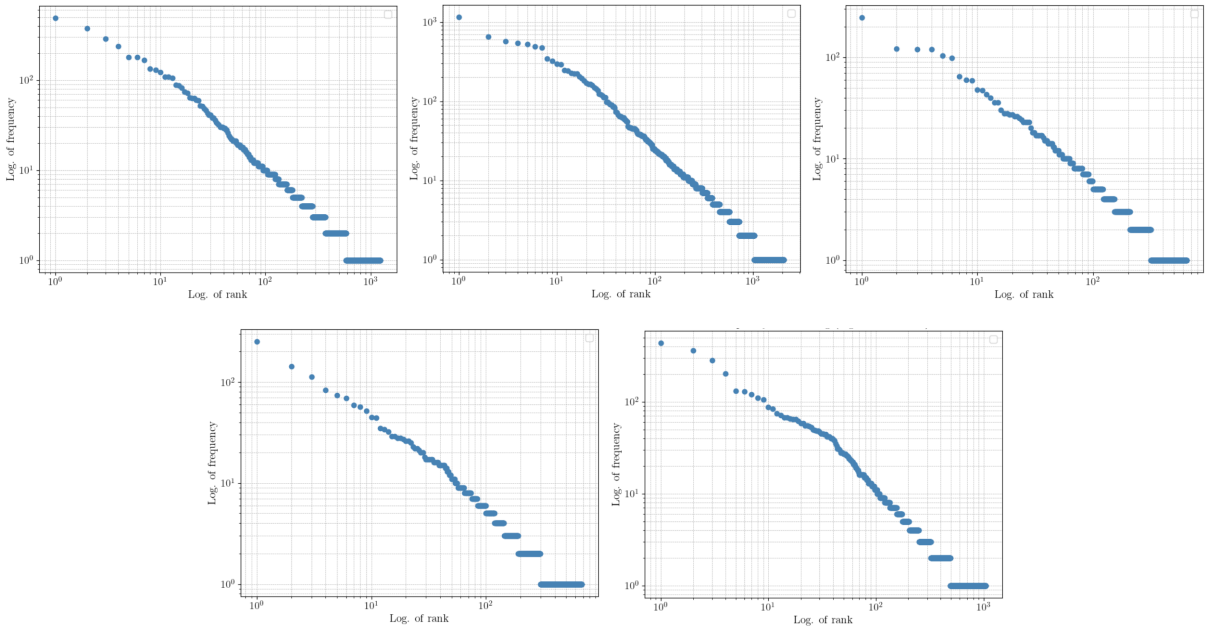

Figure 2: Zipf's law: Rank vs Frequency in logarithmic scale. Top (from left to right): age groups 2–3, 3–5, and 9–10 years old. Bottom (from left to right): age groups 12–13 and 16–18 years old.

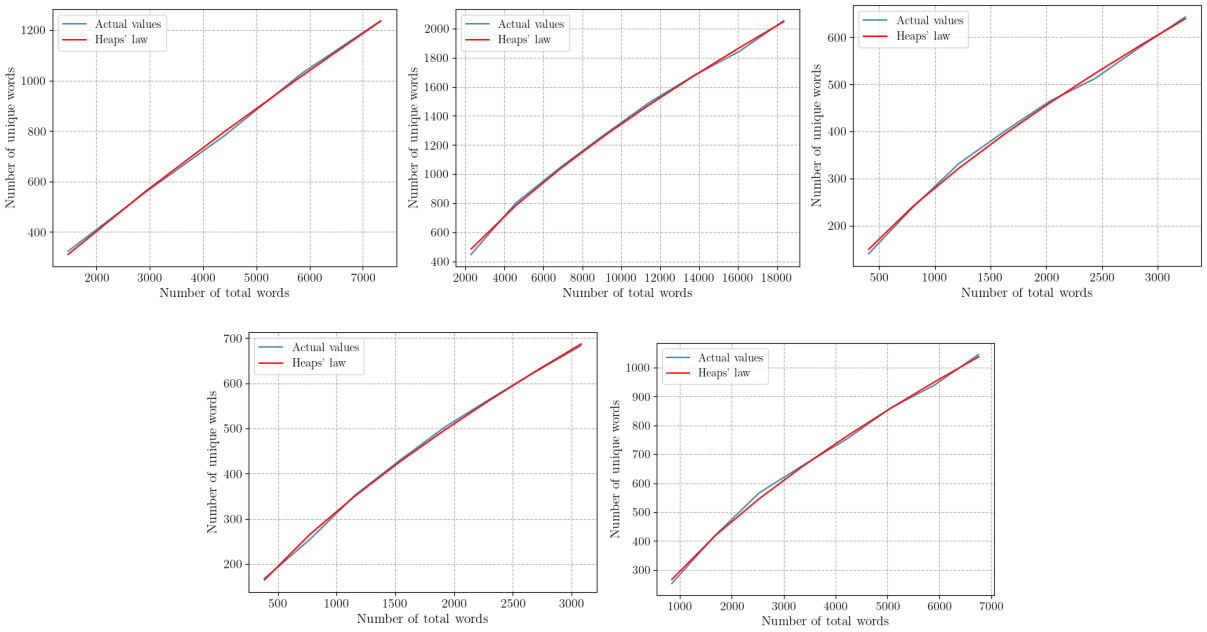

Figure 3: Heaps' law: total words vs unique words. Top (from left to right): age groups 2–3, 3–5, and 9–10 years old. Bottom (from left to right): age groups 12–13 and 16–18 years old.

Table 5: Estimated parameter values for Heaps' law obtained via function fitting.

Parameter	2–3 years	3–5 years	9–10 years	12–13 years	16–18 years
T	7335	18344	3248	3080	6758
c	0.7110 ± 0.15	2.2624 ±0.28	2.2558 ± 0.27	2.7693 ± 0.23	3.3000 ± 0.39
θ	0.8373 ± 0.03	0.6935 ± 0.01	0.6984 ± 0.02	0.6865 ± 0.01	0.6521 ± 0.01

θ value exceeds 0.80, whereas the remaining age groups show values below 0.70. Overall, with the exception of the 3–5 and 9–10 age groups, which display nearly identical parameter estimates, the c parameter (indicative of the initial vocabulary richness), demonstrates a generally increasing trend with age. This suggests that older children begin with a higher baseline lexical richness, as evidenced by the greater lexical diversity present even in smaller speech samples. This trend indicates a developmental increase in lexical knowledge and an expanded accessible vocabulary. On the other hand, the parameter θ shows a decreasing tendency. A higher θ value implies a greater probability of observing new or previously unused words as the number of spoken words increases. This suggests that younger children exhibit a higher rate of lexical innovation during speech production, whereas older individuals, having already consolidated a substantial portion of their active vocabulary, display a reduced rate of novel word introduction. This pattern is consistent with the notion that language use becomes more repetitive and automated with linguistic maturation.

As shown in Figure 3, the fitted curves of Heaps' law (in red) exhibit a strong alignment with the actual data (in blue) across the various age groups, indicating a good fit of the model.

The Brevity law was approached by using statistical correlations, in particular, Pearson's, Spearman's and Kendall's correlation tests. Pearson's correlation coefficient is a parametric measure that quantifies the strength and direction of the linear relationship between two continuous variables, and assumes normally distributed data. In contrast, Spearman's rank correlation coefficient and Kendall's Tau correlation coefficient are non-parametric measures that assess the strength and direction of monotonic associations between ranked variables, and do not assume a specific distribution (El-Hashash and Hassan, 2022). The two variables that were tested for correlation are word frequency and word length.

Table 6 presents the correlation coefficients for all three variants under investigation. In all cases, the analyses reveal statistically significant negative correlations, indicating an inverse relationship between the variables, that is, shorter words tend to occur with higher frequency. Although the absolute values of the correlation coefficients are relatively low, all associated p-values fall below the typical significance threshold of $\alpha = 0.001$, supporting the reliability of the observed associations.

4.2 Zipf's semantic laws

The law of meaning distribution characterizes the relationship between a word's number of meanings (μ) and its frequency rank (r) (see mathematical formulation in Table 1). This law formalizes the empirical finding that more frequent words tend to have more meanings.

A first observation is that all age groups follow the core pattern of the law, as all plots show a linear function with a negative slope, indicating that words in the lowest rank positions (i.e., the most frequent

Table 6: Correlation analysis between word frequency versus word length across age groups. For each correlation metric, the value of the statistic (Pearson's r, Spearman's ρ, and Kendall's τ), both the coefficient and its corresponding p-value are reported.

Correlation	2–3 years	3–5 years	9–10 years	12–13 years	16–18 years
Pearson	$r = -0.198$ (p = 6.65e-13)	$r = -0.191$ ($p <$ 2e-16)	$r = -0.279$ (p = 1.44e-13)	$r = -0.255$ (p = 4.32e-12)	$r = -0.253$ ($p <$ 2e-16)
Spearman	$\rho = -0.254$ ($p <$ 2e-16)	$\rho = -0.289$ ($p <$ 2e-16)	$\rho = -0.334$ ($p <$ 2e-16)	$\rho = -0.380$ ($p <$ 2e-16)	$\rho = -0.370$ ($p <$ 2e-16)
Kendall	$\tau = -0.205$ ($p <$ 2e-16)	$\tau = -0.232$ ($p <$ 2e-16)	$\tau = -0.272$ ($p <$ 2e-16)	$\tau = -0.308$ ($p <$ 2e-16)	$\tau = -0.299$ ($p <$ 2e-16)

words) have a greater number of meanings. Additionally, as the bin size increases, the slope becomes steeper, which corresponds to an increase in the value of γ, as shown in Table 7. It should be noted that different bin sizes were tested - including the case without binning, the results of which are provided in Table A.1 in the Appendix A - and it was observed that the most optimal γ values were obtained with the largest bin sizes. Figure 4 presents the model fit obtained using the largest bin size parameter.

Table 7: Parameter estimates for the meaning distribution and meaning–frequency laws derived from nonlinear function fitting procedures using equal size binning.

Corpus	Bin size	C_1	C_2	γ	δ
2–3 years	6	15.2589	6.0151	-0.1872	0.1682
	12	14.2111	6.0162	-0.2031	0.1674
	103	11.7892	5.7581	-0.3456	0.1983
3–5 years	13	15.7485	5.9922	-0.2040	0.1595
	26	14.6720	5.9817	-0.2246	0.1602
	79	12.8582	5.9775	-0.2710	0.1568
9–10 years	20	14.0610	9.6786	-0.1179	0.0964
	40	13.3412	9.7611	-0.1292	0.0850
	160	12.9482	9.2910	-0.2823	0.1154
12–13 years	22	12.4887	10.2097	-0.0680	0.0456
	31	12.4281	10.1949	-0.0757	0.0487
	62	12.0000	10.2616	-0.0826	0.0370
16–18 years	55	12.1917	8.8895	-0.1196	0.0823
	95	12.3021	8.8110	-0.1653	0.0889
	209	11.8638	8.6676	-0.2755	0.0950

The meaning-frequency law establishes a relationship between the frequency of a word (f) and its number of meanings (μ) (see mathematical formulation in Table 1). This law also formalizes the distributional tendency whereby words with higher usage frequency typically exhibit greater semantic ambiguity or polysemy.

In the case of the meaning-frequency law, all plots show the same tendency: as a word's frequency increases, the same happens with the average number of meanings. Similarly to the law of meaning

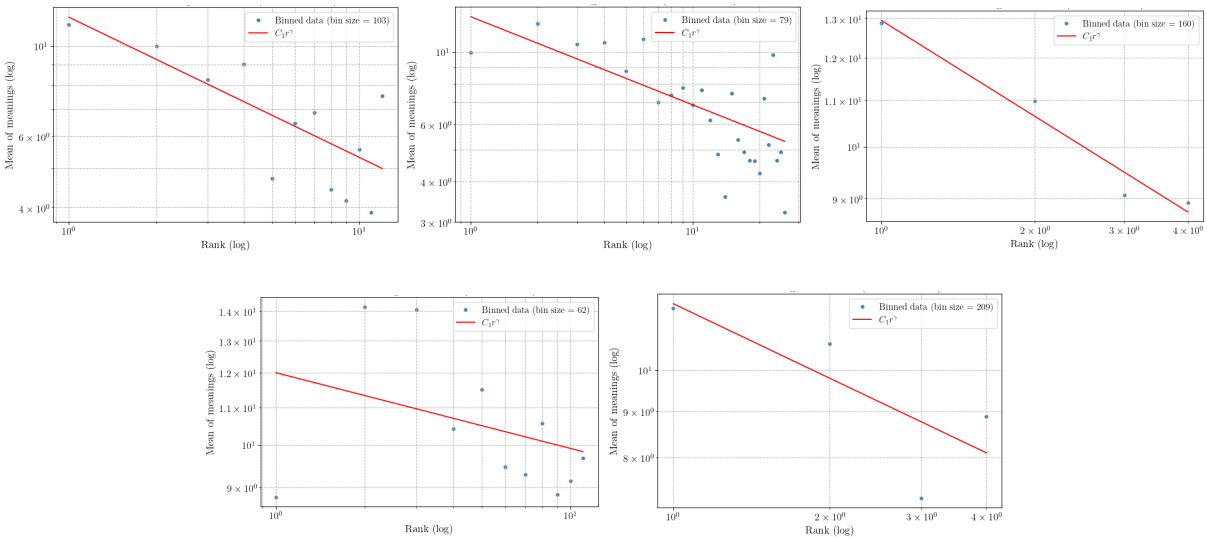

Figure 4: Zipf's law of meaning distribution: Average number of word meanings as a function of frequency rank (r), using equal size binning (blue). The red curve represents the best fitting power-law model. Top (from left to right): age groups 2–3, 3–5, and 9–10 years old. Bottom (from left to right): age groups 12–13 and 16–18 years old.

distribution, as the bin size increases, the linear function also gets steeper. The results are consistent with the results obtained before, as a low rank is equivalent to a high frequency. Figure 5 presents the model fit corresponding to the largest bin size parameter used in the analysis.

Although γ values from the meaning distribution law are very divergent from the typical value -0.5 established by Zipf (Table 1), the results showed that as the binning size increases the results get more proper to it. It must be taken into account that the original study was done taking 1000-word bins, which in this study could not be done due to the lack of data. This increase (or decrease, considering the negative sign) in the γ values is observed in all age groups. However, the 12–13 year age group exhibits parameter values that deviate significantly from those of the other groups. A similar pattern is observed for the δ parameter in the meaning-frequency law, with all estimated values showing an even greater divergence from the reference Zipfian value of 0.5. Once more, the highest parameter estimates generally correspond to the largest bin size, except in the case of the 12–13 year age group which remains an outlier.

A remarkable observation from the results of this law is that the outcome most consistent with the original studies comes from the younger children, while the results for the 12–13 year age group appear to be the poorest, despite the expectation that they would perform better due to their proximity to adulthood. A primary explanation for this lies in both the language and the size of the dataset. The established value was derived from English data using bins of size 1000, whereas this study used a much smaller

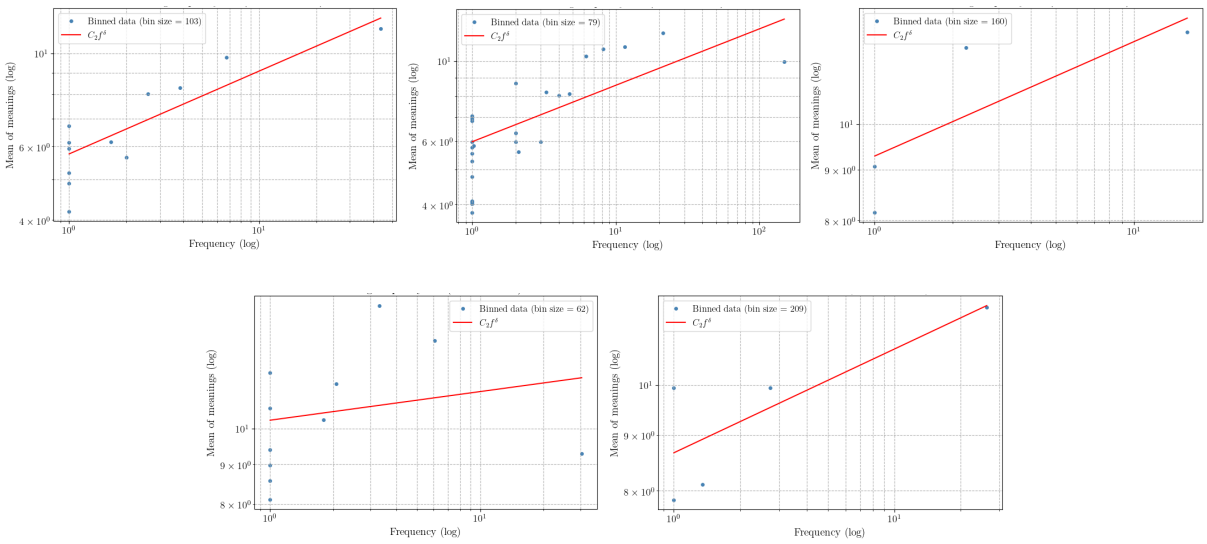

Figure 5: Zipf's meaning-frequency law: Average number of word meanings as a function of frequency (f), using equal size binning (blue). The red curve represents the best fitting power-law model. Top (from left to right): age groups 2–3, 3–5, and 9–10 years old. Bottom (from left to right): age groups 12–13 and 16–18 years old.

dataset in Catalan. One possible reason for the strong results in the 2–3 year age group is the particular nature of the data: it was obtained from videos of children interacting with their families at home. This context likely led to many of the children's words being repetitions of what parents or older siblings said. Furthermore, the data used for the analysis were not the original audiovisual recordings, but rather the available transcriptions, as explained in earlier sections. This introduces a bias, as the transcriber may have recorded the intended words rather than the exact utterances of the children. Conversely, the poor results observed in the 12–13 year age group may stem from increased data variability. The datasets corresponding to both this group and the 16–18 year group are derived from transcribed interviews with multiple speakers (children). Consequently, these datasets reflect greater linguistic heterogeneity characterized by variation in speakers' specific language use, topics, perspectives, and even verbal morphology, compared to the more homogeneous data from younger age groups.

On the other hand, a small value of δ in Zipf's meaning-frequency law indicates that the number of meanings a word has increases only slowly as its frequency increases. In other words, even if a word is used very often, it does not necessarily accumulate many additional meanings. This suggests that the relationship between frequency and polysemy is weak: frequent words are not dramatically more polysemous than less frequent ones. This pattern is consistent with the use of language by children (Casas et al., 2019). Young children tend not to exploit the full range of possible meanings of a word, often using each word with a single concrete sense. This may be due both to their limited knowledge of multiple meanings and to the nature of their speech, which is typically composed of short, closed sentences with highly predictable contexts (Tomasello, 2001).

4.3 Semanticity

Before analyzing the semanticity for the different word classes, a division between content and function words was performed in each age group (Català et al., 2024). Although function words (such as pronouns or conjunctions) constitute the most frequently lexical items in terms of token frequency, as seen in the previous section's study of Zipf's law in different corpus, this distribution changes when considering lexical types rather than tokens. Across all age group corpora, content words account for approximately 80% of the data, whereas function words comprise only the remaining 20%.

To study the relationship between semanticity and frequency rank for co-occurrences, a linear regression (LS) fit was applied to both word classes on the logarithmic scale data. An initial analysis was conducted using distance $d = 1$ to evaluate the behaviour of the semanticity measure under different normalization strategies: no normalization, normalization of the numerator, normalization of the denominator, and normalization of both components. Based on these results, subsequent analyses employed the fully normalized formulation, accounting for both the number of connections and the number of meanings (see Equation 6) across distances ranging from $d = 1$ to $d = 4$. Figure 6 presents the semanticity distributions for all subsets at $d = 4$, while Table 8 reports the corresponding linear regression slopes for each distance.

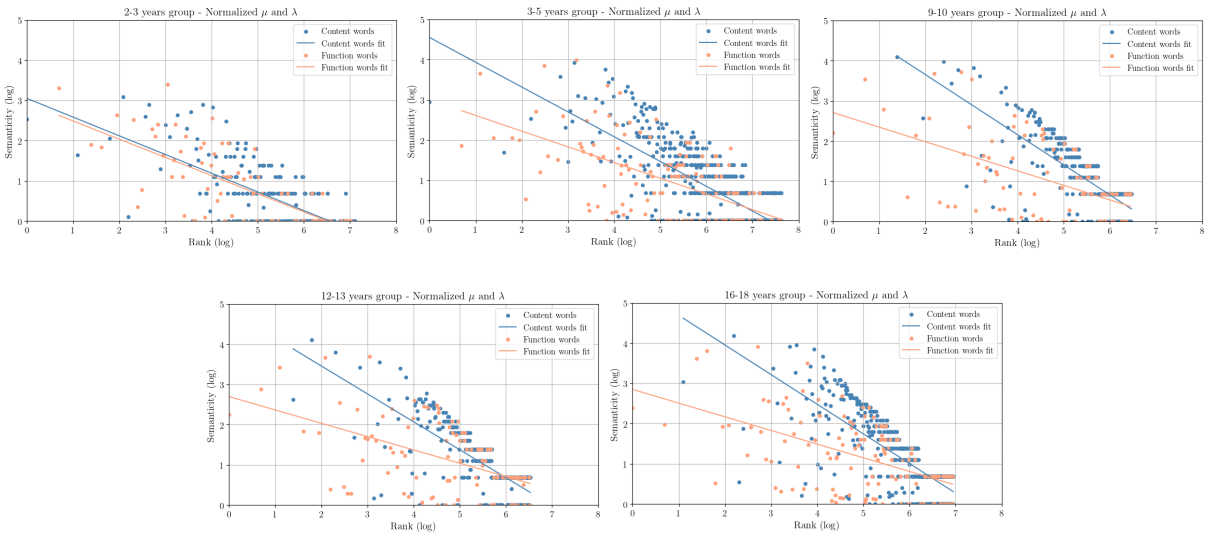

Figure 6: Frequency rank vs semanticity with μ and λ normalizations, at $d = 4$. Top (from left to right): age groups 2–3, 3–5, and 9–10 years old. Bottom (from left to right): age groups 12–13 and 16–18 years old.

Content words, irrespective of their rank or frequency, present semanticity values notably higher than those of function words. For large ranges, i.e., for low occurrence frequencies, both word classes show

Table 8: Linear regression slope coefficients quantifying the relationship between semanticity and frequency rank for content words (CW) and function words (FW), evaluated across lexical network distances ranging from 1 to 4.

Age group	Word class	Slope at d=1	Slope at d=2	Slope at d = 3	Slope at d = 4
2–3 years	CW	-0.7133	-0.6041	-0.5397	-0.4644
	FW	-0.5088	-0.4959	-0.4751	-0.4515
3–5 years	CW	-0.8447	-0.7523	-0.6809	-0.6163
	FW	-0.4527	-0.4185	-0.4023	-0.3901
9–10 years	CW	-0.8072	-0.7670	-0.7572	-0.7476
	FW	-0.3281	-0.3641	-0.3541	-0.3617
12–13 years	CW	-0.7405	-0.7036	-0.6864	-0.6939
	FW	-0.3511	-0.3322	-0.3314	-0.3308
16–18 years	CW	-0.7917	-0.7380	-0.7367	-0.7364
	FW	-0.3613	-0.3427	-0.3439	-0.3395

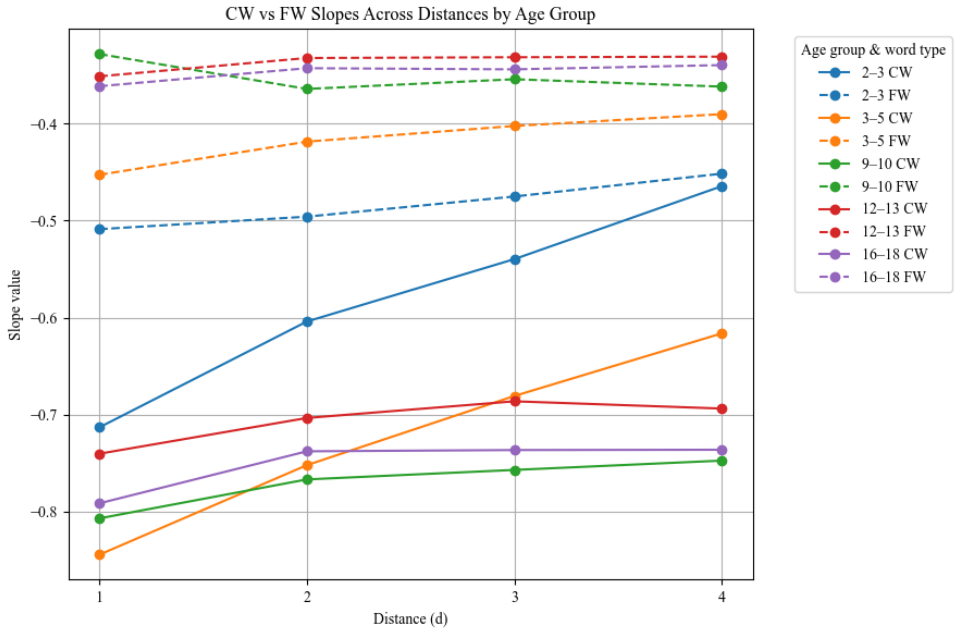

Figure 7: Comparison between the slope coefficients of linear regression of frequency versus semanticity for content words (CW) and function words (FW), evaluated across lexical network distances (d) ranging from 1 to 4.

low semanticity values, as expected. This behavior persists across co-occurrence distances ranging from 1 to 4.

Notably, in the younger age groups (2–3 years old), the distinction between content and function words diminishes as the co-occurrence distance increases. The respective linear regression trends for both word classes converge progressively, such that at $d = 4$, their slopes become nearly indistinguishable. This suggests that, at greater semantic distances, the distributional behavior of content words increasingly aligns with that of function words. This convergence is also reflected in their slope coefficients (see Table 8 and Figure 7). While the slope for function words (FW) has a subtle change from -0.5088 at $d = 1$ to -0.4515 at $d = 4$, the slope for content words (CW) shows a more pronounced shift from -0.7133 to -0.4644 over the same distance range. The resulting proximity of the two slope values at $d = 4$ accounts for the near overlap of the corresponding regression lines.

This observation can be seen as younger kids not distinguishing between content and function words, which is consistent with the usage-based theory of language acquisition (Tomasello, 2001). In it, children are seen as active participants in communication who learn language through repeated exposure to meaningful interactions. Rather than acquiring language through innate grammatical knowledge, children gradually build linguistic competence by recognizing and generalizing patterns from the input they receive. This process begins with very early expressions known as *holophrases*, single words or word-like utterances, that convey the meaning of an entire sentence. For example, a child might say "Water!" to mean "I want water" or "Ball?" to express "Where is the ball?" (Barrett, 1982). These expressions reflect the child's attempt to reproduce the full communicative intention of an adult utterance, even though they can only manage to articulate part of it. Closely related are *frozen expressions*; phrases that are learned as holophrases but will at some point be broken down into their constituent elements (Lieven et al., 1992; Tomasello, 2001). Over developmental time, children progressively segment these down into constituent units, identifying recurrent structural patterns. This process reflects the gradual emergence of grammar, a core idea in usage-based theory. Through repeated exposure, the child learns to both decompose multiword chunks into meaningful units and generate new utterances by recombining these learned components.

During these early stages of acquisition, children typically do not distinguish between content words and function words. This lack of differentiation is evident in both holophrastic and frozen expressions (Tomasello, 2001), where functional elements are either absent or embedded in unanalyzed chunks. The usage-based perspective explains this by emphasizing that children's learning is usage-driven and input-dependent. Since function words are often less salient (shorter, unstressed, less meaningful on their own), they may not initially stand out to the child. Only with increased exposure and pattern recognition

do children start to understand their grammatical role.

An additional observation regarding content words pertains to the divergence in linear regression slope behavior across two broad developmental stages: younger (ages 2–5) and older (ages 9 and above) children. This distinction becomes evident when analyzing the change in slope values across co-occurrence distances from $d = 1$ to $d = 4$. For the younger groups (2–3 and 3–5 years), the slope variations is relatively modest, approximately 0.25 in absolute magnitude. In contrast, the older groups exhibit substantially greater variations, with slope differences approaching to 0.6. This pattern suggests that the influence of semantic distance on lexical organization becomes more pronounced with age, indicating a more refined and differentiated lexical organization as language development progresses.

5 Conclusion

The analysis confirms that Zipf's rank-frequency law is consistently observed across all age groups, as evidenced by the linear patterns in the frequency versus rank double-logarithmic plots. This supports the conclusion that the law is not dependent on age or stage of linguistic development. Although the estimated exponent α was systematically lower than the canonical value of 1 (Zipf, 1949), the results align with previous findings in child language corpora (Baixeries et al., 2013). A lower exponent reflects a flatter distribution of word frequencies, meaning that the difference between high-frequency and low-frequency words is less pronounced than in adult corpora. In addition, the data also supported both the Brevity law and Herdan-Heaps' law. Regarding the Brevity law, both statistical correlation analysis and qualitative observation confirmed that the most frequent words tend to be shorter in length. This pattern was also evident in the list of the top 10 most frequent words, none of which exceeded four characters. As for Heaps' Law, a generally linear growth pattern was observed between the total number of words and the number of unique words. However, the parameters obtained differed significantly from the commonly accepted values in the literature (Herdan, 1960; Hernández-Fernández and Ferrer-i-Cancho, 2019). This suggests that younger children tend to introduce new words at a faster rate as they speak, whereas older speakers, having already developed a larger vocabulary, encounter fewer new words—consistent with the idea that language use becomes more fluent and repetitive over time.

On the other hand, the semantic laws yielded results that diverged more noticeably from those originally formulated by Zipf (Zipf, 1932, 1935, 1949). These laws highlighted the differences between children and adults in terms of lexical knowledge, particularly regarding the understanding of multiple meanings and polysemous words (Casas et al., 2019; Català et al., 2021). While the general functional trends aligned with expectations—namely, that more frequently used words tend to have more dictionary meanings—the estimated exponents deviated significantly from the canonical values reported in the literature. This discrepancy suggests that the relationship between frequency and meaning is less

pronounced in developing language users. Moreover, the analysis underscored the importance of data binning when applying these laws, as it led to smoother distributions and more interpretable patterns, reinforcing its role as a key step in semantic data analysis.

Lastly, the semanticity analysis revealed some interesting patterns when distinguishing between function and content words. Consistent with prior research (Català et al., 2023; Català et al., 2024) content words exhibited systematically higher semanticity scores, reflecting their association with a greater number of semantic interpretations or meanings. This pattern persisted even under normalization, where the semanticity measure was computed using corpus-derived sense (i.e. the observed contextual usage), rather than relying on dictionary-based counts. In the case of younger children, it was observed that as the distance considered in computing semanticity increased, content words began to behave more like function words. This trend can be explained through the usage-based theory of language acquisition (Tomasello, 2001), which posits that language is learned through repeated exposure to interactions. According to this theory, children do not initially acquire grammatical structures explicitly, but rather learn patterns of use through communication. As a result, they may rely on content words in a more functional way, using them as scaffolds for meaning before fully developing grammatical awareness. Our result suggests that semanticity may also capture aspects of cognitive and linguistic development, providing insight into how meaning and structure emerge in tandem during early language acquisition. In the age group of 2-3 year olds, semanticity-rank slopes clearly differs from that of older children and adults (Table 8), and at distance $d = 4$ there is no difference between content words and function words, which could be indicative of an indiscriminate use of words without considering syntax, something typical of frozen-type utterances in young children.

In summary, this work assessed the applicability of multiple linguistic laws in the context of Catalan language acquisition by leveraging computational methodologies and natural language processing techniques. Although empirical findings largely corroborated expected patterns for frequency-based laws, particularly Zipfian distributions, semantic-level analyses exhibited significant deviations from established formulations, suggesting variability in vocabulary and semantic development. These differences can be attributed to a variety of factors, including age, data collection methods, language-specific features, and corpus size. Despite these challenges, the findings offer valuable insights into how linguistic patterns emerge and evolve in early speakers, and highlight the importance of methodological awareness in quantitative linguistic research.

This work offers many possibilities for future research. Firstly, utilizing a larger and more uniformly collected dataset would enhance the robustness of statistical analyses and mitigate biases arising from heterogeneous data collection procedures. Furthermore, ensuring a more continuous and balanced

age distribution across the full developmental span from ages 2 to 18, would facilitate a more precise characterization of linguistic progression and transitional phases. The current study is limited by a significant data gap between ages 5 and 9, and incorporating additional data within this interval would likely yield more reliable insights, particularly with respect to syntactic measures such as average sentence length.

Furthermore, extending the analysis to include corpora from additional languages would of course enable cross-linguistic comparisons, facilitating the assessment of whether the observed phenomena are specific to Catalan or indicative of universal principles in language acquisition. Such an approach would also help to separate age-related developmental effects from typological features intrinsic to each languages. The implementation of multilingual analyses would require standardized linguistic resources, including language-specific tokenizers, morphological analyzers, lemmatizers, and lexical databases that annotate words with their respective polysemy counts.

Finally, the quantitative characterization of typical language acquisition processes in children improves the early identification and diagnosis of developmental language disorders and could also guide the design of more effective intervention strategies in both educational and clinical contexts. A comprehensive understanding of standard linguistic development provides a crucial reference point for recognizing and addressing atypical linguistic patterns, which should be explored in future research.

Acknowledgments

This paper is a revised and updated version of part of the research from the first author's Bachelor's Thesis (*Linguistic laws in language acquisition*), supervised by the second and third authors. We would like to thank here, the evaluation committee for their comments on this work, they were very useful.

Funding

A.H-F acknowledges the support received through grant PID2024- 155946NB-I00, funded by the Spanish Ministerio de Ciencia, Innovación y Universidades (MICIU), the Agencia Estatal de Investigación (AEI/10.13039/ 501100011033), and the European Social Fund Plus (ESF+), and also the project PRO2023-S03-HERNANDEZ (Semàntica de les paraules del català) of l'Institut d'Estudis Catalans. M.T.S. acknowledges the support received through grant AGRUPS-2025 from Universitat Politècnica de Catalunya.

References

Ambridge, B., Pine, J. M., Rowland, C. F., Chang, F., Bidgood, A. (2013). The retreat from overgeneralization in child language acquisition: Word learning, morphology, and verb argument structure. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(1), 47–62. https://doi.org/10.1002/wcs.1207

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press. https://books.google.es/books?id=UvWkIg5E4foC

Baayen, R. H., Tweedie, F. J. (1998). Sample-size invariance of LNRE model parameters: Problems and opportunities. *Journal of Quantitative Linguistics*, *5*(3), 145–154. https://doi.org/10.1080/09296179808590121

Baixeries, J., Elvevag, B., Ferrer-i-Cancho, R. (2013). The evolution of the exponent of Zipf's law in language ontogeny. *PLoS ONE 8(3): e53227*. https://doi.org/10.1371/journal.pone.0053227

Barrett, M. (1982). The holophrastic hypothesis: Conceptual and empirical issues. *Cognition*, *11*(1), 47–76. https://doi.org/10.1016/0010-0277(82)90004-X

Bel, A. (2001). Teoria lingüística i adquisició del llenguatge. Anàlisi comparada dels trets morfològics en català i en castellà [PhD thesis]. Departament de Filologia Catalana. Universitat Autònoma de Barcelona [Institut d'Estudis Catalans]. https://doi.org/10.21415/T5Q30M

Bentz, C., Ruzsics, T., Koplenig, A., Samardžić, T. (2016, December). A comparison between morphological complexity measures: typological data vs. language corpora. In D. Brunato, F. Dell'Orletta, G. Venturi, T. François, P. Blache (Eds.), *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)* (pp. 142–153). The COLING 2016 Organizing Committee. https://aclanthology.org/W16-4117/

Bentz, C., Alikaniotis, D., Cysouw, M., Ferrer-i-Cancho, R. (2017). The entropy of words—learnability and expressivity across more than 1000 languages. *Entropy*, *19*(6), 275. https://doi.org/10.3390/e19060275

Bentz, C., Ferrer-i-Cancho, R. (2016). Zipf's law of abbreviation as a language universal. *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. https://doi.org/10.15496/publikation-10057

Casas, B., Hernández-Fernández, A., Català, N., Ferrer-i-Cancho, R., Baixeries, J. (2019). Polysemy and brevity versus frequency in language. *Computer Speech & Language*, *58*, 19–50. https://doi.org/10.1016/j.csl.2019.03.007

Català, N., Baixeries, J., Lacasa, L., Hernández-Fernández, A. (2023). Semanticity, a new concept in quantitative linguistics: An analysis of Catalan. *Qualico 2023, 12th International Quantitative Linguistics Conference. Lausanne, Switzerland, June 28–30*.

Català, N., Baixeries, J., Ferrer-i-Cancho, R., Padró, L., Hernández-Fernández, A. (2021). Zipf's laws of meaning in Catalan. *PLoS ONE*, *16*(12), e0260849. https://doi.org/10.1371/journal.pone.0260849

Català, N., Baixeries, J., Hernández-Fernández, A. (2024). Exploring semanticity for content and function word distinction in Catalan. *Languages*, *9*(5), 179. https://doi.org/10.3390/languages9050179

El-Hashash, E., Hassan, R. (2022). A comparison of the Pearson, Spearman rank and Kendall Tau correlation coefficients using quantitative variables. *Asian Journal of Probability and Statistics 20(3):36-48*. https://doi.org/10.9734/ajpas/2022/v20i3425

Ferrer i Cancho, R. (2005). Zipf's law from a communicative phase transition. *The European Physical Journal B-Condensed Matter and Complex Systems*, *47*(3), 449–457. https://doi.org/10.1140/epjb/e2005-00340-y

Ferrer i Cancho, R., Solé, R. V., Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E*, *69*, 051915. https://doi.org/10.1103/PhysRevE.69.051915

Ferrer-i-Cancho, R., Gómez-Rodríguez, C., Esteban, J., Alemany-Puig, L. (2022). Optimality of syntactic dependency distances. *Physical Review E*, *105*, 014308. https://doi.org/10.1103/PhysRevE.105.014308

Ferrer-i-Cancho, R., Riordan, O., Bollobás, B. (2005). The consequences of Zipf's law for syntax and symbolic reference. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1562), 561–565. https://doi.org/10.1098/rspb.2004.2957

Ferrer-i-Cancho, R., Vitevitch, M. (2018). The origins of Zipf's meaning-frequency law. *Journal of the Association for Information Science and Technology*, *69*(11), 1369–1379. https://doi.org/10.1002/asi.24057

Ferrer-i-Cancho, R. (2015). The placement of the head that minimizes online memory: A complex systems approach. *Language Dynamics and Change*, *5*(1), 114–137. https://doi.org/10.1163/22105832-00501007

Ferrer-i-Cancho, R., Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society B: Biological Sciences*, *268*(1482), 2261–2265. https://doi.org/10.1098/rspb.2001.1800

Gaztambide-Fernández, R., Cairns, K., Kawashima, Y., Menna, L., VanderDussen, E. (2011). Portraiture as pedagogy: Learning research through the exploration of context and methodology. *International Journal of Education & the Arts*, *12*(4), 1–29. http://www.ijea.org/v12n4/v12n4.pdf

Heaps, H. S. (1978). *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Inc.

Herdan, G. (1960). *Type-Token Mathematics: A Textbook of Mathematical Linguistics*. Mouton & Co., s-Gravenhage.

Hernández-Fernández, A., Ferrer-i-Cancho, R. (2019, August). *Lingüística cuantitativa: la estadística de las palabras*. EMSE EDAPP / Prisanoticias.

Hernández-Fernández, A., Garrido, J., Luque, B., Torre, I. G. (2023). Linguistic laws in Catalan. In M. Yamazaki, H. Sanada, R. Köhler, S. Embleton, R. Vulanović, E. Wheeler (Eds.), *Quantitative Approaches to Universality and Individuality in Language* (pp. 49–62). De Gruyter Mouton. https://doi.org/10.1515/9783110763560-005

Hernández-Fernández, A., Torre, I., Garrido, J., Lacasa, L. (2019). Linguistic laws in speech: The case of Catalan and Spanish. *Entropy, 21(12), 1153*. https://doi.org/10.3390/e21121153

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*(3), 88–97. https://www.jstor.org/stable/24940617

Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303

Institut d'Estudis Catalans. (2007). Diccionari de la llengua catalana. Online version. https://dlc.iec.cat/

Jones, E., Oliphant, T., Peterson, P., Et al. (2001). SciPy: Open source scientific tools for Python. http://www.scipy.org/

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences*, *97*(22), 11850–11857. https://doi.org/10.1073/pnas.97.22.11850

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831–843. https://doi.org/10.1038/nrn1533

Lieven, E., Pine, J., Barnes, H. D. (1992). Individual differences in early vocabulary development: Redefining the referential-expressive distinction. *Journal of Child Language*, *19*(2), 287–310. https://doi.org/10.1017/S0305000900011429

Llinàs-Grau, M. (1998). The GRERLI corpus. https://doi.org/10.21415/YME2-PD42

Llinàs-Grau, M. (2000). The Jordina corpus. https://doi.org/10.21415/T52313

Llinàs-Grau, M., Bel, A., Torras, M. C., Capdevila, M., Coll, M., Domínguez, J., Ojea, A., Pladevall, E., Rosselló, J., Tubau, S. (2003). El desarrollo de las categorías gramaticales: Análisis contrastivo de la adquisición lingüística temprana del inglés, castellano y catalán [Research project].

Llinàs-Grau, M., Coll-Alfonso, M. (2001). Telic verbs in early Catalan. *Probus*, *13*(1), 69–79. https://doi.org/10.1515/prbs.13.1.69

MacWhinney, B. (1999). Talkbank. *TalkBank online resource*.

MacWhinney, B. (2000). *The CHILDES project: The database, Vol. 2, 3rd ed.* Lawrence Erlbaum Associates Publishers.

Miller, G. A. (1994). WordNet: A lexical database for English. *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*. https://aclanthology.org/H94-1111/

Milojević, S. (2010). Power law distributions in information science: Making the case for logarithmic binning. *Journal of the American Society for Information Science and Technology*, *61*(12), 2417–2425. https://doi.org/10.1002/asi.21426

Moskowitz, B. (1978). The acquisition of language. *Scientific American*, *239*(5), 92–109. https://www.jstor.org/stable/24955849

Nowak, P., Santolini, M., Singh, C., Siudem, G., Tupikina, L. (2024). Beyond Zipf's law: Exploring the discrete generalized beta distribution in open-source repositories. *Physica A: Statistical Mechanics and its Applications*, *649*, 129927. https://doi.org/10.1016/j.physa.2024.129927

Piantadosi, S. T., Tily, H., Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529. https://doi.org/10.1073/pnas.1012551108

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663–12668. https://doi.org/10.1073/pnas.1419773112

Serra, M., Solé, R. (1986). Language acquisition in Catalan and Spanish children [Universitat de Barcelona and Universitat Autonoma de Barcelona]. https://talkbank.org/childes/access/Biling/Serra.html

Stevens, L. (Ed.). (2020). *Introduction to Psychology & Neuroscience*. Dalhousie University Libraries - Digital Editions. https://digitaleditions.library.dal.ca/intropsychneuro/

Tomasello, M. (2001). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, *11*(1-2), 61–82. https://doi.org/10.1515/cogl.2001.012

Torre, I., Luque, B., Lacasa, L., Kello, C., Hernández-Fernández, A. (2019). On the physical origin of linguistic laws and lognormality in speech. *Royal Society Open Science*, *6*(8), 191023. https://doi.org/10.1098/rsos.191023

Tubella Salinas, M. (2025). Linguistic laws in language acquisition [Bachelor's Thesis]. Bachelor's Degree in Data Science and Engineering. Facultat d'Informàtica de Barcelona. Universitat Politècnica de Catalunya.

Watts, D. J., Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

Zipf, G. K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press. https://doi.org/10.4159/harvard.9780674434929

Zipf, G. K. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin.

Zipf, G. K. (1945). The meaning-frequency relationship of words. *The Journal of General Psychology*, *33*(2), 251–256. https://doi.org/10.1080/00221309.1945.10544509

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press.

A Appendix

A.1 Parameter estimates using different binning methods

Table A.1: Parameter estimates for the meaning distribution and meaning–frequency laws derived from nonlinear function fitting procedures using both no binning and equal size binning.

Binning	Corpus	Bin size	C_1	C_2	γ	δ
No binning	2–3 years	-	19.3741	6.0276	-0.1711	0.1652
	3–5 years	-	21.6303	5.9953	-0.1728	0.1594
	9–10 years	-	16.8492	9.6563	-0.0881	0.0965
	12–13 years	-	13.7048	10.1753	-0.0476	0.0506
	16–18 years	-	15.3196	8.9134	-0.0806	0.0814
Equal size	2–3 years	6	15.2589	6.0151	-0.1872	0.1682
		12	14.2111	6.0162	-0.2031	0.1674
		103	11.7892	5.7581	-0.3456	0.1983
	3–5 years	13	15.7485	5.9922	-0.2040	0.1595
		26	14.6720	5.9817	-0.2246	0.1602
		79	12.8582	5.9775	-0.2710	0.1568
	9–10 years	20	14.0610	9.6786	-0.1179	0.0964
		40	13.3412	9.7611	-0.1292	0.0850
		160	12.9482	9.2910	-0.2823	0.1154
	12–13 years	22	12.4887	10.2097	-0.0680	0.0456
		31	12.4281	10.1949	-0.0757	0.0487
		62	12.0000	10.2616	-0.0826	0.0370
	16–18 years	55	12.1917	8.8895	-0.1196	0.0823
		95	12.3021	8.8110	-0.1653	0.0889
		209	11.8638	8.6676	-0.2755	0.0950