# Glottometrics

59/2025

International Quantitative Linguistics Association

Glottometrics

Vol. 59, 2025

Glottometrics is an open access scientific journal for the quantitative research of language and text published twice a year by International Quantitative Linguistics Association (IQLA). The journal was established in 2001 by a pioneer of Quantitative Linguistics Gabriel Altmann.

Manuscripts can be submitted by email to glottometrics@gmail.com.
Submission guideline is available at https://glottometrics.iqla.org/.

International Quantitative Linguistics Association (IQLA)

Friedmanngasse 50
1160 Vienna
Austria

eISSN 2625-8226

# Contents

# The paradox of SOV: A case for token-based typology

Natalia Levshina[1*]

[1] Radboud University
[*] Corresponding author's email: natalia.levshina@ru.nl

DOI: https://doi.org/10.53482/2025_59_425

**ABSTRACT**

This study addresses a paradox in word order typology. On the one hand, the SOV order has longer dependency distances and therefore higher processing costs compared to verb-medial order. On the other hand, it is the most frequent word order in languages of the world. How come? A study of corpus data annotated with Universal Dependencies provides a simple answer: the costly long distances occur more rarely than one would assume because SOV clauses are infrequent in language use. A quanitative analysis of 150 Universal Dependencies corpora shows that the proportions of verb-final clauses with two overt core arguments are low across languages, including predominantly verb-final languages. Moreover, a series of Bayesian phylogenetic models based on comparable corpora in thirty-two languages show a negative correlation between the proportion of verb-final clauses in a language and the average number of arguments in a clause, while controlling for argument indexing and high- and low-context culture. A closer examination of argument configurations reveals a positive correlation between proportions of verb-final clauses and proportions of subjectless clauses; as for proportions of objectless clauses, the evidence is less clear. The study highlights the importance of the token-based, gradient approach to typology, which gives us insights into what kind of structures language users prefer, and what they avoid.

**Keywords:** dependency distances, word order, communicative efficiency, token-based typology.

## 1 Introduction

Efficient communication means minimizing the costs of language use, while maximizing its benefits (Hawkins 2004; Gibson et al. 2019; Levshina 2022, *inter alia*). One way of being efficient is to produce semantically and/or syntactically related units close to each other because it reduces working memory costs. According to dependency locality theory (Gibson 1998, 2000), if two syntactically linked words are far away from each other, the memory costs will be high because the processor must store too many structures and predictions about the following elements, which presents a challenge for our limited working memory (cf. Yngve 1960). Moreover, the representation of the word that appears first will be hard to retrieve by the time the second word is produced, due to interference from other words and decay of the first word's representation (Futrell et al. 2020). Although minimization of distances has been most actively studied for syntactic heads and dependents, which are usually represented by corpus tokens (Ferrer-i-Cancho 2004; Liu 2008; Futrell et al. 2015), one can also apply similar principles to syntactic constituents (Hawkins 2004) and morphemes (Bybee 1985; cf. Levshina 2022: Ch. 3). A more general principle about minimization of distances between pairs of any strongly associated elements (that is, not necessarily syntactic heads and dependents), is called information locality (Futrell and Levy 2017). All these principles lead to maximization of accessibility in communication (Levshina 2022: Ch. 3), which, in its turn, is an adaptation to the "Now-or-Never" bottleneck, a fundamental constraint on language sustem, which means that the brain must process linguistic input as rapidly as possible (Cristiansen and Charter 2016).

The present paper addresses a paradox in word order typology. It has to do with the SOV order, which has longer sum and mean dependencies compared to verb-medial transitive clauses. Dependency distances are minimized when the head verb is placed at the centre of a clause (Gildea and Temperley 2007), not at the end. Compare an SVO sentence *Linguists love languages* with an imaginary SOV sentence *Linguists languages love*. As shown in Figure 1, the sum dependency distances in the SVO variant are two words. There are two dependencies: between the verb and the subject and between the verb and the object, and both distances are one word. In the SOV variant, the sum distances are three: two words from the verb to the subject, and one word from the verb to the object. The mean dependency distance, which has been proposed as a metric of processing difficulty (Liu 2008), is also longer in the SOV clause (3/2 = 1.5) than in the SVO clause (2/2 = 1). As a result, SOV has higher memory costs.

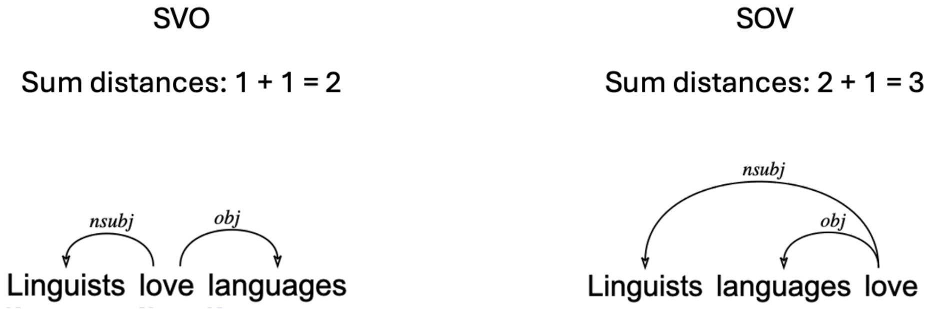

**Figure 1**: Dependency distances in SVO and SOV orders.

Yet, despite these extra costs, the SOV order enjoys massive popularity across human languages (Dryer 2013) as the most widely spread word order. According to Hammarström (2016), SOV is the dominant order in 2,275 languages, which constitute 43.3% of all languages we have information about. If we control for the genealogical relationships between languages, the leadership of SOV becomes even more obvious: it is the most popular order in 239 language families out of 366, or 56.6%. For comparison, the next most popular order, SVO, is dominant in 2,117 languages (40.3%) and represents the majority value in only 55 families (13%). It has also been claimed that SOV was the most likely order of the common ancestor of all existing languages, if one accepts the monogenesis view (Newmeyer 2000; Gell-Mann and Ruhlen 2011; Maurits and Griffiths 2013). Experiments in spontaneous gestural communication also reveal a preference for the agent-patient-action order, which corresponds to SOV (Goldin-Meadow et al. 2008), although not for all types of meanings (Hall et al. 2013; Schouwstra and de Swart 2014).

So why is SOV so widely spread, and even considered cognitively and evolutionarily basic, if it is less efficient than verb-medial order? How to explain this paradox? Some factors could be named. For example, verb-final order may bring processing benefits, due to higher predictability of the verb (Ferrer-i-Cancho 2017). Also, the cognitive costs of switching word order can be too high for language users (Ferrer-i-Cancho and Namboodiripad 2023), which is why SOV languages resist change. Jing et al. (2021) argue that the pressure for dependency distance minimization is weaker in head-final languages, which means that the principle may be less universal than many believe. There is some empirical support for this claim: for example, Futrell et al. (2020) report longer distances in head-final languages for sentences of the same length, whereas Liu (2020) finds that the relative order of adpositional phrases is not systematically constrained by dependency distance minimization in verb-final languages.

However, there may also be a simpler explanation. In a corpus-based study, Ueno and Polinsky (2009) found that spoken Japanese and Turkish (SOV) had fewer arguments than English and Spanish (SVO), due to the fact that the SOV languages contained more one-place predicates, manifesting intransitive bias. This would mean that the processing of verb-final clauses is less costly in reality than one would assume because the longer dependencies associated with two-argument clauses simply do not occur often. At the same time, pro-drop of subject and object, another strategy for reducing the number of overt arguments, was not associated with word order.

We can formulate a research hypothesis then: language users normally avoid using clauses with two overt arguments if the verb is produced at the end, across different languages and types of texts. This expectation, which is explored here using 150 diverse corpora from the Universal Dependencies collection, v2.15 (Zeman et al. 2024), is similar to the maxims of preferred argument structure (Du Bois 1987; Du Bois et al. 2003), such as “Avoid more than one lexical core argument” (Du Bois 1987: 829). However, Du Bois’ theory emphasizes cognitive effort required of the addressee when activating new referents, following Chafe’s (1987) ideas exemplified by the well-known dictum, “one new concept at a time”. In the present study, the expectation that language users avoid SOV (or OSV) clauses with two overt arguments reflects memory constraints shared by both the speaker and the addressee.

Based on Ueno and Polinsky’s (2009) observations, we can also expect that languages with a higher proportion of verb-final clauses will have on average a lower number of overt arguments in a clause. Ueno and Polinsky (2009) investigated only four languages, which is not enough for testing a cross-linguistic correlation. The aim of the present study is to investigate the relationship between verb-final order and the number of core arguments on a larger number of languages. Moreover, the number of overt arguments may also depend on other factors, which should be controlled for in order to avoid confounding effects. One of such factors is the cultural reliance on context, known as Hall's (1976) classification of high- and low-context cultures. Communicators in some cultures may be more used to implicit communication than in others, which enables a lower level of grammatical and lexical specification. Therefore, the need for overt encoding of referents and other information may be lower. Another factor is the presence of argument indexing on the verb (also known as agreement, or cross-referencing). If it is possible to recover an argument from the verb form, it may be efficient to omit the argument as a full form (Haig 2018; Berdicevskis et al. 2020; but see Bickel 2003). Such omission would not only help to save articulatory costs and time, but also the costs of keeping the argument in working memory.

In addition, the number of arguments may be influenced by the degree of formality and modality of communication. We can expect, for example, that the omission rate would be higher in informal and face-to-face communication than in formal and distant communication, because informal style allows for more reduction and the referents can be more accessible from shared linguistic and extralinguistic information, which is part of common ground. This is why it is necessary to control for the register. I use thirty-two corpora of online news from the Leipzig Corpora Collection (Goldhahn et al. 2012), parsed with Universal Dependencies (UD) (Zeman et al. 2024) to test the correlational hypothesis. Because of the phylogenetic and areal dependencies between the languages represented by the corpora, Bayesian phylogenetic models with a two-dimensional Gaussian Process were fitted (Guzmán Naranjo and Becker 2022).

To summarize, the main hypotheses of this study are as follows:

1) Language users disprefer verb-final clauses with two overt core arguments, across all languages and text types;

2) There is a negative correlation between the proportion of verb-final clauses in a language and an average number of overt arguments in a clause, other factors (more exactly, the type of culture and the presence of argument indexing) being controlled for.

These hypotheses represent a perfect case for a token-based and gradient approach to typology (Liu 2010; Levshina 2019; Levshina, Nambodiripad et al. 2023; Gerdes et al. 2021; Yan and Liu 2023). This direction of research aims to describe what people do with language frequently, and what they do only occasionally. It allows us to formulate and test new linguistic universals and explain them from a cognitive and communicative perspective.

The rest of the paper is organized as follows. Section 2 provides details about the corpus data and how the dependencies were counted. Section 3 presents exploratory analyses of the number of SOV clauses in 150 UD corpora. Sections 4 and 5 are dedicated to testing and interpreting the correlation between the proportion of verb-final clauses and the average number of core arguments based on the newspaper data. Section 6 provides a discussion of the results and an outlook.

# 2 Data and method

## 2.1 Corpora

As mentioned in Section 1, two corpora collections were used. For the exploratory analyses of how many SOV clauses actually occur in language use, I used Universal Dependencies corpora, v2.15 (Zeman et al. 2024). The corpora are very heterogeneous, representing many different types of texts and communication modalities. Since the purpose of this exploratory analysis was to establish the upper limit of SOV proportions, the diversity of genres was advantageous, as it enabled more generalizable claims about language use. To make the statistics reliable, I only used the training datasets with the total number of verbal clauses greater than 100. There were in total 150 corpora representing 77 languages.

For testing the correlation between the number of arguments and proportion of verb-final clauses, I selected news corpora from the Leipzig Corpora Collection (Goldhahn et al. 2012) in thirty-two languages, for which Universal Dependencies annotation tools (UDPipe in R package udpipe, Wijffels 2020) were available. As explained above, the main reason for the use of newspaper texts was the fact that argument omission rates can vary by register and text type. Controlling for register was therefore essential to avoid confounding the results. Although the Universal Dependencies corpora include different registers and text types, there were not enough comparable corpora within similar registers. Additionally, many of the corpora consist of mixed data, further limiting comparability. Online newspaper articles were selected due to their wide availability across many languages in the Leipzig Corpora Collection. The languages represented different families and genera, according to the World Atlas of Language Structures Online (Dryer and Haspelmath 2013):

- Indo-European: Baltic (Latvian, Lithuanian), Germanic (Danish, Dutch, English, German, Norwegian, Swedish), Greek (Modern Greek), Indic (Hindi), Iranian (Persian), Romance (French, Italian, Portuguese, Romanian, Spanish) and Slavic (Bulgarian, Croatian, Czech, Russian, Slovenian);
- Afro-Asiatic: Semitic (Arabic);
- Altaic: Turkic (Turkish);
- Austro-Asiatic: Vietic (Vietnamese);
- Austronesian: Malayo-Sumbawan (Indonesian);
- Dravidian: Dravidian (Tamil);
- Japanese (Japanese);
- Korean;
- Sino-Tibetan: Chinese (Chinese [traditional]);

- Uralic: Finnic (Estonian, Finnish), Ugric (Hungarian).

Each corpus represented sentences from online news (when available) or newscrawl categories in the Leipzig Corpora Collection. For every language, 200,000 sentences were parsed and analyzed.

## 2.2 Extraction of dependencies

I searched for all main clauses with a verbal head (dependency relation *root* and Universal Part of Speech ‘VERB’). Main clauses were chosen because subordinate clauses often have relative pronouns as arguments, which should be overt for the sentence to be grammatical.

Two versions of identifying the core arguments were used, depending on what kind of overt arguments were allowed: a) non-clausal ones, represented by the UD dependencies *nsubj* and *obj* only, and b) non-clausal ones plus all possible types of finite and non-finite clausal complements, which are represented by the UD dependency relations *csubj*, *ccomp* and *xcomp*. More details and simplified examples are provided in Table 1.

**Table 1:** Two approaches to defining overt core arguments.

| Approach | Possible values | Examples |
|---|---|---|
| Only non-clausal dependencies<br>(nsubj and obj) | 0 | *Will do!* |
| | 1 | ***She** (nsubj) is running. Just do **it** (obj).* |
| | 2 | *The **students** (nsubj) do their **homework** (obj).* |
| Non-clausal and clausal dependencies<br>(nsubj, obj, csubj, ccomp and xcomp) | 0 | *Will do!* |
| | 1 | ***She** (nsubj) is running. What she **said** (csubj) was surprising. Just do **it** (obj). Just try to **do** (xcomp) it. Remember that you **have** responsibilities (ccomp).* |
| | 2 | *The **students** (nsubj) do their **homework** (obj). **I** (nsubj) want to **do** it (xcomp). The **students** (nsubj) know they must **do** their homework (ccomp). That his idea was **flawed** (csubj) surprised **me** (obj).* |

The two approaches were used because none of them was perfect on its own. If we take all arguments – clausal and non-clausal – we have a more comprehensive picture of the processing costs involved in communication. At the same time, a closer look at the data shows that not all clausal structures that look similarly are annotated in the same way. For instance, while a modal verb in English is usually annotated as an auxiliary of the following infinitive, Russian modal verbs are treated as main verbs, and the infinitive is a non-finite clausal complement. There are also discrepancies in the annotation of direct speech and parataxis. Because a full uniform re-annotation of all such cases for thirty-two corpora was not practically feasible, the decision was to combine the two approaches and see if the results converge.

Note that the sentences were analyzed in accordance with the UD annotation, which has some peculiarities that may affect the results. Agents of passive sentences, for example, were not considered core arguments because they are normally tagged as oblique complements with the dependency relation *obl*. Also, the presence or absence of an overt argument depended fully on whether it was analyzed as a separate token or not. Moreover, according to the UD approach, content words are regarded as heads (cf. de Marneffe et al. 2021; see also the UD documentation[1]). This means that in languages like German and Dutch, in which auxiliaries are often separated from lexical verbs by other constituents, we will find more verb-final clauses compared with if we treated auxiliaries as heads. For the purposes of this study, this is a desirable feature. Semantic microroles, which are relevant for the interpretation of arguments, can be assigned based on the information provided in the lexical verb (e.g., the roles of a breaker and a broken object, which are defined by the verb *break*). Therefore, the position of the lexical verb is relevant in the context of this study.

As for word order, it was operationalized as follows. In every main transitive clause with both overt subject and object, the position of the head verb was established based on the token IDs. Next, the proportions of SOV or OSV clauses were computed for every language. Two versions of the variable were computed, again: either with only non-clausal arguments, or with both non-clausal and clausal arguments.

### 2.3 Other factors: argument indexing, culture, genealogy and areal effects

As mentioned in Section 1, other factors that may influence the number of overt arguments were argument indexing on the verb (also known as verb agreement or cross-referencing),  and high- or low-context culture. Inferring argument indexing from a corpus is not a trivial task. Most of the languages in the sample had some form of indexing in the form of person, gender or number markers, which could help to infer at least some information about the main participants. The following languages were coded as having no indexing on the verb predicate that would help to disambiguate between A and P: Chinese, Danish, Indonesian,[2] Japanese, Korean, Norwegian, Swedish and Vietnamese (Siewierska 2013; Skirgård et al. 2023).

As for the role of context in culture, the following languages were coded as representing low-context cultures, which have a stronger preference for explicit communication: Danish, Dutch, English, German, Swedish and Finnish. This distinction is based on the aggregated classifications in Rösch and Segler (1987) and Holtbrügge et al. (2012). The remaining languages were treated as representing high-context cultures, which have a stronger preference for implicit communication. One should mention, however, that many claims about the role of context in a specific culture are still in need of more robust empirical support, and should be taken with a grain of salt (Levshina et al. 2025).

[1] http://universaldependencies.org/docs/u/overview/syntax.html

[2] Indonesian has bound object markers that are sometimes attached to a transitive verb (Sneddon 1996: 163). However, they are analyzed as pronouns and separate tokens in the UD corpora, which means that the verb carries no formal indices.

When performing a quantitative analysis of typological data, it is necessary to control for genealogical dependencies between languages and possible areal effects, which depend on the geographic distances. To account for this, I fitted several mixed Bayesian models with genealogical and geographic dependencies between the languages as random effects, based on the data from Glottolog (Hammarström et al. 2024). In all models, weakly informative Cauchy priors were used for the fixed and random effects. The warm-up period was 2,000 and the final estimation is based on four chains with 18,000 iterations in each. The *adapt_delta* parameter was 0.9999999. All R-hat values were 1.00, which means that the chains mixed and converged well.

The datasets are freely available in the OSF directory: https://osf.io/75wjx/.

## 3 Testing the avoidance of verb-final clauses with two overt core arguments

To test the hypothesis about the avoidance of verb-final clauses with two overt core arguments, we will look at the proportions of SOV/OSV clauses in different languages represented in the UD corpora. To compute the proportions, the following frequencies were obtained for every UD corpus:

(a) all main verbal clauses with subject and object (i.e., SOV or OSV, as explained above);

(b) all main verbal clauses with one subject only and without any objects (i.e., SV or VS);

(c) all main verbal clauses with one object only and without any subjects (i.e., OV or VO);

(d) all main verbal clauses without subject or objects (i.e., V).

The proportion of (a) in the total sum of (a), (b), (c) and (d) was then computed as the SOV/OSV proportion for every corpus. Because some of the languages were represented by more than one corpus, the final proportions for every language were averaged across the corpora.

The boxplots in Figure 2 represent distributions of proportions of full SOV/OSV clauses in 77 languages. The boxplot on the left represents the first approach (non-clausal subject and object only), and the one on the right corresponds to the second approach (both clausal and non-clausal arguments). The distributions show clearly that in none of the corpora or languages is the proportion of two-argument V-final clauses higher than 30% under the first approach and higher than 40% under the second approach. The languages with relatively high values (represented as outliers) are Afrikaans (Germanic), Marathi (Indic) and Telugu (Dravidian) under the first approach. The same languages also have relatively high valus under the second approach, plus Tamil (Dravidian) and Uyghur (Turkic). From this we can conclude that SOV/OSV clauses represent a minority of verbal clauses in all languages, including strongly verb-final ones.

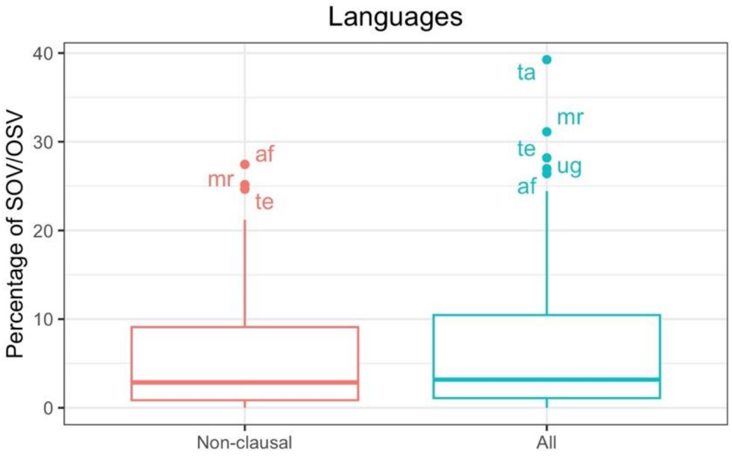

**Figure 2**: Distributions of percentages of SOV/OSV clauses in the languages represented by the UD corpora.

## 4 Testing correlation between verb-finalness and the number of overt arguments

This section tests the second research hypothesis, which predicts negative correlation between the average number of arguments in a clause, and the proportion of verb-final full transitive clauses, based on the news corpora. The relationship between the variables on the basis of non-clausal arguments only is displayed in Figure 3. Overall, the average number of overt core arguments ranges from 0.63 to 1.34. The languages on the right-hand side of the plot (Japanese, Korean, Persian, Tamil, Turkish), which have a strong preference for verb-final clauses, tend to have a low number of overt arguments compared to the languages on the left with a low proportion of verb-final clauses. Verb-medial Danish, English and Vietnamese have more than 1.3 arguments per clause on average – these are the top values. At the same time, some languages on the left-hand side also have a relatively low average number of overt arguments, less than one per clause (Chinese, Hungarian, Latvian, Lithuanian, Slovene). These languages are mostly verb-medial (SVO), but they have a fraction of verb-final clauses, especially Hungarian. Arabic, which has the highest proportion of verb-initial clauses (0.38, or 38%), according to the data, patterns together with the verb-medial languages, having a relatively high number of overt arguments.

The contrast between the strictly SOV and other languages becomes even more pronounced if we consider all possible arguments – clausal and non-clausal. These data are displayed in Figure 4. The average number of arguments is now higher, as one would expect, because clauses are counted, as well. Note that Persian has moved to the left-hand side, due to the high proportion of clausal arguments that follow the verb.

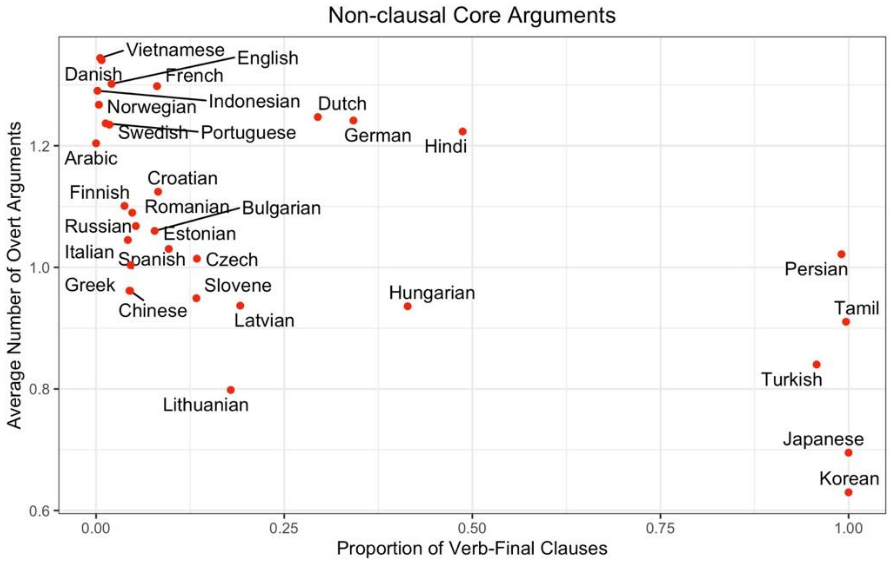

**Figure 3**: Relationship between the average number of arguments and the proportion of verb-final clauses. Non-clausal arguments only.

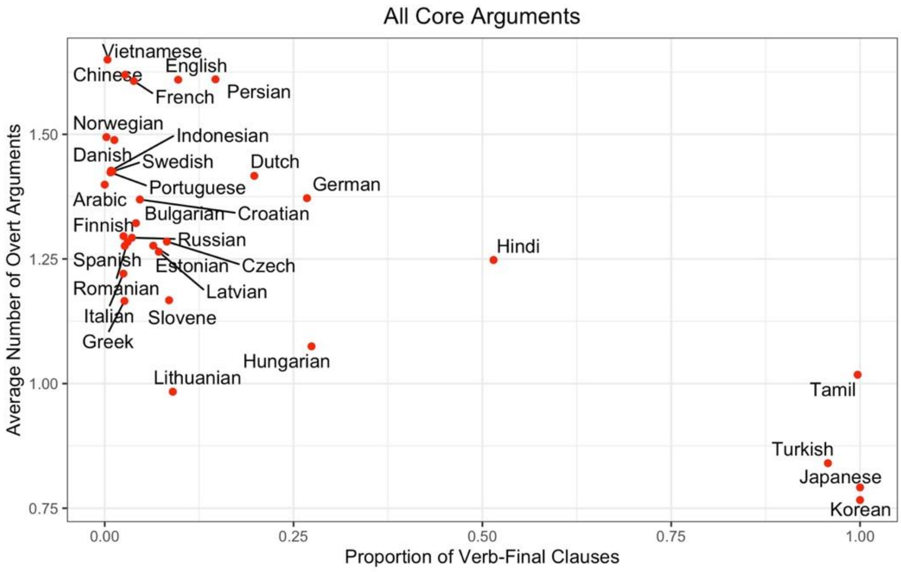

**Figure 4**: Relationship between the average number of arguments and the proportion of verb-final clauses. Non-clausal and clausal arguments.

To test the hypothesis more systematically, Bayesian phylogenetic regression models with a Gaussian process based on the genealogical and geographical information were fitted. The response variable was the average number of arguments. The predictors were the proportion of verb-final clauses, as well as the presence of verb indexing on verbs and the culture.

Table 2 displays the coefficients of the fixed effects. The intercept is 1.15, which represents the average number of non-clausal arguments for a language without any verb-final clauses, high context and no indexing (controlling for the random effects) – a language similar to Vietnamese. The intercept for all kinds of arguments, clausal and non-clausal, is higher (1.46), as one can infer from the y-axis values in Figure 4.

The effect of verb-finalness is negative in both cases. When a language has exclusively verb-final full transitive clauses, all its verbal clauses will have on average 0.29 non-clausal arguments less, and 0.58 arguments of any type less, compared to a language without any verb-final full transitive clauses, the other variables controlled for.  The effect is clearly supported by the data: The 95% credible interval does not include zero, and the posterior probability of a negative effect, which is computed as the proportion of the posterior distribution with a negative sign (Makowski et al. 2019), is nearly 100%. Therefore, the hypothesis about the negative correlation between verb-finalness and the average number of arguments is borne out.

As for the culture, languages from low-context cultures seem to have more overt arguments than languages from high-context cultures, but this effect is sufficiently credible when we count non-clausal arguments only. The probability of a positive effect based on the posterior distribution is then 97.5%, and the 95% credible interval includes mostly positive values. However, the effect is weaker and less credible if we count all types of arguments. The probability of a positive effect decreases to 86.6%.

The effect of argument indexing is not supported by the data. For both approaches, the 95% Credible Interval includes 0, ranging between -0.21 and 0.10, and the posterior probability of a negative effect is only 76.4% for non-clausal arguments and even less, 70.3%, for all arguments. This means a lack of evidence for this effect[3].

**Table 2:** Table of coefficients of the Bayesian phylogenetic regression models. Without parentheses: non-clausal arguments only. In parentheses: clausal and non-clausal arguments.

| Regression term | Posterior mean coefficient | Lower boundary of 95% Credible Interval | Upper boundary of 95% Credible Interval | P of an effect in given direction |
|---|---|---|---|---|
| Intercept | 1.15 (1.46) | 0.74 (1.19) | 1.42 (1.68) | 99.4% (99.9%) |
| Proportion of verb-final clauses | -0.29 (-0.58) | -0.45 (-0.79) | -0.14 (-0.38) | 99.9% (≈100%) |
| Culture = Low-context | 0.13 (0.09) | 0.00 (-0.07) | 0.26 (0.24) | 97.5% (86.6%) |
| Indexing = Yes | -0.06 (-0.04) | -0.21 (-0.20) | 0.10 (0.13) | 76.4% (70.3%) |

Note that the role of verb-initial clauses for processing is not completely clear. On the one hand, the sum dependencies should be as long in verb-initial clauses as in verb-final clauses. On the other hand, the assignment of thematic roles to the arguments in verb-initial clauses can happen very early because

[3] This measure often concentrates around 70% under the null hypothesis of no effect (Kelter 2020).

the verb is already there, which may facilitate processing. An important question is whether the verb-final word order only, or verb-final AND verb-initial orders are difficult for processing. We can answer this question by comparing the predictive power of the models with different sets of independent variables. The models with proportions of verb-final clauses, which are reported in above, had the following performance: the Bayesian $R^2$ of the model with non-clausal arguments only was 0.85, with the 95% credible interval [0.63, 0.99], whereas the model with all arguments had 0.82 [0.63, 0.98]. I also fit models with proportions of verb-medial clauses (thus treating verb-initial and verb-final clauses together), which had somewhat weaker performance: the Bayesian $R^2$ was 0.79 [0.55, 0.98] for the model with non-clausal arguments only, and 0.80 [0.58, 0.99] for the model with all arguments. Although the credible intervals are very wide, this suggests that the verb-final order has a stronger correlation with the number of arguments and is therefore more relevant for processing effort. Model comparisons using the Leave-One-Out and Watanabe Information Criteria also suggest that the models with the verb-final proportions are slightly better than the models with the verb-medial proportions, although the difference is very small. In any event, there is no evidence of the models with verb-medial proportions outperforming the models with verb-final proportions only. Note, however, that the sample did not contain predominantly verb-initial languages, and the proportions of verb-initial clauses were generally low (with the exception of Arabic), which means that no final conclusion can be drawn at the moment about the effects of verb-initial order.

To summarize, the analyses presented in this section provide support for the cross-linguistic correlation: the proportion of verb-final clauses is negatively correlated with the average number of arguments. More exactly, verb-final languages tend to have on average one overt core argument per clause, whereas the other languages display more variation. This means that users of SOV languages experience long dependencies between the verb and core arguments only rarely, which saves their memory costs. But which arguments are usually omitted? This question is addressed in next section.

## 5 How to keep the number of arguments low

### 5.1 Clauses without subject

This section digs deeper into the data from the newspaper corpora, with the aim of understanding better the strategies that help language users to keep the number of overt arguments low in verb-final languages. We begin with the proportions of clauses without subject. Is there a correlation between verb-finalness and preference for subjectless clauses? Figures 5 and 6 show that there is a positive correlation between proportions of verb-final full transitive clauses and proportions of subjectless clauses, both for non-clausal arguments and for all kinds of arguments.

To test these correlations, Bayesian phylogenetic beta regressions were fitted. The response variable was the proportion of subjectless verbal clauses, and the predictors were the proportion of the verb-final full transitive clauses, the culture type and the presence or absence of argument indexing in a language. The beta family was chosen because the response variable is a proportion, which can only be between 0 and 1. The coefficients are displayed in Table 3. According to the model based on the data with non-clausal arguments only, the log-odds is 1.06, which corresponds to the factor of 2.89. The 95% credible interval [0.26, 1.83] does not contain zero, and the probability of a positive effect is 99.4%. In addition, low-context culture has a negative effect with a probability of almost 95%. As for agreement, it does not play any important role, again.

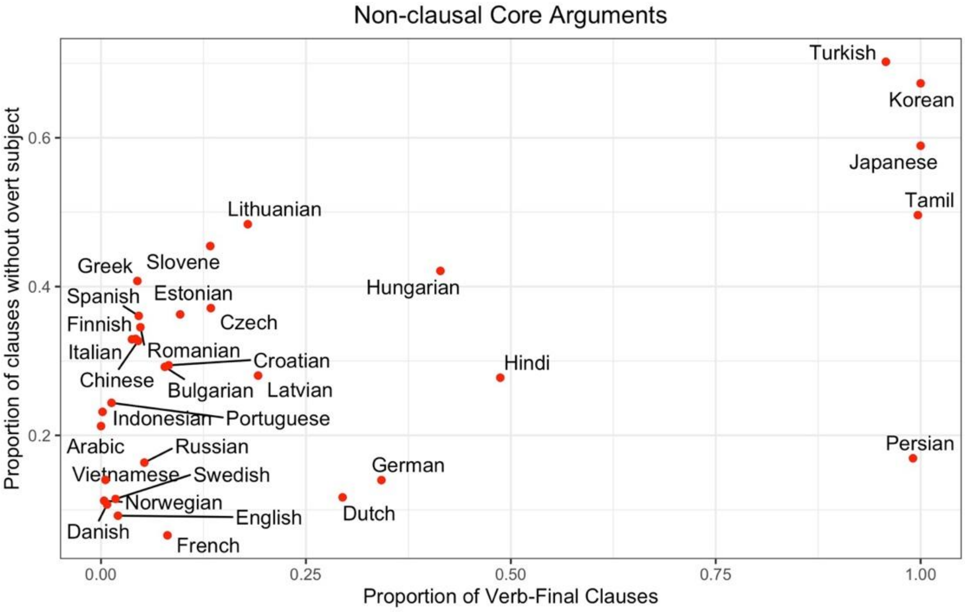

**Figure 5**: Relationship between the proportion of clauses without overt subject and the proportion of verb-final clauses. Non-clausal arguments only.

If we consider all types of arguments, clausal and non-clausal, the effect of verb-finalness is again positive and credible. It is even stronger than in the previous model: the log-odds is 1.59 (factor of 4.9) with the 95% credible interval [0.83, 2.35]. The probability of a positive effect is almost 100%. There is also a credible negative effect of low-context culture with a probability of 97%, somewhat higher than in the previous model. As in the previous case, indexing does not play any important role.

To summarize, we observe a credible positive correlation between the proportion of verb-final full transitive clauses and the proportion of clauses without overt subject. This suggests that subject omission (e.g., due to high accessibility) can be a strategy for saving processing costs when the verb comes at the end of a clause. Since the verb-final languages in the sample are predominantly SOV languages, omission of a clause-initial subject in such cases provides an substantial reduction of costs in terms of both sum and mean dependency distances.

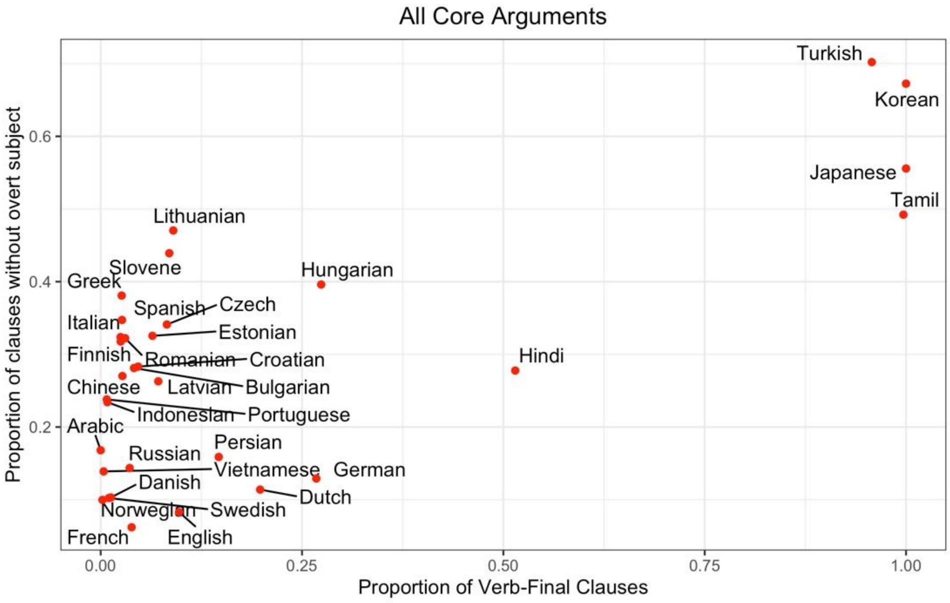

**Figure 6**: Relationship between the proportion of clauses without overt subject and the proportion of verb-final clauses. Clausal and non-clausal arguments.

**Table 3**: Coefficients of the Bayesian phylogenetic beta regression models with proportion of clauses without overt subject as the response variable. Units: log-odds. Without parentheses: non-clausal arguments only. In parentheses: clausal and non-clausal arguments.

| Regression term | Posterior mean coefficient | Lower boundary of 95% Credible Interval | Upper boundary of 95% Credible Interval | P of an effect in given direction |
|---|---|---|---|---|
| Intercept | -1.17 (-1.41) | -1.97 (-2.13) | -0.38 (-0.73) | 99.2% (99.8%) |
| Proportion of verb-final clauses | 1.06 (1.59) | 0.26 (0.83) | 1.83 (2.35) | 99.4% (≈100%) |
| Culture = Low-context | -0.55 (-0.60) | -1.25 (-1.23) | 0.13 (0.03) | 94.5% (97%) |
| Indexing = Yes | 0.15 (0.22) | -0.56 (-0.40) | 0.92 (0.89) | 66.1% (76.1%) |

## 5.2 Clauses without object

This subsection zooms in on the proportions of clauses without overt object. Figure 7 displays the proportion of verb-final full transitive clauses against the proportion of clauses without an overt non-clausal object. The picture is not very clear. As one can see in Table 4, the Bayesian regression model reveals only a weak positive effect of word order: 0.26 in log-odds (or the factor of 1.3). However, the 95% credible interval from -0.2 to 0.75 includes zero, and the posterior probability of a positive effect is 0.87, which does not represent strong evidence.

However, if we include all arguments, clausal and non-clausal, we see an expected credible effect with 0.91 log-odds (the factor of 2.5), a very high probability based on the posterior distribution (99.7%) and a 95% credible interval that does not include zero. Figure 8 illustrates the correlation. The strongly verb-final languages have a high proportion of objectless clauses. In neither model, the type of culture or indexing have any substantial effect.

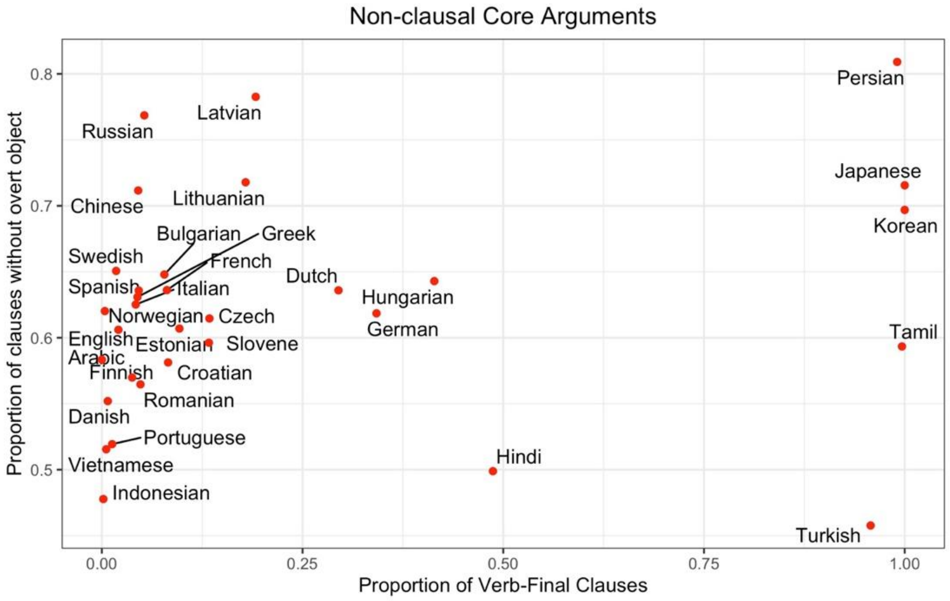

**Figure 7**: Relationship between the proportion of clauses without overt object and the proportion of verb-final clauses. Non-clausal arguments only.

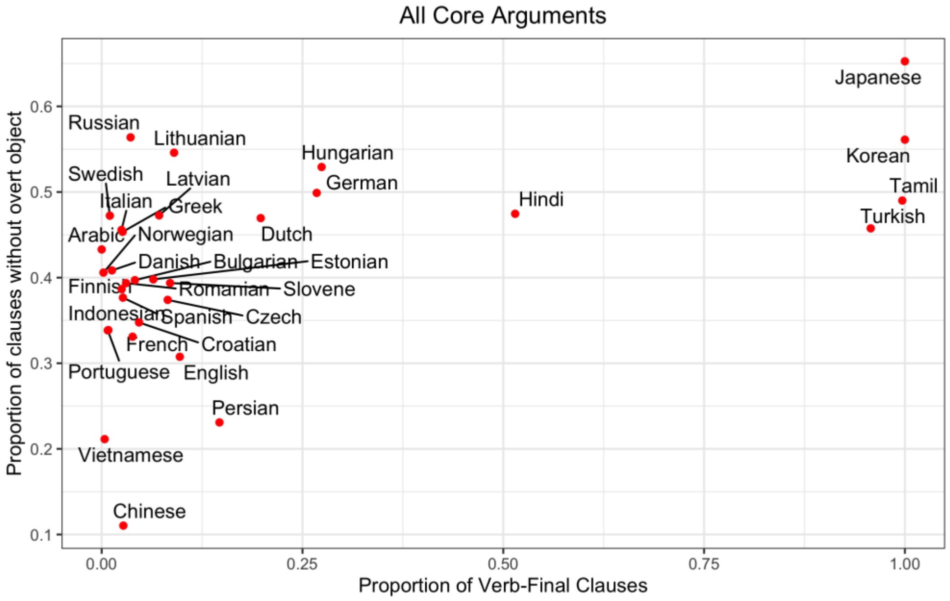

**Figure 8**: Relationship between the proportion of clauses without overt object and the proportion of verb-final clauses. Clausal and non-clausal arguments only.

To conclude, there is no clear support for a correlation between the proportion of verb-final full transitive clauses and the proportion of clauses without overt object. Only when we include clausal arguments can we see a credible positive correlation. Also, the effect sizes are smaller than in the case of subjectless clauses. This may be due to the mixture of two different strategies – object pro-drop and intransitivization (e.g., due to passivization, reflexivization and other types of A-demotion), which may obscure the patterns. For example, a preliminary analysis of the Russian data reveals an abundance of intransitive clauses with reflexive constructions, in which the A-argument is demoted.

**Table 4**: Table of coefficients of the Bayesian phylogenetic beta regression models with proportion of clauses without overt objects as the response variable. Units: log-odds. Without parentheses: non-clausal arguments only. In parentheses: clausal and non-clausal arguments.

| Regression term | Posterior mean coefficient | Lower boundary of 95% Credible Interval | Upper boundary of 95% Credible Interval | P of an effect in given direction |
|---|---|---|---|---|
| Intercept | 0.35 (-0.79) | -0.52 (-1.33) | 1.05 (-0.17) | 90.5% (98.7%) |
| Proportion of verb-final clauses | 0.26 (0.91) | -0.20 (0.29) | 0.75 (1.53) | 86.9% (99.7%) |
| Culture = Low-context | -0.09 (0.13) | -0.47 (-0.31) | 0.30 (0.55) | 68.4% (72.5%) |
| Indexing = Yes | 0.03 (0.15) | -0.38 (-0.35) | 0.46 (0.63) | 54.8% (73.5%) |

This section has demonstrated that the main strategy for reduction of memory costs in verb-final languages is the use of transitive clauses without subjects. The most likely scenario in this case is omission of subjects due to their high accessibility. As for the expression of objects, the differences between verb-final and other languages are less clear. This is not surprising, as objects usually represent new information and cannot be easily omitted.

# 6 Discussion

The goal of this study was to explain the paradoxical popularity of SOV order, given the fact that verb-final order means greater processing costs due to longer sum dependency distances, compared to verb-medial orders. The main expectation was that actual SOV clauses with two overt arguments are not frequently used and therefore cause no significant processing costs. The analysis of the 150 UD corpora supports this hypothesis: verb-final clauses with two overt arguments constitute less than 30% of all verbal main clauses if we count only non-clausal arguments, and less than 40% if we count both non-clausal and clausal ones. In the main bulk of languages, however, these numbers are much lower. Of course, we should keep in mind that the UD corpora are biased towards European languages, which tend to have SVO as the dominant word order. Still, verb-final clauses with long dependencies due to overt subject and object are clearly in the minority in all the languages.

The quantitative analyses of news corpora in thirty-two languages also reveal that languages with a high proportion of verb-final clauses (SOV or OSV) have on average a smaller number of overt core arguments (subject and object) in a verbal main clause. The negative correlation between the proportion of verb-final clauses and the average number of overt arguments, tested in a series of Bayesian spatiophylogenetic regression models, is stable and independent from the type of core arguments – only non-clausal (nominal and pronominal) or both clausal and non-clausal. The correlation holds in the presence of other covariates – high- or low-context culture, and argument indexing on the verb. If a language is spoken in a low-context culture, which relies more on explicit communication, it also seems to have more overt arguments, although the effect is weaker if one counts all types of core arguments, clausal and non-clausal. There was no convincing effect of argument indexing on the number of overt arguments. This may be explained by the scarceness of languages with object indexing, which may be more likely to be in a trade-off relationship with overt objects (Haig 2018), in the sample.

At the same time, we see from the plots that languages with a low proportion of verb-final clauses have substantial variation in the number of core arguments. It seems that the relationship is implicational, rather than correlational: if the verb comes late in the clause, relatively few overt core arguments are expected, but if the verb is in the middle, as in most languages in our sample, the language can have many or few core arguments.

Zooming in on the specific arguments, there is a positive correlation between verb-finalness and the proportion of clauses without subject in a language. The expression of subject also seems to depend on the type of culture: there are fewer subjectless clauses in low-context cultures, which sounds plausible. As for objectless clauses, the pattern is less clear: The correlation with word order is observed only when we count both clausal and non-clausal core arguments. When analyzing large corpora, it is impossible to establish whether an object is absent because the predicate is one-place, or because of pro-drop. A more detailed analysis with manual annotation of individual clauses would be necessary for a precise identification of the strategies that help language users save processing effort.

Therefore, we find evidence of the pressure for maximization of accessibility, or rather, minimization of inaccessibility associated with longer dependencies. A use of an SV or OV clause instead of SVO means that both sum dependency distances and mean dependency distances are lower. Also, a lower number of arguments can also reduce the total costs of language use, including articulation, processing and time costs. The question arises then, why use two-argument clauses if one can do very well with only one core argument? As discussed above, one potential booster for the number of arguments may be a low-context culture. Another possible explanation may have to do with the fact that two-argument clauses in my dataset are mostly observed in languages with loose associations between the grammatical roles and the semantics of the arguments (Hawkins 1986; Levshina 2021), such as English and Indonesian, which may allow for greater semantic flexibility of a transitive clause.

The results of this study have implications not only for the study of communicative efficiency, but also for typology. There is a certain irony in the fact that the SOV order, which is claimed to be the most frequent in human languages, is relatively rare in actual language use. Given the low number of arguments, especially in verb-final languages, one may wonder if it makes sense to use the six-way typology of word order with SOV, SVO, OSV, and so on, if only one core argument is typically overt in verb-final languages and some verb-medial ones. Although this problem was made clear already by Dryer (1997), who discussed the rarity of transitive clauses with nominal subject and object, alongside with other issues, the six-way typology remains very popular nowadays. Even if one includes pronominal and clausal arguments, the six-way typology does not do justice to most languages, failing verb-final ones particularly badly, as one can see from the results of this study. The recent advances in token-based typology will hopefully speed up the movement towards comparative concepts that are closer to actual language use.

The results of this study may be relevant for explaining variation in referential density described by Bickel (2003) and Stoll and Bickel (2009). Referential density is the number of overt lexical/NP arguments in comparison with the total number of possible arguments in a clause. The present study shows that the number of overtly expressed referents may depend on word order, *pace* Bickel (2003). We can also add another maxim of preferred argument structure similar to the ones formulated by Du Bois (1987): "Avoid more than one overt core argument if the verb comes at the end".

There remain many open questions to be addressed in future research. First, the results of the correlational analyses, which are based on the formal register of online news, should be corroborated on data from other registers and text types, most importantly, informal spontaneous conversations. Finding a sufficient number of comparable UD-annotated spoken corpora is not easy, however. More languages from different parts of the world should be tested, especially languages with verb-initial clauses, as well as the dependency relations beyond subject and object. Crucially, we should understand better how the processing pressures affect the language user's choices at the level of a clause in contextualized language use. Finally, it is also important to investigate the interaction between memory constraints and other cognitive and communicative pressures, such as the need for disambugation between core arguments. It has been shown that verb-finalness is correlated with nominal case marking and constraints on argument semantics (Greenberg 1966; Hawkins 1986, 2004; Dryer 2002; Sinnemäki 2010; Levshina 2021). How different pressures interact in language use, making it communicatively efficient and adjusting it to the Now-or-Never bottleneck, remains an open question.

In conclusion, this study highlights the importance of examining actual use of linguistic patterns in communication, which is the focus of the token-based, gradient approach to typology. It demonstrates how this approach can help us solve paradoxes and identify problematic categories in language comparison.

## References

**Ariel, M.** (1990). *Accessing Noun-Phrase Antecedents*. London: Routledge.

**Berdicevskis, A., Schmidtke-Bode, K., Seržant, I.** (2020). Subjects tend to be coded only once: Corpus-based and grammar-based evidence for an efficiency-driven trade-off. In: *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pp. 79–92. Düsseldorf: ACL. https://doi.org/10.18653/v1/2020.tlt-1.8

**Bickel, B.** (2003). Referential density in discourse and syntactic typology. *Language*, 79(4), pp. 708–736. https://doi.org/10.1353/lan.2003.0205

**Bybee, J. L.** (1985). *Morphology: A Study of the Relation Between Meaning and Form*. Amsterdam: John Benjamins.

**Chafe, W.** (1987). Cognitive constraints on information flow. In: Tomlin, R.S. (ed.), *Coherence and Grounding in Discourse: Outcome of a Symposium, Eugene, Oregon, June 1984*, pp. 21–51. Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.11.03cha

**Christiansen, M.H., Chater, N.** (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39, e62. https://doi.org/10.1017/S0140525X1500031X

**de Marneffe, M.C., Manning, C.D., Nivre, J., Zeman, D.** (2021). Universal Dependencies. *Computational Linguistics*, 47(2), pp. 255–308. https://doi.org/10.1162/coli_a_00402

**Dryer, M. S.** (1997). On the six-way word order typology. *Studies in Language*, 21(2), pp. 69–103. https://doi.org/10.1075/sl.21.1.04dry

**Dryer, M.S.** (2002). Case distinctions, rich verb agreement, and word order type (Comment on Hawkins' paper). *Theoretical Linguistics*, 28, pp. 151–157. https://doi.org/10.1515/thli.2002.28.2.151

**Dryer, M. S.** (2013). Order of Subject, Object and Verb. In Dryer, M. S. & Haspelmath, M. (Eds.), *WALS Online* (v2020.4) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13950591. Retrieved from http://wals.info/chapter/81.

**Dryer, M. S., Haspelmath, M.** (Eds.) (2013). WALS Online (v2020.4) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13950591. Retrieved from https://wals.info.

**Du Bois, J.** (1987). The discourse basis of ergativity. *Language*, 64, pp. 805–855.

**Du Bois, J., Kumpf, L.E., Ashby, W.J.** (Eds.). (2003). *Preferred Argument Structure: Grammar as architecture for function*. Amsterdam: John Benjamins.

**Ferrer-i-Cancho, R.** (2004). Euclidean distance between syntactically linked words. *Physical Review E*, 70(5), 056135. https://doi.org/10.1103/PhysRevE.70.056135

**Ferrer-i-Cancho, R.** (2017). The placement of the head that maximizes predictability. An Information Theoretic Approach. *Glottometrics*, 39, pp. 38–71.

**Ferrer-i-Cancho, R., Namboodiripad, S.** (2023). Swap distance minimization in SOV languages. Cognitive and mathematical foundations. *Glottometrics*, 55, pp. 59–88. https://doi.org/10.53482/2023_55_412

**Futrell, R., Levy, R.** (2017). Noisy-context surprisal as a human sentence processing cost model. In: Lapata, M., Blunsom, Ph., Koller, A. (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 688–698. Valencia: EACL. URL https://www.aclweb.org/anthology/E17-1065.

**Futrell, R, Levy, R., Gibson, E.** (2020). Dependency locality as an explanatory principle for word order. *Language*, 96(2), pp. 371–413. https://doi.org/10.1353/lan.2020.0024

**Futrell, R., Mahowald, K., Edward Gibson, E.** (2015). Large-scale evidence of dependency length minimization in 37 languages. *PNAS*, 112(33), pp. 10336–10341. https://doi.org/10.1073/pnas.1502134112

**Gell-Mann, M., Ruhlen, M.** (2011). The origin and evolution of word order. *PNAS*, 108(42), pp. 17290–17295. https://doi.org/10.1073/pnas.1113716108

**Gerdes, K., Kahane, S., Chen, X.** (2021). Typometrics: From implicational to quantitative universals in word order typology. *Glossa: A Journal of General Linguistics*, 6(1), 17. https://doi.org/10.5334/gjgl.764

**Gibson, E.** (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68, pp. 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

**Gibson, E.** (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In: Marantz, A., Miyashita, Y. O'Neil, W. (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium*, pp. 94–126. Cambridge, MA: MIT Press.

**Gibson, E., Futrell, R., Piantadosi, S., Dautriche, I., Mahowald, K., Bergen, L., Levy, R.** (2019). How efficiency shapes human language. *Trends in Cognitive Science*, 23(5), pp. 389–407. https://doi.org/10.1016/j.tics.2019.02.003

**Gildea, D., Temperley, D.** (2007). Optimizing grammars for minimum dependency length. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 184–191. Prague: ACL. URL https://aclanthology.org/P07-1024/

**Goldhahn, D., Eckart, Th., Quasthoff, U.** (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In: Calzolari, N., Choukri, Kh., Declerck, Th. et al. (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pp. 759–765. Istanbul: ELRA.

**Goldin-Meadow, S., Wing Chee, S., Ozyürek, A., Mylander, C.** (2008). The natural order of events: How speakers of different languages represent events nonverbally. *PNAS*, 105(27), pp. 9163–9168. https://doi.org/10.1073/pnas.071006010

**Greenberg, J.H.** (1966). Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, J.H. (ed.), *Universals of Language,* pp. 73–113. Cambridge, MA: MIT Press.

**Guzmán Naranjo, M., Becker, L.** (2022). Statistical bias control in typology. *Linguistic Typology*, 26(3), pp. 605–670. https://doi.org/10.1515/lingty-2021-0002

**Haig, G.** (2018). The grammaticalization of object pronouns: Why differential object indexing is an attractor state. *Linguistics*, 56(4), pp. 781–818. https://doi.org/10.1515/ling-2018-0011

**Hall, E. T.** (1976). *Beyond Culture.* Garden City, NY: Anchor Press/Doubleday.

**Hall, M. L., Mayberry, R.I., Ferreira, V.S.** (2013). Cognitive constraints on constituent order: evidence from elicited pantomime. *Cognition*, 129(1), pp. 1–17. https://doi.org/10.1016/j.cognition.2013.05.004

**Hammarström, H.** (2016). Linguistic diversity and language evolution. *Journal of Language Evolution*, 1(1), pp. 19–29. https://doi.org/10.1093/jole/lzw002

**Hammarström, H., Forkel, R., Haspelmath, M., Bank, S.** (2024). Glottolog 5.1. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.14006617. Retrieved from http://glottolog.org.

**Hawkins, J. A.** (1986). *A Comparative Typology of English and German: Unifying the contrasts*. London: Croom Helm.

**Hawkins, J. A.** (2004). *Efficiency and Complexity in Grammars.* Oxford: Oxford University Press.

**Holtbrügge, D., Weldon, A., Rogers, H.** (2012). *Cultural determinants of email communication styles. International Journal of Cross-Cultural Management*, 13(1), pp. 89–110. https://doi.org/10.1177/147059581245

**Jing, Y., Blasi, D.E., Bickel, B.** (2022). Dependency-length minimization and its limits: A possible role for a probabilistic version of the final-over-final condition. *Language*, 98(3), pp. 397–418. https://doi.org/10.1353/lan.0.0267.

**Kelter, R.** (2020). Analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. *BMC Medical Research Methodology*, 20, 88. https://doi.org/10.1186/s12874-020-00968-2

**Levshina, N.** (2019). Token-based typology and word order entropy. *Linguistic Typology*, 23(3), pp. 533–572. https://doi.org/10.1515/lingty-2019-0025

**Levshina, N.** (2021). Cross-Linguistic trade-offs and causal relationships between cues to grammatical Subject and Object, and the problem of efficiency-related explanations. *Frontiers in Psychology*, 12, 648200. https://10.3389/fpsyg.2021.648200

**Levshina, N.** (2022). *Communicative Efficiency: Language Structure and Use*. Cambridge: Cambridge University Press.

**Levshina, N., Centola, M., Gazzinelli, A., Mello, L., Morhy, R., Narayanaswamy, N.** (2025). *From German to Japanese? A corpus-based analysis of implicitness in events across languages.* Paper presented at the 58th Annual Meeting of The Societas Linguistica Europaea (SLE), 26-29 August 2025, Bordeaux.

**Levshina, N., Nambodiripad, S., Allassonnière-Tang, M., Kramer, M., Talamo, L., Verkerk, A., Wilmoth, S., Garrido Rodriguez, G., Gupton, T.M., Kidd, E., Liu, Z., Naccarato, Ch., Nordlinger, R., Panova, A., Stoynova, N.** (2023). Why we need a gradient approach to word order. *Linguistics*, 61(4), pp. 825–883. https://doi.org/10.1515/ling-2021-0098

**Liu, H.** (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2), pp. 159–191.

**Liu, H.** (2010). Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6), pp. 1567–1578. https://doi.org/10.1016/j.lingua.2009.10.001

**Liu, Z.** (2020). Mixed evidence for cross-linguistic dependency length minimization. *Sprachtypologie und Universalienforschung*, 74, pp. 605–633. https://doi.org/10.1515/stuf-2020-1020

**Maurits, L., Griffiths, Th. L.** (2013).Tracing the roots of syntax with Bayesian phylogenetics. *PNAS*, 111(37), pp. 13576–13581. www.pnas.org/cgi/doi/10.1073/pnas.1319042111

**Makowski, D., Ben-Shachar, M.S., Chen, S.J.A, Lüdecke, D.** (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology*, 10, 2767. https://doi.org/10.3389/fpsyg.2019.02767

**Newmeyer, F. J.** (2000). On the reconstruction of 'proto-world' word order. In: Knight, Ch., Studdert-Kennedy, M., Hurford, J.R. (Eds.), *The Evolutionary Emergency of Language*, pp. 372–388. Cambridge: Cambridge University Press.

**Rösch, M., Segler, K.M**. (1987). Communication with the Japanese. *Management International Review*, 27, pp. 56–67.

**Schouwstra, M., de Swart, H.** (2014). The semantic origins of word order. *Cognition*, 131, pp. 431–436. https://doi.org/10.1016/j.cognition.2014.03.004

**Siewierska, A.** (2013). Verbal Person Marking. In: Dryer, M.S. & Haspelmath, M. (eds.). WALS Online (v2020.4) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13950591 (Available online at http://wals.info/chapter/102, Accessed on 2025-05-21)

**Sinnemäki, K.** (2010). Word order in zero-marking languages. *Studies in Language*, 34, pp. 869–912. https://doi.org/10.1075/sl.34.4.04sin

**Skirgård, H., Haynie, H. J., Hammarström, H., Blasi, D. E., Collins, J., Latarche, J., Lesage, J., Weber, T., Witzlack-Makarevich, A., Dunn, M., Reesink, G., Singer, R., Bowern, C., Epps, P., Hill, J., Vesakoski, O., Abbas, N. K., Ananth, S., Auer, D., Bakker, N. A., Barbos, G., Bolls, A., Borges, R. D., Browen, M., Chevallier, L., Danielsen, S., Dohlen, S., Dorenbusch, L, Dorn, E., Duhamel, M., El Haj Ali, F., Elliott, J., Falcone, G., Fehn, A.-M., Fischer, J., Ghanggo Ate, Y., Gibson, H., Göbel, H.-P., Goodall, J. A., Gruner, V., Harvey, A., Hayes, R., Heer, L., Herrera Miranda, R. E., Hübler, N., Huntington-Rainey, B. H., Inglese, G., Ivani, J. K., Johns, M., Just, E., Kapitonov, I., Kashima, E., Kipf, C., Klingenberg, J. V., König, N., Koti, A., Kowalik, R. G. A., Krasnoukhova, O., Lindsey, K. L., Lindvall, N. L. M., Lorenzen, M., Lutzenberger, H., Marley, A., Martins, T. R. A., Mata German, C., van der Meer, S., Montoya, J., Müller, M., Muradoglu, S., HunterGatherer, Nash, D., Neely, K., Nickel, J., Norvik, M., Olsson, B., Oluoch, C. A., Osgarby, D., Peacock, J., Pearey, I. O.C., Peck, N., Peter, J., Petit, S., Pieper, S., Poblete, M., Prestipino, D., Raabe, L., Raja, A., Reimringer, J., Rey, S.C., Rizaew, J., Ruppert, E., Salmon, K.K., Sammet, J., Schembri, R., Schlabbach, L, Schmidt, F. W. P., Schokkin, D., Siegel, J., Skilton, A., de Sousa, H., Sverredal, K., Valle, D., Vera, J., Voß, J., Wikalier Smith, D., Witte, T., Wu, H., Yam, S., Ye 葉婧婷, J., Yong, M, Yuditha, T., Zariquiey, R., Forkel, R., Evans, N., Levinson, S.C., Haspelmath, M., Greenhill,**

**S.J., Atkinson, Q.D., Gray, R.D.** (2023). Grambank v1.0 (v1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7740140

**Sneddon, J.N.** 1996. *Indonesian: a Comprehensive Grammar.* London and New York: Routledge.

**Stoll, S., Bickel, B.** (2009). How deep are differences in Referential Density? In: Guo, J., Lieven, E., Ervin-Tripp, S., Budwig, N., Özçalikan, S., Nakamura, K. (Eds.), *Crosslinguistic approaches to the psychology of language: research in the tradition of Dan Isaac Slobin*, pp. 543–555. New York: Psychology Press.

**Ueno, M., Polinsky, M.** (2009). Does headedness affect processing? A new look at the VO-OV contrast. *Journal of Linguistics*, 45, pp. 675–710. https://doi.org/10.1017/S0022226709990065

**Wijffels, J.** (2020). udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the UDPipe NLP Toolkit. R package version 0.8.4-1.2020. Retrieved from https://CRAN.R-project.org/package=udpipe.

**Yan, J., Liu, H.** (2023). Basic word order typology revisited: a crosslinguistic quantitative study based on UD and WALS. *Linguistics Vanguard*, 9(1), pp. 73-85. https://doi.org/10.1515/lingvan-2021-0001

**Yngve, V. H.** (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104(5), pp. 444–466.

**Zeman, D., et al.** (2024). Universal Dependencies 2.15, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Retrieved from http://hdl.handle.net/11234/1-5787.

# Word-Frequency Distributions in Chinese- and English- Speaking Older Adults: An Analysis across Languages and Cognitive Statuses

Tongfu Yang[1], Lihe Huang[1], Tsy Yih[1#*]

[1] School of Foreign Studies, Tongji University, Shanghai, China
[#] Tsy Yih is the transliteration of the name of the first author in his mother tongue, Shanghai Wu Chinese. He is also known as Zi Ye in Mandarin pinyin.
[*] Corresponding author's email: yihtsy@outlook.com

DOI: https://doi.org/10.53482/2025_59_426

## ABSTRACT

This study investigates how the word-frequency distributions in spoken language reflect cross-linguistic and cognitive differences in older adults. We analyzed *Cookie Theft* picture descriptions from 96 older adults: 48 Mandarin speakers (24 cognitively impaired and 24 cognitively normal) and 48 English speakers (24 cognitively impaired and 24 cognitively normal) and modeled their word frequency distributions using three functions: Zipf, Zipf-Mandelbrot, and Exponential model. All three models showed excellent goodness of fit at both group and individual levels, indicating that the basic Zipfian structure of lexical distributions is preserved in late life and is not disrupted by mild cognitive impairment. As for the fitting parameters, however, the decay parameter $a$ in the Exponential and Zipf models consistently distinguished Mandarin from English, suggesting that language-specific lexical patterns are robustly encoded in the slope of the distribution but that adding a shift parameter can dampen how clearly $a$ reflects them. By contrast, differences between cognitive groups were weak and inconsistent, implying that parameter $a$ provides only a coarse and context-dependent reflection of cognitive status in short, constrained picture-description tasks.

**Keywords:** Zipf's law, Cognitive impairment, Cross-linguistic analysis.

## 1 Introduction

Human language can be viewed as a self-organizing complex system whose internal structural regularities, or potential regularities, can be effectively revealed by quantitative methods (Ferrer-i-Cancho 2018; Köhler 1987; Oudeyer 2006; Steels 2000). In the context of population aging, this quantitative perspective on language is particularly relevant. Age-related changes in cognition, especially those associated with Alzheimer's disease (AD) and its prodromal stages have been repeatedly shown to manifest systematically in spoken and written language (e.g., Liu et al. 2021; Sand Aronsson et al. 2021;

Orimaye et al. 2017; Gao and He 2025), making language an ecologically valid, multidimensional window into age-related cognitive decline and a promising tool for early detection and longitudinal monitoring. Therefore, investigating the systemic regularities of these linguistic changes holds significant clinical and theoretical value.

Among the various regularities observed in language, the power-law pattern exhibited in word-frequency distributions is one of the most robust phenomena across languages and levels of linguistic structure. It has been documented at multiple levels of analysis, from micro-level measures such as dependency distance and the length of linguistic units (e.g., Liu 2009; Sigurd et al. 2004) to more macro-level domains including language evolution (Bentz 2014) and acquisition (Ellis 2012). This regularity is most commonly captured by the classical Zipf's law (Zipf 1949), which succinctly characterizes the inverse relationship between a word's frequency and its rank.

$$f_r = Cr^{-a}$$

However, because the classical Zipf model has limitations in fitting empirical data, especially at the high-frequency end of the distribution, subsequent work has proposed refined models such as the Zipf-Mandelbrot (ZM) model (Mandelbrot 1966), which introduces an additional parameter to increase the flexibility of the fit.

$$f_r = C(r+b)^{-a}$$

As most previous studies have relied on Zipf and ZM power-law models to fit rank-ordered frequencies, Altmann (2018) introduced an Exponential function as a unified model for diversification phenomena, including rank-frequency distributions. Both families of models aim to capture the same basic empirical regularity, namely that frequency decreases monotonically as rank increases.

$$f_r = 1 + ae^{-br}$$

Taken together, these models provide powerful mathematical tools for quantifying the organizational structure of linguistic systems, yet empirical work has typically selected a single best-fitting model for a given dataset, with relatively little attention to systematic comparisons across models.

In research on older adults, most studies using Zipf's law have focused on specific characteristics of connected speech such as dependency-based measures (Liu et al. 2021; Sand Aronsson et al. 2021; Gao and He 2025) rather than on the rank-frequency distribution of words, even though works on aphasia have demonstrated that quantitative analyses of rank-frequency distributions can distinguish impaired speech from that of healthy controls (Neophytou et al. 2017; van Egmond et al. 2015). To our knowledge, only one study has directly analyzed the rank-frequency distribution of words in the spontaneous speech of older adults (Abe and Otake-Matsuura, 2021).

Furthermore, a critical gap exists at the intersection of typological linguistics and clinical research. Quantitative studies have already established that rank-frequency distributions, differ significantly across typologically distinct languages (Popescu and Altmann 2008; Jiang and Liu 2015; Neophytou et al. 2017). Yet, this comparative cross-linguistic perspective has not been extended to aging populations. It remains unknown whether the linguistic manifestations of cognitive decline are largely universal, or whether they are modulated by the typological properties of a given language, such as Mandarin Chinese versus English.

Against this background, the present study applies three rank-frequency models: Zipf's law, the ZM model, and Altmann's Exponential diversification model to picture-description narratives produced by cognitively normal and cognitively impaired older adults in Mandarin Chinese and English. By jointly modeling word-frequency distributions across models, languages and cognitive groups, we aim to firstly assess how well each model captures the word-frequency structure of older adults' spoken language, secondly examine whether model parameters are sensitive to cross-linguistic differences between Mandarin and English, and lastly test whether these parameters can discriminate between cognitively normal and cognitively impaired speakers within each language. Specifically, we address the following research questions:

**RQ1:** Do word-frequency distributions of older adults' spoken language in Mandarin Chinese and English conform to Zipf's law, the ZM model, and Exponential diversification model?

**RQ2:** Do the parameters of these three models show cross-linguistic differences between Mandarin- and English-speaking older adults, and are these differences dependent on the choice of model?

**RQ3:** Do the model parameters distinguish between cognitively impaired and cognitively normal older adults within each language, and are such group differences robust across the three models?

# 2 Material

## 2.1 Participants

The speech data analyzed in this study were drawn from two existing corpora: the DementiaBank corpus (Becker et al. 1994) and the MCGD (Multimodal Corpus of Gerontic Discourse) (Zhou 2024). On the basis of language (Mandarin Chinese vs. English) and cognitive status (CI: cognitively impaired and CN: cognitively normal), we selected 96 speakers and divided them into four groups of equal size (24 speakers per group). For each participant, we extracted basic demographic information, including age, sex, years of education, and MMSE score. Descriptive statistics for age, sex, education, and MMSE are reported in **Table 1**.

Cognitive status (CI and CN) was determined using the Mini-Mental State Examination (MMSE) with education-specific cutoff points. In line with previous work on mild cognitive impairment, we adopted the following cut-offs for MMSE total scores: MMSE ≤ 19 for illiterate individuals, MMSE ≤ 22 for participants with elementary school education, MMSE ≤ 26 for those with middle school education and above (Jia et al. 2021).

**Table 1:** Group comparisons of demographic variables in the Chinese sample.

| Language | Variable | CN | CI |
|---|---|---|---|
| Chinese | Number of participants | 24 | 24 |
| | Age (mean ± SD) | 64.25 ± 9.40 | 73.33 ± 7.69 |
| | Edu (mean ± SD) | 10.62 ± 5.15 | 9.79 ± 4.51 |
| | MMSE (mean ± SD) | 27.88 ± 1.39 | 20.75 ± 4.58 |
| | Sex (Female / Male) (n (%)) | 13 (54.2%) / 11 (45.8%) | 15 (62.5%) / 9 (37.5%) |
| English | Number of participants | 24 | 24 |
| | Age (mean ± SD) | 63.83 ± 7.70 | 73.67 ± 7.56 |
| | Edu (mean ± SD) | 13.58 ± 3.13 | 12.42 ± 2.34 |
| | MMSE (mean ± SD) | 28.21 ± 0.88 | 20.54 ± 3.98 |
| | Sex (Female / Male) (n (%)) | 15 (62.5%) / 9 (37.5%) | 17 (70.8%) / 7 (29.2%) |

Notes: CI = cognitively impaired; CN = cognitively normal.

## 2.2 Corpus

All speech samples are based on the *Cookie Theft* picture descriptions. In the original corpora, the English data are provided in CHAT (.cha) format in DementiaBank, whereas the Mandarin Chinese data from MCGD are available as plain-text (.txt) transcripts. For the purposes of uniform processing and modelling, all files were converted into plain-text format after preprocessing.

Preprocessing included **speaker selection, data cleaning, and language-specific tokenization and normalization.** First, we retained only the participant's speech (the PAR: tier in the English CHAT files and the corresponding participant lines in the Chinese transcripts) and removed all utterances produced by examiners. During data cleaning, all non-lexical items were removed, including non-lexical filled pauses in the English transcripts (e.g., *&-uh*), as well as coding brackets and meta-linguistic annotations such as [+ gram] and [//]. Regarding lexical filled pauses, we strictly distinguished between interactional signals and production-related hesitation markers. English items such as *yeah* and *oh* and the Mandarin filler *en* "yes" were excluded as they functioned as backchannel responses to the examiner rather than as part of the picture description itself. The Mandarin token *zhège* "this" was retained in all cases. While *zhège* can function as a demonstrative, its non-demonstrative use typically serves as a filled pause indicating lexical retrieval difficulty during the description process. Unlike backchannels, these hesitation markers reflect the speaker's internal cognitive planning directly related to the task. Orthographic reconstruction marks like *spillin(g)* were changed to their full canonical form (e.g., *spilling*). At the same time, we preserved repetitions (e.g., *the the cookie*) and grammatically deviant forms whenever the lexical items were still identifiable, on the grounds that these features reflect genuine production patterns and directly influence the resulting word frequency distributions. Finally, language-specific tokenization and normalization procedures were applied: for Mandarin Chinese, the cleaned transcripts were segmented into word-level tokens using the NLP tool THULAC (Sun et al. 2016); for English, participant speech was extracted from the .cha files and converted to plain text, tokenized based on space and punctuation, converted to lower case, with common contractions (e.g., *he's*) expanded (e.g., *he is*) and the resulting text further processed using the spaCy toolkit (Honnibal et al. 2020) for tokenization and lemmatization. All automated procedures were followed by manual verification. We inspected the transcripts of every participant on a case-by-case basis to ensure the highest level of accuracy.

After cleaning and tokenization, the corpus consisted of 96 *Cookie Theft* picture descriptions, corresponding to 96 processed text files (one per participant). For each file, we computed the total number of tokens (words) and types (distinct word forms) and derived a rank-frequency list in which word forms are ordered from the most to the least frequent. These rank-frequency tables constitute the primary input to the models fitted in the subsequent analyses. At the corpus level, we report the overall token and type counts separately for the Mandarin Chinese and English datasets (see **Table 2**). At the individual level, descriptive statistics regarding token counts, type counts, and Type-Token Ratio (TTR) are presented in

**Table A2**. To illustrate the structure of the derived frequency data, the first 5 tokens of each group are presented in **Table 3** (detailed data are available at https://github.com/toferyoung-wq/detailed-data).

**Table 2:** token, type, and TTR statistics by language and cognitive group.

| Language | Group | Texts | Total tokens | Total types | Tokens (mean ± SD) | Types (mean ± SD) | TTR (mean ± SD) |
|---|---|---|---|---|---|---|---|
| Chinese | CI | 24 | 2447 | 1242 | 101.96 (40.93) | 51.75 (12.62) | 0.54 (0.11) |
| Chinese | CN | 24 | 2508 | 1401 | 104.50 (46.41) | 58.38 (21.20) | 0.59 (0.10) |
| English | CI | 24 | 2139 | 1122 | 89.12 (39.90) | 46.75 (12.52) | 0.56 (0.10) |
| English | CN | 24 | 2470 | 1379 | 102.92 (37.65) | 57.46 (15.10) | 0.58 (0.08) |

Notes: CI = cognitively impaired; CN = cognitively normal.

**Table 3:** The first 5 tokens of each group

| Language | Cognitive Status | Rank | Token | Frequency |
|---|---|---|---|---|
| Chinese | Cognitively Impaired | 1 | *zhège* | 141 |
| | | 2 | *shì* | 131 |
| | | 3 | *le* | 104 |
| | | 4 | *zhè* | 95 |
| | | 5 | *zài* | 72 |
| | Cognitively Normal | 1 | *zhège* | 128 |
| | | 2 | *shì* | 109 |
| | | 3 | *le* | 81 |
| | | 4 | *de* | 80 |
| | | 5 | *zài* | 63 |
| English | Cognitively Impaired | 1 | *the* | 218 |
| | | 2 | *be* | 196 |
| | | 3 | *and* | 123 |
| | | 4 | *she* | 51 |
| | | 5 | *i* | 50 |
| | Cognitively Normal | 1 | *the* | 275 |
| | | 2 | *be* | 233 |
| | | 3 | *and* | 112 |
| | | 4 | *a* | 69 |
| | | 5 | *to* | 54 |

# 3 Methodology

## 3.1 Model Formulas and Fitting

Building on rank-frequency tables, we modeled the resulting word frequency distributions in Python (van Rossum and Drake 2009) at both group level (aggregated by language and cognitive status) and individual level (one model per participant). For each cleaned transcript and each aggregated corpus, the corresponding sequence (rank, frequency) was used as input to a series of non-linear fits implemented with SciPy (Virtanen et al. 2020).

At the modeling level, we considered three mathematical models of the rank-frequency distribution.

**The Zipf model (1)** assumes a power-law decay of frequency with rank, where $C$ is a scale parameter and $a$ is the Zipf exponent controlling the slope.

(1) $$f_r = Cr^{-a}$$

**The ZM model (2)** introduces *a* shift parameter *b*, which captures additional curvature in the high-frequency region.

(2) $$f_r = C(r + b)^{-a}$$

**The Exponential model (3)** includes *a* and *b*. *a* determines the height of the curve at low ranks and *b* controls the rate of exponential decay with rank. In the implementation, the Zipf and ZM models explicitly include the scale parameter *C*, whereas the Exponential model does not contain a separate *C* term[1].

(3) $$f_r = 1 + be^{-ar}$$

The parameter settings in the fitting procedure reflected the model structures. For each text, in the Zipf and ZM models the initial value of *C* was set to the frequency of the most frequent word in that text (i.e., the rank-1 token), and the remaining parameters were initialized at 1.0. Parameter bounds were imposed to avoid degenerate solutions: in the Zipf model, the exponent a was constrained to the interval [0, 10]; in the ZM model, *a* and *b* were constrained to [0, 10] and [0, 100], respectively. For the Exponential model, which does not contain a separate scale parameter *C*, the parameters *a* and *b* were both initialized at 1.0 and constrained to [0, 1000] and [0, 100]. These ranges were chosen to be wide enough to cover all empirically plausible values observed in preliminary fits and in previous work on Zipfian word-frequency distributions, but to exclude clearly implausible or numerically unstable solutions (e.g., negative exponents). Importantly, all best-fitting parameter estimates in our data fell well inside these bounds, indicating that the constraints served only to stabilize the optimization rather than to artificially restrict the models. The maximum number of function evaluations in curve_fit was set to 20,000 to improve convergence for long-tailed distributions.

For each model and each text, the optimization returned a set of best-fitting parameters and the corresponding predicted frequencies. Model performance was summarized using the coefficient of determination $R^2$ (4) which is computed from the squared deviations between the observed frequencies $f_r$ and the model predictions $\hat{f}_r$ relative to the variance of $f_r$.

[1] The model was originally expressed as $f_r = 1 + ae^{-br}$. For comparability with the Zipf and ZM models, we simply swap the positions of *a* and *b*, using *a* denote the decay parameter.

(4) $$R^2 = 1 - \frac{\sum_{r=1}^{R}(f_r - \hat{f}_r)^2}{\sum_{r=1}^{R}(f_r - \bar{f})^2}$$

Thus, for every participant and for each of the three models we obtained a set of fitted parameters (*C*, *a*, *b*) together with an associated $R^2$ value. All individual-level parameter estimates and goodness-of-fit indices were exported as CSV files, along with detailed rank-wise tables that include, for every word, its rank, observed frequency, predicted frequency, and residual. These exported tables form the basis for the subsequent statistical analyses and visualizations.

### 3.2 Statistical Analysis and Visualization

All statistical analyses were carried out in R (R Core Team 2023). We used the packages car (Fox and Weisberg 2019) for analysis of variance, emmeans (Lenth 2020) for estimated marginal means and post-hoc contrasts, and dplyr (Wickham et al. 2023) for data handling. For each model parameter of interest, we analyzed the individual-level estimates obtained from the Python fitting procedure with Language (Chinese vs. English) and Cognitive Status (CI vs. CN) as between-subject factors.

Because the raw parameter distributions were typically right-skewed and strictly positive, we applied a natural-log transformation to the parameter estimates so that the data better met the assumptions of linear modelling and ANOVA. For each parameter, we then analyzed the effects of language (Chinese vs. English) and cognitive status (CI vs. CN) in a two-way framework. Whenever the language × cognitive status interaction was significant, we conducted planned post-hoc comparisons using the emmeans package. Specifically, we focused on four a priori contrasts: CI vs. CN within each language, and Chinese vs. English within each cognitive-status group. Statistical significance for these planned contrasts was determined on the basis of Holm-corrected *p*-values, in order to control the family-wise error rate within this set of comparisons.

The visualization was carried out in Python. Data were processed by Pandas (McKinney 2010) and NumPy (Harris et al. 2020) and visualized by Matplotlib (Hunter 2007) with significance markers manually added based on the results of the statistical tests.

## 4 Results and Discussion

### 4.1 Goodness of fit

The first aim of this study was to examine whether the rank-frequency distributions of older adults' *Cookie Theft* descriptions in English and Mandarin can be captured by three standard diversification models (Exponential, Zipf, ZM). At the group level, the $R^2$ values (see **Figure 1** and **Figure 2**) indicate excellent fits in both languages, with values ranging from 0.924 to 0.995 for Chinese and from 0.859 to 0.955 for English, and no clear differences in goodness of fit between cognitively impaired and cognitively normal speakers within each language. At the individual level (see **Figure 3**), the models also performed well: most participants showed $R^2$ values above 0.800, with only a few exceptions (e.g., one Chinese CN speaker with $R^2$ = 0.769 under the Zipf model), and again no systematic differences in $R^2$ between CI and CN groups within language group for the three models.

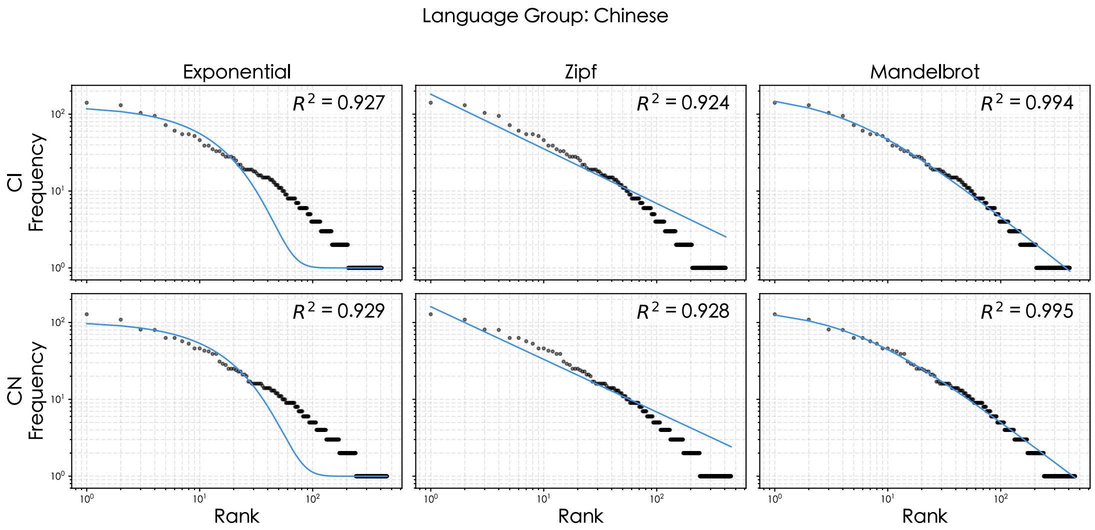

**Figure 1:** Log-log rank-frequency plots for Zipf, ZM and Exponential models of the Chinese group at group level, with $R^2$ values shown for each.

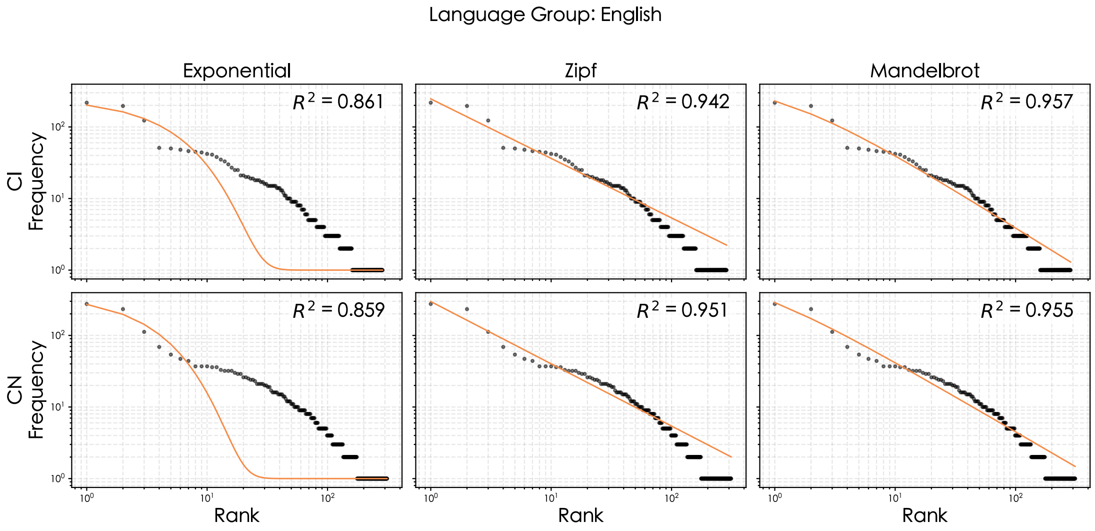

**Figure 2:** Log-log rank-frequency plots for Zipf, ZM and Exponential models of the English group at group level, with $R^2$ values shown for each.

More fine-grained inspection of the group-level fits reveals systematic language differences in how the three models capture the distributions. The Chinese data are better fitted by the Exponential and ZM models than the English data, whereas English achieves slightly higher $R^2$ than Chinese under the basic Zipf model. For the exponential model, this difference appears to be driven by the overall shape of the distributions: in Chinese, the high-frequency head and mid-frequency range closely follow the exponential curve and the low-frequency tail deviates only mildly, whereas in English the mid and low ranks show larger deviations from the Exponential prediction. Comparing the Zipf and ZM fits leads to a similar conclusion. In Chinese, adding the ZM shift term (parameter *b*) yields a marked improvement over the basic Zipf model, with $R^2$ increasing from 0.924 to 0.994 (CI) and from 0.928 to 0.995, which points to a more pronounced high-frequency head in the Chinese distributions. In English, by contrast, the head correction improves the fit only marginally with $R^2$ increasing from 0.942 to 0.957 (CI) and from 0.951 to 0.955 (CN); the English curves tend to show a very steep drop at the top ranks followed by a relative flattening or slight rise in the mid-frequency range.

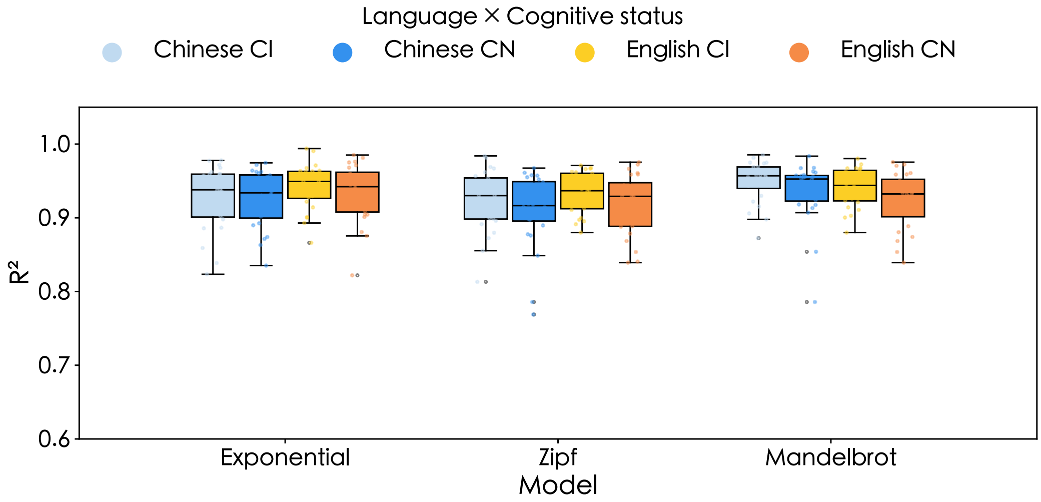

**Figure 3:** Individual-level $R^2$ values for the three models, with model names on the x-axis and colors indicating groups (from left to right: Chinese CI, Chinese CN, English CI, English CN).

Overall, these results suggest that, although Chinese and English differ slightly in the detailed shape of their word-frequency distributions, all three models capture the word-frequency structure of older adults' *Cookie Theft* descriptions well. Cognitive impairment does not appear to substantially alter the overall rank-frequency pattern.

### 4.2 Differences in Model Parameters

To address RQ 2 and 3, we conducted two-way ANOVAs on the log-transformed decay parameter *a* for all three models, with language and cognitive status as between-subject factors. The detailed results of two-way ANOVA are presented in **Table A1.** Significant language × cognitive status interactions were found in Exponential and Zipf, respectively F (1,92) = 6.210, $p$ = 0.015, $\eta p^2$= 0.060; F (1,92) = 7.820, $p$ = 0.006, $\eta p^2$ = 0.080. We therefore focus on the corresponding simple effects based on post-hoc tests (see **Table 4** for details). For the ZM model, however, the log-transformed parameters still exhibited clear violations of the homogeneity-of-variance assumption (Levene's test for *a*: F (3, 92) = 7.310, $p <$ 0.001). Moreover, neither the main effects ($p$ = 0.344 for language and $p$ = 0.151 for cognitive status) nor their interaction (*p* = 0.296) reached statistical significance for the ZM parameters. Therefore, we do not pursue further inferential analyses for this model and treat its results as purely supplementary.

**Table 4:** Simple-effects analyses of parameters *a* (log-transformed) of Exponential and Zipf model across four groups.

| **Models** | **Contrast** | **Estimate (log scale)** | **SE** | **t (92)** | ***p* (Holm)** |
|---|---|---|---|---|---|
| Exponential | Chinese_CI − Chinese_CN | 0.213 | 0.129 | 1.654 | 0.130 |
| | English_CI − English_CN | −0.241 | 0.129 | 1.869 | 0.130 |
| | Chinese_CI − English_CI | −0.392 | 0.129 | −3.038 | **0.009**** |
| | Chinese_CN − English_CN | −0.847 | 0.129 | −6.561 | **< 0.001***** |
| Zipf | Chinese_CI − Chinese_CN | 0.113 | 0.048 | 2.338 | **0.045*** |
| | English_CI − English_CN | −0.078 | 0.048 | −1.616 | 0.110 |
| | Chinese_CI − English_CI | −0.120 | 0.048 | −2.479 | **0.045*** |
| | Chinese_CN − English_CN | −0.311 | 0.048 | −6.433 | **< 0.001***** |

Notes: CI = cognitively impaired; CN = cognitively normal. Estimates are contrasts on the log-transformed decay parameter a (first group minus second group), based on estimated marginal means from two-way ANOVAs with language (Chinese vs. English) and cognitive status (CI vs. CN) as between-subject factors. SE = standard error of the contrast; t(92) = t-statistic with 92 residual degrees of freedom; *p* (Holm) = Holm-adjusted *p*-value for the planned comparisons. * *p* < .05; ** *p* < .01; *** *p* < .001.

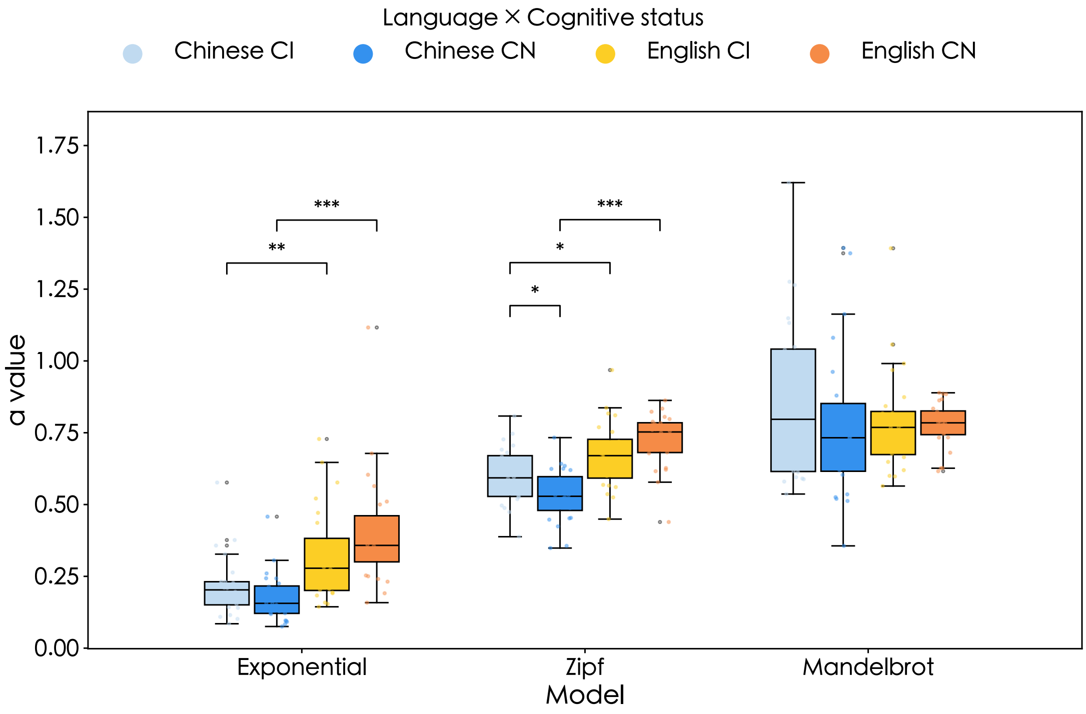

**Figure 4:** Individual-level values of the decay parameter *a* for each model. Asterisks indicate significance (**p* < .05, ***p* < .01, **p* < .001).

### 4.2.1 Changes in Parameter *a* across Languages

In both the Exponential and Zipf models, after controlling for cognitive status, both the boxplots (see **Figure 4**) and the simple-effects analyses (see **Table 4**) indicate that across all comparable contrasts, the English groups exhibited significantly larger decay parameters *a* (Exponential: $p = 0.009$ in CI and $p < 0.001$ in CN; Zipf: $p = 0.045$ in CI and $p < 0.001$), indicating steeper distributional slopes than the Chinese groups, with log-scale differences ranging from 0.120 to 0.847. This discrepancy was particularly pronounced for the Exponential model in the cognitively normal group (Chinese_CN − English_CN = − 0.847).

At first glance, this finding seems to contradict previous work. Earlier studies have reported that morphologically poorer languages show steeper Zipfian slopes: for instance, Bentz et al. (2014) found that Modern English has a steeper slope than Old English, and Neophytou et al. (2017) reported steeper slopes for English than for morphologically richer Greek. On this view, Mandarin Chinese which is typically described as an analytic language with virtually no inflectional morphology would be expected to have a steeper slope than English. Our data point in the opposite direction.

We argue that this apparent contradiction is unlikely to reflect macro-level typological differences, but rather arises from the specific task (*Cookie Theft* picture description) and the language-specific properties of the resulting corpora. Inspection of the data shows that, in the English group, the highest-frequency items are predominantly *be, and* and *the*, whereas in the Chinese group they are mainly *zhègè* "this", *zhè* "this" and *shì* "is". At the task level, the *Cookie Theft* picture invites both object naming and event description. Differences in familiarity with the scene may lead Chinese- and English-speaking older adults to focus on different aspects: English speakers tend to attend more to the actions and frequently use *be + V-ing* constructions to describe ongoing events, which raises the frequency of the lemma *be*, whereas many Chinese speakers, who are less familiar with the depicted setting, place greater emphasis on object identification and often rely on *shì* "be"[2]. At the language level, English speakers also tend to use overt conjunctions such as *and* to maintain discourse coherence, while Mandarin speakers often do not employ an explicit coordinator in analogous contexts. Likewise, the rich article system of English routinely pushes the into the very top ranks, whereas the use of Chinese classifiers such as *yígè* "one" is more optional and more strongly influenced by individual stylistic preferences than by grammatical obligation. Together, these task- and corpus-specific factors lead to extremely high frequencies of *and*, *the* and *be* in the English data, and a much larger gap between the top ranks and the mid ranks, whereas in the Chinese data the contrast between the head and the middle of the distribution is less pronounced (can also be observed at group level, see **Figure 1**, **Figure 2** and **Table 3**). This offers a plausible explanation for why the English distributions in our data appear steeper than the Chinese ones, despite typological expectations based on morphological richness.

Taken together, these results suggest that the decay parameter a can indeed capture specific characteristics of the rank-frequency distribution. First, language-specific properties, especially at the lexical level, are clearly reflected in the shape of the distributions. At the same time, model properties also matter: for example, no significant effects were detected for the ZM model, indicating that introducing additional parameters may increase $R^2$ but at the cost of masking corpus-specific structure. This underscores the value of applying more than one model in future work rather than relying on a single specification. Finally, our conclusions are constrained by the limited size of the corpus and the relatively simple, picture-description task. Studies with larger samples and more diverse, cognitively demanding tasks may yield more robust and generalizable insights into how *a* relates to language use.

### 4.2.2 Changes in Parameter *a* across Cognitive Statuses

Across cognitive-status groups, the associations between the decay parameter *a* and cognition appear relatively weak. Compared with the highly pronounced differences across languages, the differences between CI and CN are much less distinct: simple-effects analyses show that only one significant effect was detected in the Chinese group under the Zipf model ($p$ =0.045, see **Table 4**) and the cross-cognitive trends even go in opposite directions across languages (from CI to CN, *a* generally decreases in Chinese but increases in English, see **Figure 4**).

[2] In Mandarin, *zhègè* "this" does not always function as a genuine demonstrative; it is often used as a filled pause that buffers lexical retrieval difficulty. When *zhè* "this" and *shì* "is" occur together, the phrase typically has a clear demonstrative-copular function which means *this is* or *there is* in English, and in this sense the copula *shì* "is" is the item that more directly reflects Chinese older adults' focus on object naming in the task.

Several factors may underlie this pattern. First, the differences between cognitively impaired and normal group in our corpora are relatively subtle, making them difficult to capture with a single slope parameter. Previous work suggests that *a* has good discriminative power for severe language disorders, but is much less sensitive to finer-grained variation: van Egmond et al. (2015) showed that non-fluent aphasic speech has a reliably steeper Zipfian slope than healthy control speech, and Neophytou et al. (2017) replicated this pattern for both fluent and non-fluent aphasia in English, while finding no systematic differences between the two aphasic subtypes. In addition, Abe and Otake-Matsuura (2021) reported no robust association between Zipf's exponent and cognitive scores in cognitively screened Japanese elders. Second, the behavior of *a* is model-dependent: although the decay parameter plays a comparable conceptual role across the Exponential, Zipf and ZM models, the cross-model discrepancies indicate that each model emphasizes different aspects of the rank-frequency shape. Third, in relatively short texts the parameter *a* is easily distorted by idiosyncratic high-frequency patterns, so the way *a* changes across cognitive-status groups can differ substantially between languages. For example, in the English group, the higher *a* values observed in the CN speakers may stem from their longer descriptions (see **Table 2**, CN has longer TTR than CI) which naturally contain more instances of function words such as *and* and *the*, thereby raising the head of the distribution, whereas the higher *a* in Chinese CI speakers may be driven by frequent use of *zhège* "this" as a filled pause in contexts of word-finding difficulty.

Overall, the parameter *a* appears to distinguish impaired from unimpaired language only when the underlying differences are relatively pronounced. Moreover, given the limited text length and the complexity of the models, our findings suggest that the decay parameter *a* offers only a coarse, context-dependent reflection of cognitive status, and its usefulness as a marker of cognition in short, constrained tasks such as *Cookie Theft* descriptions requires further investigation.

## 5 Conclusion

In this study, we modeled the word-frequency distributions of *Cookie Theft* picture descriptions produced by older Mandarin and English speakers, with and without cognitive impairment, using three models: Zipf, ZM and Exponential model. All three models provided excellent goodness of fit at both the group and individual levels, indicating that Zipfian structure is preserved across languages and cognitive statuses in this task. At the parameter level, however, the models behaved somewhat differently and were only weakly sensitive to subtle contrasts. The decay parameter *a* in the exponential and Zipf models reliably distinguished Mandarin from English, whereas the Zipf-Mandelbrot model did not, suggesting that lexical-level language-typological differences are indeed captured in the shape of the distributions. By contrast, the ability of *a* to differentiate cognitive status was limited: only one significant difference was found for Mandarin under the Zipf model, which is likely due to the relatively mild degree of language impairment in our CI group.

However, this study has several limitations. First, the sample size is relatively small, which limits the statistical power of the analyses and reduces the robustness of group comparisons. With only 24 speakers per group, subtle effects may go undetected and parameter estimates may be more vulnerable to individual variability. Second, the *Cookie Theft* picture-description task is simple and may not elicit representative speech that fully captures the properties of participant's speech production, nor does it strongly amplify differences between cognitively healthy and impaired speakers. Third, cognitive status was dichotomized into CI and CN based solely on MMSE, which is a relatively coarse grouping. Future work should therefore use longer speech samples, adopt more fine-grained cognitive classifications, and employ more complex and varied elicitation tasks in order to better assess the extent to which the decay parameter a can discriminate between intact and impaired language.

## Acknowledgements

This research was supported by the National Social Science Fund of China Grant: [24BYY120] awarded to Lihe Huang.

## References

**Abe, M. S., Otake-Matsuura, M.** (2021). Scaling laws in natural conversations among elderly people. *PLOS ONE*, 16(2), e0246884. https://doi.org/10.1371/journal.pone.0246884.

**Altmann, G.** (2018). *Unified modeling of diversification in language*. Lüdenscheid: RAM-Verlag.

**Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., McGonigle, K. L.** (1994). The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6), pp. 585-594.

**Bentz, C., Kiela, D., Hill, F., Buttery, P.** (2014). Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. *Corpus Linguistics and Linguistic Theory*, 10(2), 2014, pp. 175-211. https://doi.org/10.1515/cllt-2014-0009.

**Ellis, N. C.** (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, 32, pp. 17-44. https://doi.org/10.1017/S0267190512000025.

**Ferrer-i-Cancho, R.** (2018). Optimization models of natural communication. *Journal of Quantitative Linguistics*, 25(3), pp. 207-237. https://doi.org/10.1080/09296174.2017.1366095.

**Fox, J., Weisberg, S.** (2019). *An R companion to applied regression* (3rd ed.). Thousand Oaks, CA: Sage.

**Gao, N., He, Q.** (2025). A corpus-based dependency study of the syntactic complexity in the connected speech of Alzheimer's disease. *Aphasiology*, 39(11), pp. 1456-1479. https://doi.org/10.1080/02687038.2024.2434858.

**Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E.** (2020). Array programming with NumPy. *Nature*, 585, pp. 357-362.

**Hunter, J. D.** (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), pp. 90-95.

**Jia, X., Wang, Z., Huang, F., et al.** (2021). A comparison of the Mini-Mental State Examination (MMSE) with the Montreal Cognitive Assessment (MoCA) for mild cognitive impairment screening in Chinese middle-aged and older population: A cross-sectional study. *BMC Psychiatry*, 21, 485. https://doi.org/10.1186/s12888-021-03495-6.

**Jiang, J., Liu, H.** (2015). The effects of sentence length on dependency distance, dependency direction and the implications—Based on a parallel English Chinese dependency treebank. *Language Sciences*, 50, pp. 93-104. https://doi.org/10.1016/j.langsci.2015.04.002.

**Köhler, R.** (1987). System theoretical linguistics. *Theoretical Linguistics*, 14(2–3), pp. 241-247. https://doi.org/10.1515/thli.1987.14.2-3.241.

**Lanzi, A. M., Saylor, A. K., Fromm, D., Liu, H., MacWhinney, B., Cohen, M.** (2023). DementiaBank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology*, 32(2), pp. 426-438. https://doi.org/10.1044/2022_AJSLP-22-00281.

**Lenth, R. V.** (2020). *emmeans: Estimated marginal means, aka least-squares means* [R package]. Retrieved from https://CRAN.R-project.org/package=emmeans.

**Liu, H.** (2009). Probability distribution of dependencies based on a Chinese dependency treebank. *Journal of Quantitative Linguistics*, 16(3), pp. 256-273. https://doi.org/10.1080/09296170902975742.

**Liu, J., Zhao, J., Bai, X.** (2021). Syntactic impairments of Chinese Alzheimer's disease patients from a language dependency network perspective. *Journal of Quantitative Linguistics*, 28(3), pp. 253-281. https://doi.org/10.1080/09296174.2019.1703485.

**Mandelbrot, B.** (1966). Information theory and psycholinguistics: A theory of word frequencies. In: P. F. Lazarsfield, N. W. Henry (Eds.). *Readings in Mathematical Social Sciences* (pp. 350-368). Cambridge, MA: MIT Press.

**McKinney, W.** (2010). Data structures for statistical computing in Python. In: S. van der Walt, J. Millman (Eds.). *Proceedings of the 9th Python in Science Conference* (pp. 51-56). https://doi.org/10.25080/Majora-92bf1922-00a.

**Neophytou, K., van Egmond, M., Avrutin, S.** (2017). Zipf's law in aphasia across languages: A comparison of English, Hungarian and Greek. *Journal of Quantitative Linguistics*, 24(2-3), pp. 178-196. https://doi.org/10.1080/09296174.2016.1263786.

**Orimaye, S. O., Wong, J. S.-M., Golden, K. J., Wong, C. P., Soyiri, I. N.** (2017). Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics*, 18, 34. https://doi.org/10.1186/s12859-016-1456-0.

**Oudeyer, P.-Y.** (2006). *Self-organization in the evolution of speech*. Oxford: Oxford University Press.

**Popescu, I. I., Altmann, G.** (2008). Hapax legomena and language typology. *Journal of Quantitative Linguistics*, 15(4), pp. 370-378. https://doi.org/10.1080/09296170802326699.

**R Core Team.** (2023). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

**Sand Aronsson, F., Kuhlmann, M., Jelic, V., Östberg, P.** (2021). Is cognitive impairment associated with reduced syntactic complexity in writing? Evidence from automated text analysis. *Aphasiology*, 35(7), pp. 900-913. https://doi.org/10.1080/02687038.2020.1742282.

**Sigurd, B., Eeg-Olofsson, M., Van Weijer, J.** (2004). Word length, sentence length and frequency – Zipf revisited. Studia Linguistica, 58(1), pp. 37-52. https://doi.org/10.1111/j.0039-3193.2004.00109.x.

**Steels, L.** (2000). Language as a Complex Adaptive System. In: Schoenauer, M. et al. (Eds.). *Parallel Problem Solving from Nature PPSN VI. Lecture Notes in Computer Science* (Vol. 1917), pp. 17-26. Berlin, Heidelberg: Springer-Verlag. https://doi.org/10.1007/3-540-45356-3_2.

**van Egmond, M., van Ewijk, L., Avrutin, S.** (2015). Zipf's law in non-fluent aphasia. Journal of Quantitative Linguistics, 22(3), pp. 233-249. https://doi.org/10.1080/09296174.2015.1037158.

**van Rossum, G., Drake, F. L.** (2009). *Python 3 reference manual*. Scotts Valley, CA: CreateSpace Independent Publishing Platform.

**Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … van Mulbregt, P.** (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17, pp. 261-272.

**Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D.** (2023). *dplyr: A grammar of data manipulation* [R package]. Retrieved from https://dplyr.tidyverse.org.

**Zhou, D.** (2024). Multimodal corpus of geronto discourse: Construction and reflection. *Linguistic Research*, 36, pp. 20-34.

**Zipf, G. K.** (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.

# Appendix

**Table A1:** Two-way ANOVA results for the log-transformed decay parameter *a* across models

| **Model** | **Effect** | **$df_1$** | **$df_2$** | **F** | ***p*** | **partial $\eta^2$** |
|---|---|---|---|---|---|---|
| Exponential | Language | 1 | 92 | 9.232 | 0.003** | 0.091 |
| | Cognitive Status | 1 | 92 | 2.736 | 0.102 | 0.029 |
| | Language × Cognitive Status | 1 | 92 | 6.205 | 0.015** | 0.063 |
| Zipf | Language | 1 | 92 | 6.144 | 0.015** | 0.063 |
| | Cognitive Status | 1 | 92 | 5.465 | 0.022* | 0.056 |
| | Language × Cognitive Status | 1 | 92 | 7.817 | 0.006** | 0.078 |
| Mandelbrot | Language | 1 | 92 | 0.906 | 0.344 | 0.01 |
| | Cognitive Status | 1 | 92 | 2.1 | 0.151 | 0.022 |
| | Language × Cognitive Status | 1 | 92 | 1.103 | 0.296 | 0.012 |

Notes: *$p < .05$, **$p < .01$, *$p < .001$

**Table A2:** Token, type, and TTR statistics at individual level.

| Language | Cognitive Status | Participants | Token | Type | TTR |
|---|---|---|---|---|---|
| Chinese | Cognitively Impaired | CHI001 | 84 | 44 | 0.52 |
| Chinese | Cognitively Impaired | CHI002 | 64 | 44 | 0.69 |
| Chinese | Cognitively Impaired | CHI003 | 150 | 77 | 0.51 |
| Chinese | Cognitively Impaired | CHI004 | 64 | 39 | 0.61 |
| Chinese | Cognitively Impaired | CHI005 | 180 | 77 | 0.43 |
| Chinese | Cognitively Impaired | CHI006 | 97 | 49 | 0.51 |
| Chinese | Cognitively Impaired | CHI007 | 84 | 55 | 0.65 |
| Chinese | Cognitively Impaired | CHI008 | 96 | 43 | 0.45 |
| Chinese | Cognitively Impaired | CHI009 | 149 | 62 | 0.42 |
| Chinese | Cognitively Impaired | CHI010 | 85 | 58 | 0.68 |
| Chinese | Cognitively Impaired | CHI011 | 78 | 53 | 0.68 |
| Chinese | Cognitively Impaired | CHI012 | 85 | 44 | 0.52 |
| Chinese | Cognitively Impaired | CHI013 | 65 | 34 | 0.52 |
| Chinese | Cognitively Impaired | CHI014 | 75 | 48 | 0.64 |
| Chinese | Cognitively Impaired | CHI015 | 95 | 56 | 0.59 |
| Chinese | Cognitively Impaired | CHI016 | 221 | 75 | 0.34 |
| Chinese | Cognitively Impaired | CHI017 | 69 | 32 | 0.46 |
| Chinese | Cognitively Impaired | CHI018 | 88 | 56 | 0.64 |
| Chinese | Cognitively Impaired | CHI019 | 114 | 46 | 0.40 |
| Chinese | Cognitively Impaired | CHI020 | 74 | 49 | 0.66 |
| Chinese | Cognitively Impaired | CHI021 | 141 | 67 | 0.48 |
| Chinese | Cognitively Impaired | CHI022 | 84 | 45 | 0.54 |
| Chinese | Cognitively Impaired | CHI023 | 140 | 51 | 0.36 |
| Chinese | Cognitively Impaired | CHI024 | 65 | 38 | 0.58 |
| Chinese | Cognitively Normal | CHI025 | 91 | 60 | 0.66 |
| Chinese | Cognitively Normal | CHI026 | 139 | 71 | 0.51 |
| Chinese | Cognitively Normal | CHI027 | 64 | 42 | 0.66 |
| Chinese | Cognitively Normal | CHI028 | 174 | 92 | 0.53 |
| Chinese | Cognitively Normal | CHI029 | 56 | 35 | 0.62 |
| Chinese | Cognitively Normal | CHI030 | 101 | 65 | 0.64 |
| Chinese | Cognitively Normal | CHI031 | 92 | 61 | 0.66 |
| Chinese | Cognitively Normal | CHI032 | 61 | 28 | 0.46 |
| Chinese | Cognitively Normal | CHI033 | 139 | 75 | 0.54 |
| Chinese | Cognitively Normal | CHI034 | 107 | 50 | 0.47 |
| Chinese | Cognitively Normal | CHI035 | 201 | 106 | 0.53 |
| Chinese | Cognitively Normal | CHI036 | 124 | 61 | 0.49 |
| Chinese | Cognitively Normal | CHI037 | 59 | 37 | 0.63 |
| Chinese | Cognitively Normal | CHI038 | 116 | 68 | 0.59 |
| Chinese | Cognitively Normal | CHI039 | 169 | 93 | 0.55 |

| | | | | | |
|---|---|---|---|---|---|
| Chinese | Cognitively Normal | CHI040 | 108 | 58 | 0.54 |
| Chinese | Cognitively Normal | CHI041 | 83 | 45 | 0.54 |
| Chinese | Cognitively Normal | CHI042 | 36 | 31 | 0.86 |
| Chinese | Cognitively Normal | CHI043 | 47 | 30 | 0.64 |
| Chinese | Cognitively Normal | CHI044 | 187 | 79 | 0.42 |
| Chinese | Cognitively Normal | CHI045 | 77 | 47 | 0.61 |
| Chinese | Cognitively Normal | CHI046 | 68 | 49 | 0.72 |
| Chinese | Cognitively Normal | CHI047 | 68 | 43 | 0.63 |
| Chinese | Cognitively Normal | CHI048 | 141 | 75 | 0.53 |
| English | Cognitively Impaired | ENG001 | 123 | 62 | 0.50 |
| English | Cognitively Impaired | ENG002 | 109 | 44 | 0.40 |
| English | Cognitively Impaired | ENG003 | 66 | 41 | 0.62 |
| English | Cognitively Impaired | ENG004 | 139 | 68 | 0.49 |
| English | Cognitively Impaired | ENG005 | 56 | 40 | 0.71 |
| English | Cognitively Impaired | ENG006 | 123 | 52 | 0.42 |
| English | Cognitively Impaired | ENG007 | 75 | 49 | 0.65 |
| English | Cognitively Impaired | ENG008 | 51 | 33 | 0.65 |
| English | Cognitively Impaired | ENG009 | 109 | 63 | 0.58 |
| English | Cognitively Impaired | ENG010 | 49 | 28 | 0.57 |
| English | Cognitively Impaired | ENG011 | 105 | 54 | 0.51 |
| English | Cognitively Impaired | ENG012 | 68 | 45 | 0.66 |
| English | Cognitively Impaired | ENG013 | 33 | 24 | 0.73 |
| English | Cognitively Impaired | ENG014 | 97 | 54 | 0.56 |
| English | Cognitively Impaired | ENG015 | 70 | 36 | 0.51 |
| English | Cognitively Impaired | ENG016 | 99 | 50 | 0.51 |
| English | Cognitively Impaired | ENG017 | 48 | 31 | 0.65 |
| English | Cognitively Impaired | ENG018 | 73 | 43 | 0.59 |
| English | Cognitively Impaired | ENG019 | 85 | 52 | 0.61 |
| English | Cognitively Impaired | ENG020 | 189 | 68 | 0.36 |
| English | Cognitively Impaired | ENG021 | 78 | 47 | 0.60 |
| English | Cognitively Impaired | ENG022 | 51 | 34 | 0.67 |
| English | Cognitively Impaired | ENG023 | 177 | 64 | 0.36 |
| English | Cognitively Impaired | ENG024 | 66 | 40 | 0.61 |
| English | Cognitively Normal | ENG025 | 113 | 56 | 0.50 |
| English | Cognitively Normal | ENG026 | 97 | 55 | 0.57 |
| English | Cognitively Normal | ENG027 | 67 | 43 | 0.64 |
| English | Cognitively Normal | ENG028 | 143 | 71 | 0.50 |
| English | Cognitively Normal | ENG029 | 60 | 41 | 0.68 |
| English | Cognitively Normal | ENG030 | 78 | 46 | 0.59 |
| English | Cognitively Normal | ENG031 | 114 | 68 | 0.60 |
| English | Cognitively Normal | ENG032 | 95 | 52 | 0.55 |

| | | | | | |
|---|---|---|---|---|---|
| English | Cognitively Normal | ENG033 | 153 | 79 | 0.52 |
| English | Cognitively Normal | ENG034 | 89 | 56 | 0.63 |
| English | Cognitively Normal | ENG035 | 77 | 49 | 0.64 |
| English | Cognitively Normal | ENG036 | 78 | 49 | 0.63 |
| English | Cognitively Normal | ENG037 | 52 | 38 | 0.73 |
| English | Cognitively Normal | ENG038 | 197 | 91 | 0.46 |
| English | Cognitively Normal | ENG039 | 116 | 71 | 0.61 |
| English | Cognitively Normal | ENG040 | 109 | 59 | 0.54 |
| English | Cognitively Normal | ENG041 | 157 | 76 | 0.48 |
| English | Cognitively Normal | ENG042 | 100 | 55 | 0.55 |
| English | Cognitively Normal | ENG043 | 121 | 69 | 0.57 |
| English | Cognitively Normal | ENG044 | 158 | 77 | 0.49 |
| English | Cognitively Normal | ENG045 | 42 | 28 | 0.67 |
| English | Cognitively Normal | ENG046 | 80 | 58 | 0.72 |
| English | Cognitively Normal | ENG047 | 98 | 48 | 0.49 |
| English | Cognitively Normal | ENG048 | 76 | 44 | 0.58 |

# Lexical Diversity of Czech L2 Texts at Different Proficiency Levels

Michaela Hanušková[1*], Miroslav Kubát[1], Michaela Nogolová[1]

[1] University of Ostrava
[*] Corresponding author's email: mi.hanuskova@gmail.com

DOI: https://doi.org/10.53482/2025_59_427

## ABSTRACT

The study valuates how lexical diversity differs across language proficiency levels (A1–C1 according to the CEFR). The material used in the research comes from the CzeSL-SGT learner corpus belonging to the Czech National Corpus. This dataset contains more than 8,000 Czech texts written by non-native speakers of different proficiency levels. Moving Average Type-Token Ratio (MATTR) is used to calculate lexical diversity in this study. The results indicate that lexical diversity increases with writers' proficiency. There is also a significant difference between the development of lexical diversity of Slavic and non-Slavic native speakers.

**Keywords:** lexical diversity, MATTR, corpus linguistics, second language acquisition

## 1 Introduction

Lexical diversity (LD) is a characteristic that reflects the extent of the lexical knowledge of the writer speaker. LD refers to the variety of unique words (types) used in a spoken or written text. It assumes that every person has their own vocabulary reflected in their language production. (Kubát, 2016) A specific situation occurs with foreign language learners. Their vocabulary is gradually evolving, making lexical diversity a useful tool for better understanding the development of language acquisition. Simply put, a less proficient speaker tends to use a small number of lexical units and cannot achieve significant variability. In contrast, a more proficient speaker uses more variable lexical items to accomplish a task (Webb, 2019). Hence, LD is one of the most reliable indicators of linguistic proficiency and development in second language acquisition (SLA) (cf. Nasseri & Thompson, 2021).

This study aims to analyse development of lexical diversity of texts written by non-native speakers of the Czech language across different levels of language proficiency, from beginners (A1) to advanced learners (C1)[1]. We are motivated by the fact that the lexical diversity of Czech as L2 (and other Slavic languages) has not been studied quantitatively yet.

[1] Language proficiency levels in this study refer to the Common European Framework of Reference for Languages (CEFR).

Slavic languages still share a part of the lexicon, which comes from both a common protolanguage and close contact between different Slavic languages speakers (Karlíková et al., 2017). Therefore, we will also pay attention to the cross-linguistic influence of native Slavic and non-Slavic languages on the development of lexical diversity in Czech as L2. Compared to non-Slavic native speakers, Slavic native speakers are expected to use a considerably wider vocabulary in their texts. Although this aspect might play a role in SLA research, only a few studies cover cross-linguistic influence in lexical diversity research (cf. Shatz, 2021).

Lexical diversity will be measured by the Moving Average Type-Token Ratio (MATTR) (Covington & McFall, 2010). Although this indicator has been shown to be a suitable lexical diversity indicator in various fields of linguistics (especially stylometry), its application in SLA research is still relatively rare. However, recent studies have shown that MATTR is a suitable measure in SLA research. The strength of MATTR in quantitative text analysis lies in its independence from text size as well as its simplicity and straightforward interpretation.

The analysis is based on material from Czech National Corpus, namely the corpus CzeSL-SGT (Czech as a Second Language with Spelling, Grammar and Tags) (Šebesta et al., 2014) consisting of more twihan 8,000 texts ten by about 2,000 different authors with 54 different first languages. Furthermore, the corpus covers a wide range of language proficiency levels, from beginners to advanced learners. Thus, this material can be considered a substantial corpus for SLA research.

The study aims to answer two research questions. First, how do lexical diversity values develop across Czech L2 proficiency levels? The second question is whether speakers with Slavic L1 backgrounds differ from speakers with non-Slavic L1 backgrounds regarding the evolution of their lexical diversity values. If so, what are the differences?

## 2 Material

The CzeSL-SGT (Czech as a Second Language with Spelling, Grammar and Tags) corpus of non-native speakers of Czech with automatic annotation (Šebesta et al., 2014) is used in this study as the language material. It is a part of the Czech National Corpus. The raw CzeSL-SGT corpus consists of 8,617 texts written by 1,965 authors with 54 different first languages and was collected from 2009 to 2013. Each of the essays was equipped with metadata about a text (e.g. topic, size limit in the assignment, text length) and student (e.g. sex, age, L1, language proficiency level) (cf. Rosen, 2015). Information on language proficiency level and mother tongue was used for this study. The number of texts in each proficiency level can be found in Table 1. A detailed description of the corpus can be found on the Czech National Corpus website.

**Table 1:** Number of texts by proficiency level.

| Proficiency Level | Number of Texts |
|---|---|
| A1 | 2609 |
| A1+ | 315 |
| A2 | 2098 |
| A2+ | 570 |
| B1 | 1481 |
| B2 | 745 |
| C1 | 123 |
| C2 | 1 |
| total | 7942 |
| unknown | 675 |

As can be seen in Table 1, the numbers of texts at each proficiency level are unbalanced, and some texts are not even assigned to any proficiency level. Therefore, the following changes were made prior to the analysis.

- Since there is only one text at the C2 level, this level was excluded from the analysis.
- Texts with 'unknown' proficiency levels were also removed.
- Texts labelled A1+ and A2+ were excluded from the study because (a) the corpus documentation does not state on which parameters these levels are determined, and (b) these additional levels do not correspond to the CEFR framework.
- Based on Zenker and Kyle's (2021) findings, texts shorter than 55 words were removed.

In total, 6,073 texts covering levels A1, A2, B1, B2, and C1 were analysed in this research. Since we also focus on a potential cross-linguistic mother tongue influence, texts were also divided into the Slavic or non-Slavic groups. The final adjusted structure of the corpus used in this study can be found in Table 2.

**Table 2:** Number of analysed texts by proficiency level.

| Proficiency Level | Number of Texts | | |
|---|---|---|---|
| | Slavic | non-Slavic | Mix |
| A1 | 1466 | 556 | 2022 |
| A2 | 1215 | 625 | 1840 |
| B1 | 879 | 492 | 1371 |
| B2 | 511 | 211 | 722 |
| C1 | 80 | 38 | 118 |
| total | 4151 | 1922 | 6073 |

# 3 Methodology

## 3.1 Lexical Diversity

Lexical diversity can be examined with multiple indices. Indices usually express the relationship between the number of different words (types) and the number of all words (tokens) in a text. It is well known that fundamental indices like type-token ratio (TTR Johnson, 1994) and its variations Root TTR (Guiard, 1960), Log TTR (Chotlos, 1944; Herdan, 1960) are sensitive to text length. The longer the text, the lower the LD score (cf. Čech et al., 2014). In an attempt to address this issue, several revised indices have been proposed, such as Maas' index (Maas, 1972), Moving-average TTR (MATTR; Covington & McFall, 2010), the hypergeometric distribution density index (HD-D; McCarthy & Jarvis, 2007), standardized type-token ratio (zTTR) based on comparing the observed TTR with the referential TTR values representing texts of identical size (Cvrček & Chlumská, 2015), or the measure of textual lexical diversity (MTLD; McCarthy & Jarvis, 2010). However, none of them has gained universal acceptance.

Previous research indicated that measures like MATTR, HD-D, or MTLD are more stable than TTR on large texts that are similar in length (McCarthy & Jarvis, 2007; McCarthy & Jarvis, 2010) but do not mention the stability in short texts. A severe concern in SLA research is the sample length that varies widely across proficiency levels, and particularly problematic are texts at the lowest CEFR levels. Their length is usually less than 100 words.

Following previous studies (Koizumi, 2012; Koizumi & In'nami, 2012), Zenker & Kyle (2021) tested the stability of the latter indices on short texts. The study demonstrates the negligible correlation between LD value and text length for MATTR, HD-D or MTLD indices. They suggest that these are suitable for working with samples as short as 50 words. They claim that, in particular, MATTR appears to be the most stable of all indices. Zenker & Kyle (2021) also focused on the relationship between LD values and proficiency levels. The analysis of MATTR values confirmed statistically significant increases through the proficiency levels. Based on these findings, we decided to use the MATTR index in our research.

In addition to its statistical robustness on short texts, MATTR has proven useful across a range of empirical contexts. It has been applied to track longitudinal lexical development in L2 learning (Lissón & Ballier, 2018), to compare lexical proficiency in academic writing across L1 and L2 writers (Nasseri & Thompson, 2021), and to document gradual gains in ESL proficiency during university study (Vidal & Jarvis). Other research has shown that MATTR can differentiate learners across proficiency bands in diverse instructional settings, including Moroccan EFL (Ait Hammou, Larouz, & Fagroud, 2021). Taken together, these applications indicate that MATTR is sensitive both to developmental change and to proficiency-related differences while remaining stable with relatively short samples. This converging evidence supports our decision to employ MATTR in the present study.

## 3.2 Moving Average Type-Token Ratio (MATTR)

MATTR is defined as the mean of the TTR values of overlapping subtexts (the so-called windows) of the same length ($L$) in a text. The formula of TTR is defined as follows:

$$TTR = \frac{V}{N}$$

Where $V$ is vocabulary (number of types) and $N$ is text size (number of tokens).

The calculation procedure of MATTR is as follows:

1. A text is split into windows with an arbitrarily chosen size $L$.
2. The window moves forward one token at a time.
3. The TTR is calculated for every single window in the text.
4. Finally, the MATTR is calculated as an arithmetic mean of all the TTR values.

Let us demonstrate MATTR computation on a simple example of a sequence of 7 characters: a, b, c, d, a, a, b. Text length $N$ = 7, vocabulary V = *4*. If we choose a window size of 4 tokens ($L$ = 4), we obtain 4 overlapping windows:

1. a, b, c, d (TTR = 4 / 4 = 1)
2. b, c, d, a (TTR = 4 / 4 = 1)
3. c, d, a, a (TTR = 3 / 4 = 0.75)
4. d, a, a, b (TTR = 3 / 4 = 0.75)

The resulting MATTR value is calculated as the mean of the four obtained TTR values: MATTR = (1 + 1 + 0.75 + 0.75) / 4 = 0.875.

The important setting of the MATTR measurement is the window size ($L$). There is no ideal value suitable for every research. The length is usually set according to the shortest text in the corpus. The window size obviously cannot be longer than the shortest analysed text. On the other hand, the window length should be long enough to cover a sufficient text sample. Language ability, such as lexical diversity, can be barely expressed in a sequence of 5 or 10 words. It is necessary to measure such a characteristic on a longer sample. The window size can therefore vary significantly in different studies based on the analysed material. For example, stylometric analysis of novels can work with samples of hundreds of words, while investigating newspaper articles requires a window size of about 50 to 100 words. Consequently, the choice of L balances comparability and representativeness. Given that we analyse rather short texts (especially at beginner proficiency levels), we set the window size to 50 tokens, which can be considered a sufficient value for detecting lexical diversity in L2 texts.

Czech has a rich morphology in which nouns, adjectives, pronouns, numerals, and verbs are inflected to modify grammatical functions. For example, a lemma 'pes' (a dog) consists of several different word forms based on seven grammatical cases indicating their function in a sentence and two numbers (singular and plural) (see Table 3). Therefore, a lemma is a basic unit for calculating lexical diversity in this research.

**Table 3:** Grammatical paradigm of Lemma 'pes' (a dog).

| Case / Number | singular | plural |
|---|---|---|
| nominative | pes | psi, psové |
| genitive | psa | psů |
| dative | psovi, psu | psům |
| accusative | psa | psy |
| vocative | pse | psi, psové |
| locative | psovi, psu | psech |
| instrumental | psem | psy |

Software MATTR developed by Covington & McFall (2010) was used for the computation of lexical diversity in this research.

# 4 Results

In this chapter, the resulting MATTR values are presented as follows. First, the general results of different levels of language proficiency, regardless of the native language, are visualized by the graphs in Figures 1 and 2. Then the differences between the Slavic and non-Slavic groups are shown in Figures 3–5.

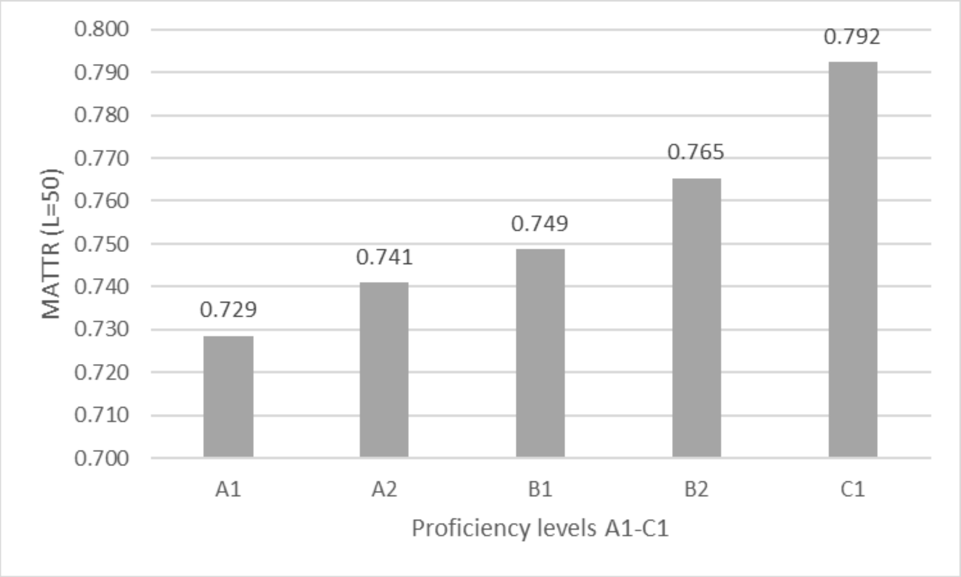

**Figure 1:** Average MATTR values at proficiency levels A1–C1 regardless of the native language.

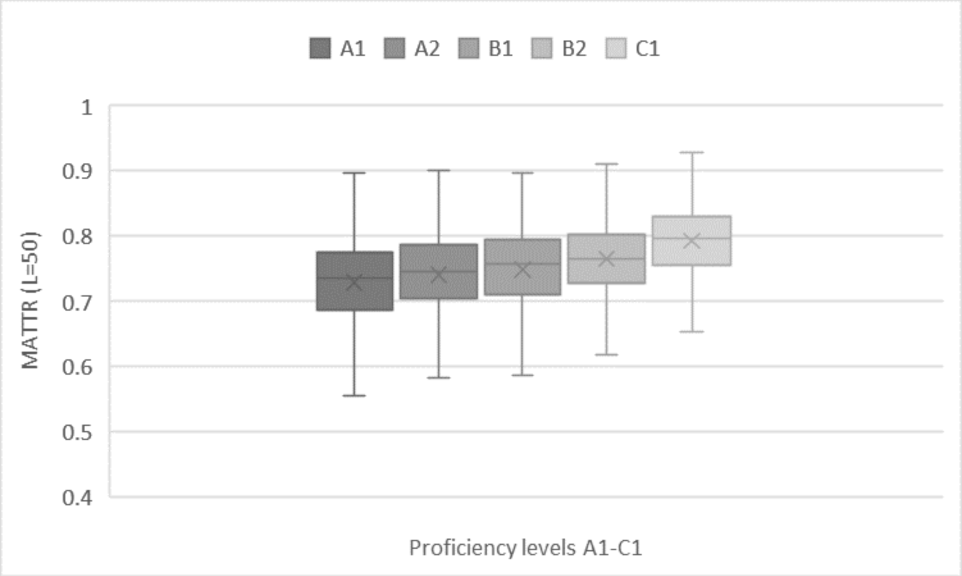

**Figure 2:** MATTR values at proficiency levels A1–C1 regardless of native language (boxplot without outliers).

The average MATTR values in Figure 1 show an evident increasing trend for lexical diversity in all levels of language proficiency analysed (A1–C1). Since the arithmetic means could be misleading, the dispersion of the resulting values obtained is visualized by a boxplot in Figure 2, where the tendency is also evident. All pairs of levels were statistically tested to ensure that the differences between the levels are significant. Since the obtained data are not normally distributed, we decided to apply the Wilcoxon-Mann-Whitney test, which is generally considered a non-parametric alternative to the t-test. The results show that all differences between all levels of language proficiency are statistically significant (p-value ≤ 0.05). Therefore, we can conclude that lexical diversity increases significantly with each proficiency level.

Our findings agree with previous research focused on the development of lexical diversity in other languages. Similar results can be found in Shatz (2021), who analysed lexical diversity on a large material of several thousand English texts across different CEFR levels (A1–B2).

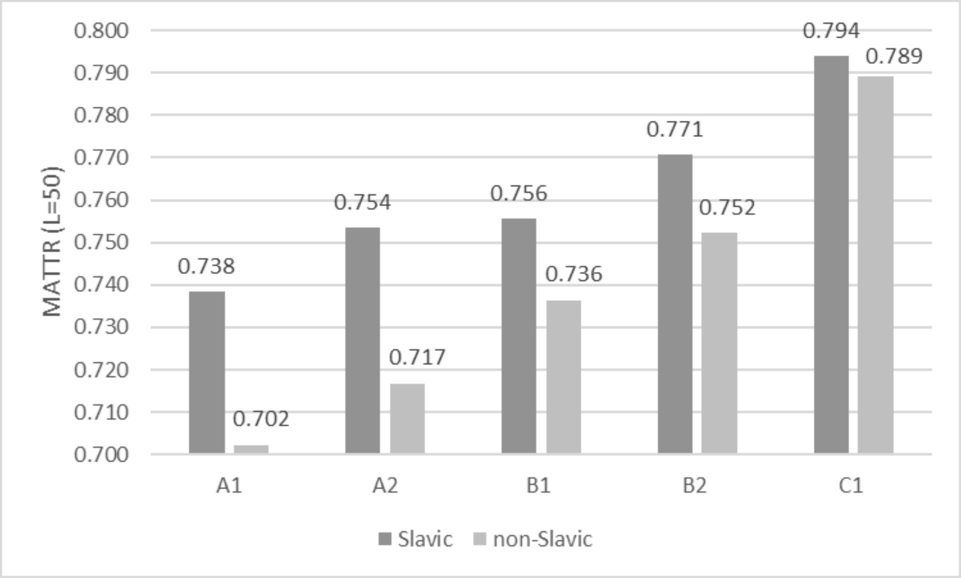

**Figure 3:** Comparison of average MATTR values at proficiency levels A1–C1 between Slavic and non-Slavic learners.

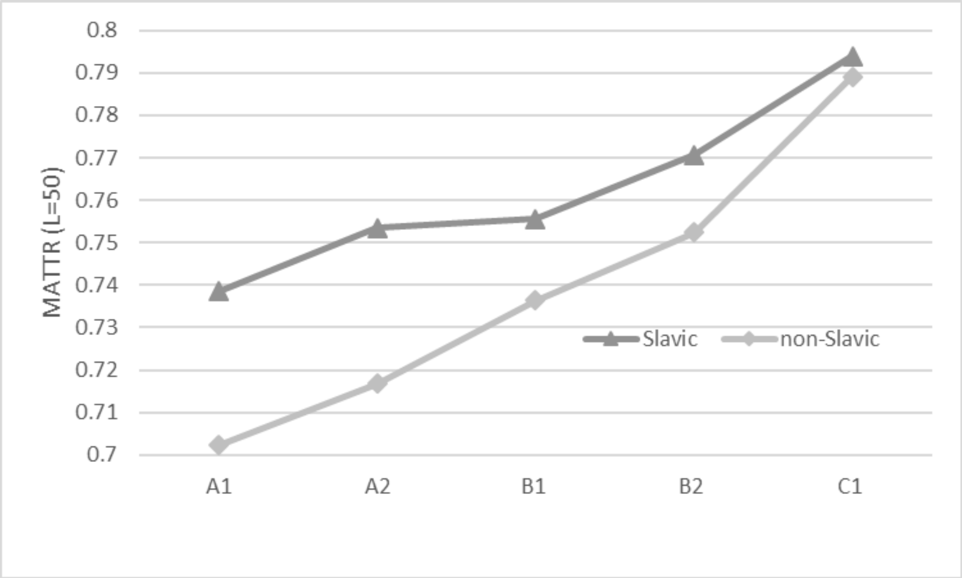

**Figure 4:** Comparison of average MATTR values at proficiency levels A1–C1 between Slavic and non-Slavic learners.

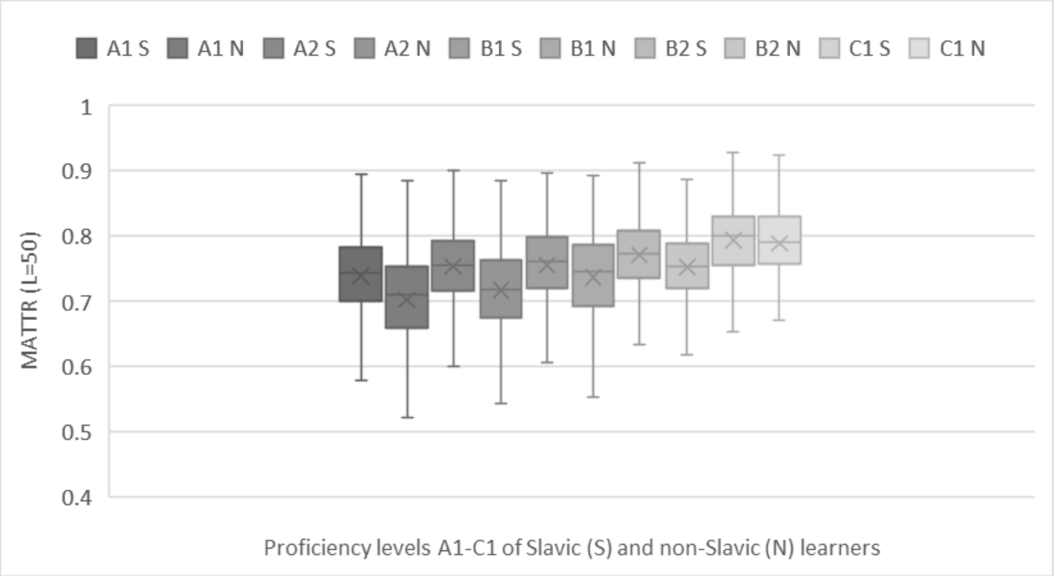

**Figure 5:** Comparison of MATTR values at proficiency levels A1–C1 between Slavic and non-Slavic learners (boxplot without outliers).

As can be seen in Figures 3–5, the resulting MATTR values of texts written by learners with Slavic and non-Slavic L1 show three general findings. First, the increasing tendency of lexical diversity is present in both groups (Slavic and non-Slavic). Second, students with a Slavic mother tongue reach higher average MATTR values at all levels of language proficiency. Third, the higher the level of language proficiency, the smaller the gap between Slavic and non-Slavic speakers. This shrinking gap is most visible in Figure 4, where the difference is minimal at the advanced level C1. To verify our findings statistically, we applied the Wilcoxon-Mann-Whitney test to test differences between Slavic and non-Slavic groups. The results show that the difference is always statistically significant (p-value ≤ 0.05) except for the C1 level. The statistical test, therefore, confirms the preliminary conclusions based on the graphs.

Our findings confirm our expectations that Slavic native speakers use a considerably wider vocabulary in their texts than their non-Slavic counterparts. The influence of overlapping vocabulary of one language family (Slavic languages) seems to be very strong and intuitive. Interestingly, Shatz (2021) concludes in his research on English as a second language that lexical similarity between the L1 and the L2 does not influence L2 lexical diversity, regardless of learners' L2 proficiency. His findings are based on the lexical distance between languages measured by similarities using Swadesh lists (Swadesh, 1971), which suggests that this type of calculation may have limitations that can bias the results. At the same time, the overall influence of cross-linguistic factors on lexical diversity has been examined only to a limited extent. While our results suggest an effect related to typological proximity, further studies across different languages and contexts are necessary to assess its generalisability.

## 5 Conclusion

Based on the obtained data, we can answer the research questions stated in the Introduction of the study. We discovered that lexical diversity increases significantly with the writer's proficiency of Czech L2 across the whole scale of analysed levels (A1–C1). The increasing tendency was also confirmed by the statistical test, where all differences between individual levels were statistically significant. We can conclude that lexical diversity is a crucial feature in learning a second language.

Besides the overall tendency of lexical diversity development across different proficiency levels, we also focused on the differences between Slavic and non-Slavic native speakers. The results show a significant difference between the development of the lexical diversity of the two groups. According to the expectation, Slavic native speakers reached significantly higher MATTR values at all proficiency levels except for advanced level C1. We can also state that the higher the level of language proficiency, the smaller the difference between Slavic and non-Slavic speakers. The cross-linguistic influence is, therefore, most visible at beginner levels.

We can also conclude that MATTR is a suitable method for measuring the lexical diversity across CEFR levels, given the strong association between L2 proficiency level and lexical diversity found in this study and previous research (e.g. Zenker & Kyle, 2021). MATTR seems to be a reliable tool for measuring lexical diversity of texts with various text lengths, which is essential, especially in the case of very short texts typical for beginner L2 learners.

## Acknowledgements

The research was supported by University of Ostrava, project No. SGS06/FF/2022.

## References

**Ait Hammou, B., Larouz, M., Fagroud, M.** (2021). Word frequency, Range and Lexical diversity: Picking Out Changes in Lexical Proficiency among University Learners in an EFL Context. *International Journal of Linguistics and Translation Studies*, 2(2), pp. 22–38. https://doi.org/10.36892/ijlts.v2i2.131

**Council of Europe. Council for Cultural Co-operation**. Education Committee. Modern Languages Division. (2001). *Common European Framework of Reference for Languages Learning, teaching, assessment.* Cambridge University Press.

**Čech, R., Popescu, I.-I., Altmann, G**. (2014). *Metody kvantitativní analýzy (nejen) básnických textů*. Olomouc: Univerzita Palackého v Olomouci.

**Cvrček, V., Chlumská, L.** (2015). Simplification in translated Czech: a new approach to type-token ratio. *Russian Linguist 39*, pp. 309–325. https://doi.org/10.1007/s11185-015-9151-8

**Karlíková, H., Skalka, B. & Večerka, R.** (2017). SLOVANSKÉ JAZYKY. In: Karlík, P, Nekula, M, Pleskalová, J. (Eds.). *CzechEncy - Nový encyklopedický slovník češtiny*. URL: https://www.czechency.org/slovnik/SLOVANSKÉ JAZYKY

**Kubát, Miroslav**. *Kvantitativní analýza žánrů*. Ostrava: Ostravská univerzita, 2016.

**Nasseri, M., Thompson, P.** (2021). *Lexical Density and Diversity in Dissertation Abstracts: Revisiting English L1 vs. L2 text differences*. Assessing Writing, 47, 100511. https://doi.org/10.1016/j.asw.2020.100511

**Rackevičienė, S., Utka, A., Bielinskienė, A., Rokas, A.** (2022). Distribution of Terms across Genres in the Annotated Lithuanian Cybersecurity Corpus. *Respectus Philologicus*, 41(46), 26–42. http://dx.doi.org/10.15388/RESPECTUS.2022.41.46.105

**Rosen, A.** (2015). *CzeSL-SGT – a corpus of non-native speakers' Czech with automatic annotation*. https://doi.org/10.13140/RG.2.1.1906.2487

**Šebesta, K. et al.** (2014). *AKCES 5 (CzeSL-SGT) Release 2*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL) http://hdl.handle.net/11234/1-162.

**Shatz, I.** (2022). *The Potential Influence of Crosslinguistic Similarity on Lexical Transfer: Examining Vocabulary Use in L2 English* (Doctoral dissertation, University of Cambridge).

**Zenker, F., Kyle, K.** (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, 47, 100505.

**Lissón, P., Ballier, N.** (2018). Investigating Lexical Progression through Lexical Diversity Metrics in a Corpus of French L3. *Discours. Revue de linguistique, psycholinguistique et informatique. A Journal of Linguistics, Psycholinguistics and Computational Linguistics,* 23. https://doi.org/10.4000/discours.9950

**Swadesh, M.** (1971). *The Origin and Diversification of Language*. Ed. post mortem by Joel Sherzer. Chicago: Aldine. Contains final 100-word list on p. 283.

**Treffers-Daller, J.** (2013). Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability. In: Jarvis, S., Daller, M. (Eds.), *Vocabulary Knowledge: Human Ratings and Automated Measures*, pp. 79–103. John Benjamins Publishing Company.

**Vidal, K., Jarvis, S.** (2020). Effects of English-medium instruction on Spanish students' proficiency and lexical diversity in English. *Language Teaching Research*, 24(5), 568–587. https://doi.org/10.1177/1362168818817945

**Webb, S.** (Ed.). (2020). *The Routledge Handbook of Vocabulary Studies (Vol. 2)*. London: Routledge.

# 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 $\mu$ 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 ($\lambda_{norm}$) and normalizing the number of meanings ($\mu_{min}$), 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<br>$r$: word rank<br>$\alpha, A$: parameters | (Zipf, 1932, 1935, 1949) |
| Brevity law | $f \sim \exp(-\lambda \ell), \quad \lambda > 0$ | $f$: frequency<br>$l$: length<br>$\lambda$: parameter | (Torre et al., 2019)<br>(Bentz and Ferrer-i-Cancho, 2016) |
| Herdan-Heaps' law | $n = cT^{\theta}$ | $n$ : word types<br>$T$ : word tokens<br>$c, \theta$ :parameters | (Herdan, 1960)<br>(Heaps, 1978) |
| Zipf's law of meaning distribution | $\mu = C_1 r^{\gamma}$ | $\mu$ : number of meanings<br>$r$ : word rank<br>$C_1, \gamma$ : parameters | (Zipf, 1945)<br>(Ferrer-i-Cancho and Vitevitch, 2018) |
| Zipf's meaning-frequency law | $\mu = C_2 f^{\delta}$ | $\mu$ : number of meanings<br>$f$ : frequency<br>$C_2, \delta$ : parameters | (Zipf, 1945)<br>(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.

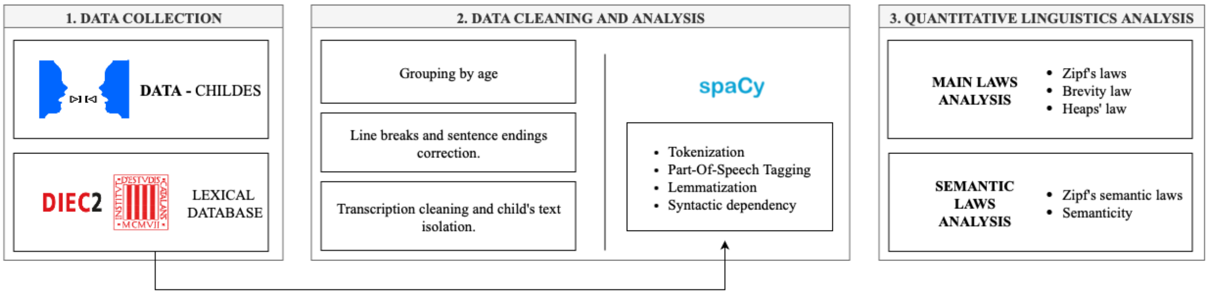

**Figure 1:** Schematic of the proposed analysis workflow. Source: Tubella Salinas, 2025, with permission.

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 $\alpha$ 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 $\theta$ 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 $\theta$ 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

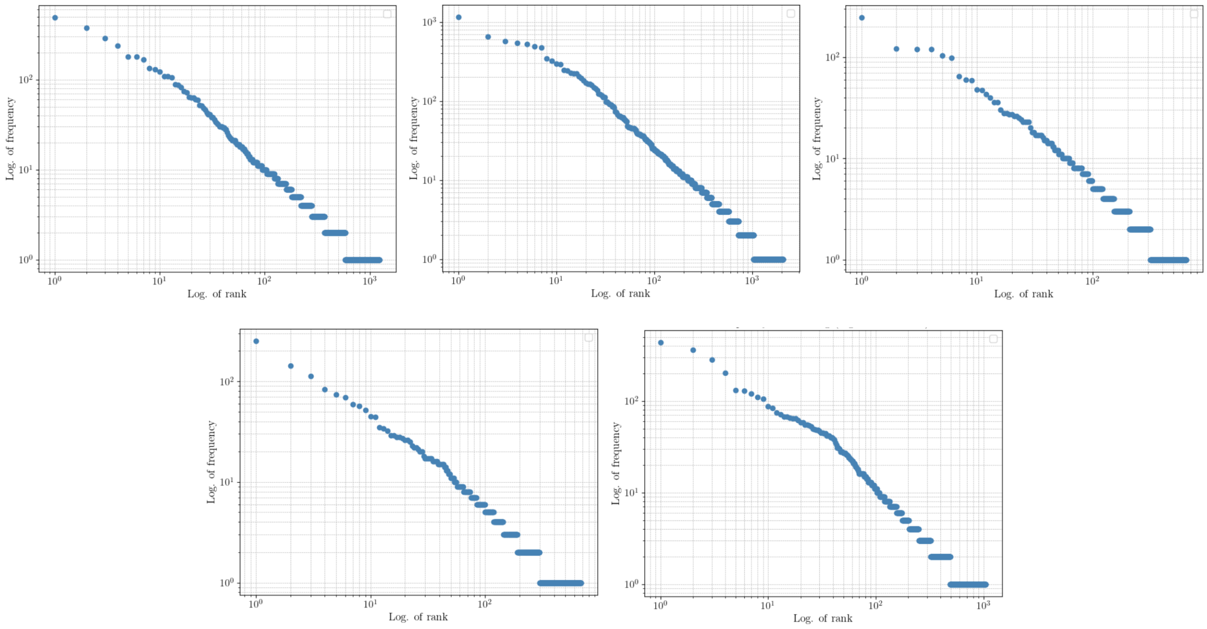

**Figure 2:** Zipf's law: Rank vs Frequency in logarithmic scale. Top (from left to right): age groups 2–3, 3–5, and 9–10 years old. Bottom (from left to right): age groups 12–13 and 16–18 years old.

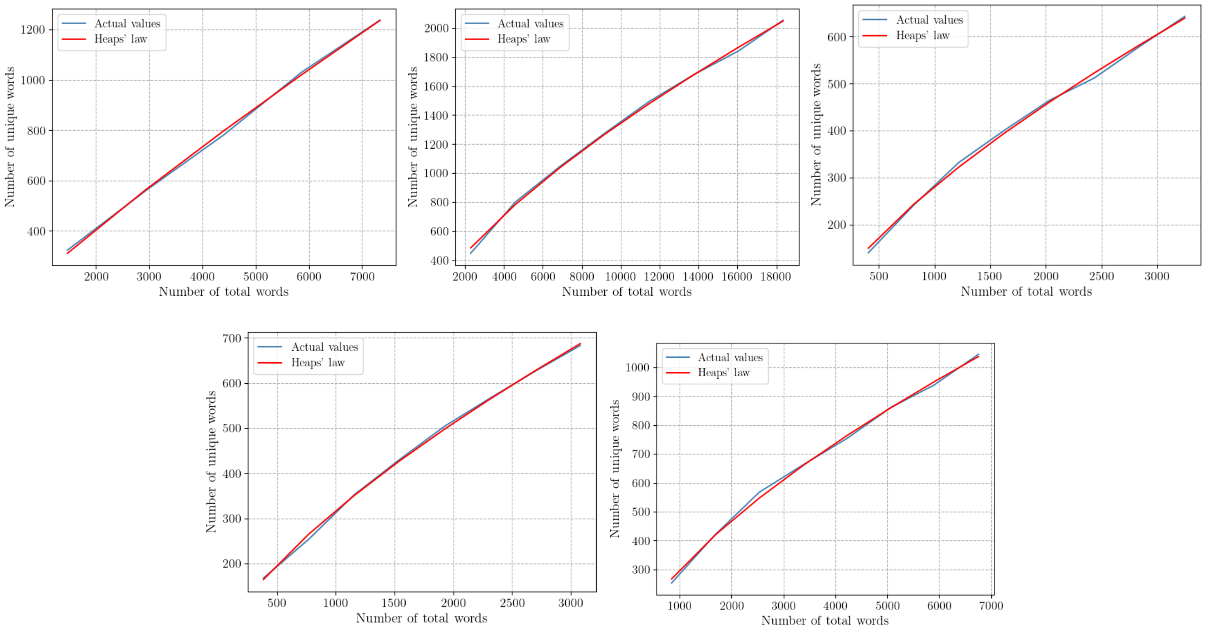

**Figure 3:** Heaps' law: total words vs unique words. Top (from left to right): age groups 2–3, 3–5, and 9–10 years old. Bottom (from left to right): age groups 12–13 and 16–18 years old.

**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 |
| $\theta$ | 0.8373 ± 0.03 | 0.6935 ± 0.01 | 0.6984 ± 0.02 | 0.6865 ± 0.01 | 0.6521 ± 0.01 |

$\theta$ 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 $\theta$ shows a decreasing tendency. A higher $\theta$ 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 ($\mu$) 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 $\rho$, and Kendall's $\tau$), 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 | $r = -0.198$<br>$(p = 6.65\text{e-}13)$ | $r = -0.191$<br>$(p < 2\text{e-}16)$ | $r = -0.279$<br>$(p = 1.44\text{e-}13)$ | $r = -0.255$<br>$(p = 4.32\text{e-}12)$ | $r = -0.253$<br>$(p < 2\text{e-}16)$ |
| Spearman | $\rho = -0.254$<br>$(p < 2\text{e-}16)$ | $\rho = -0.289$<br>$(p < 2\text{e-}16)$ | $\rho = -0.334$<br>$(p < 2\text{e-}16)$ | $\rho = -0.380$<br>$(p < 2\text{e-}16)$ | $\rho = -0.370$<br>$(p < 2\text{e-}16)$ |
| Kendall | $\tau = -0.205$<br>$(p < 2\text{e-}16)$ | $\tau = -0.232$<br>$(p < 2\text{e-}16)$ | $\tau = -0.272$<br>$(p < 2\text{e-}16)$ | $\tau = -0.308$<br>$(p < 2\text{e-}16)$ | $\tau = -0.299$<br>$(p < 2\text{e-}16)$ |

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 $\gamma$, 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 $\gamma$ 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** | $C_1$ | $C_2$ | $\gamma$ | $\delta$ |
|---|---|---|---|---|---|
| 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 ($\mu$) (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

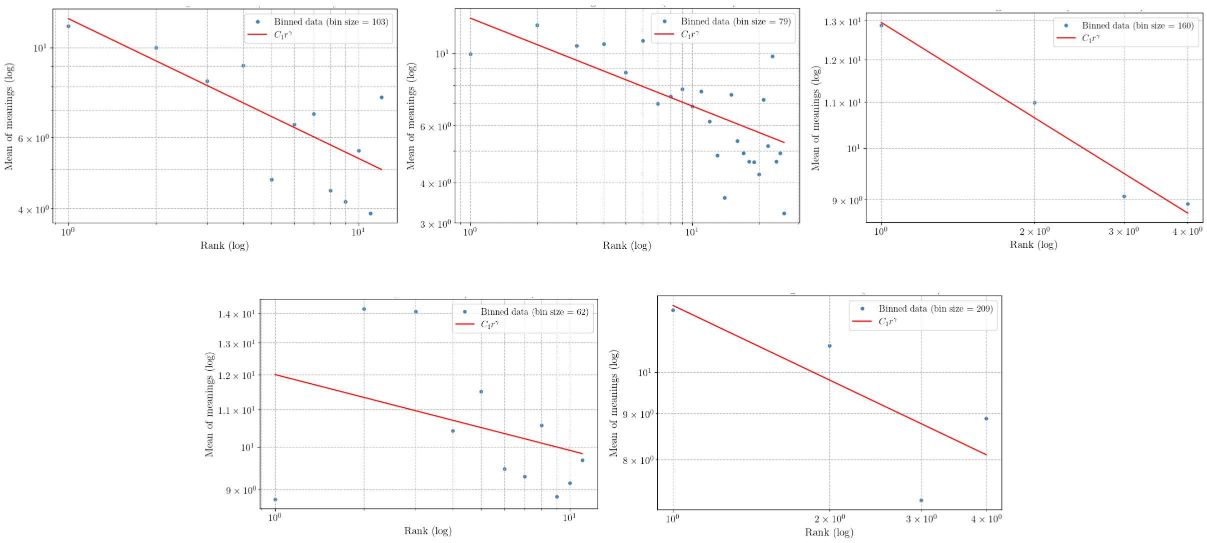

**Figure 4:** Zipf's law of meaning distribution: Average number of word meanings as a function of frequency rank ($r$), using equal size binning (blue). The red curve represents the best fitting power-law model. Top (from left to right): age groups 2–3, 3–5, and 9–10 years old. Bottom (from left to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

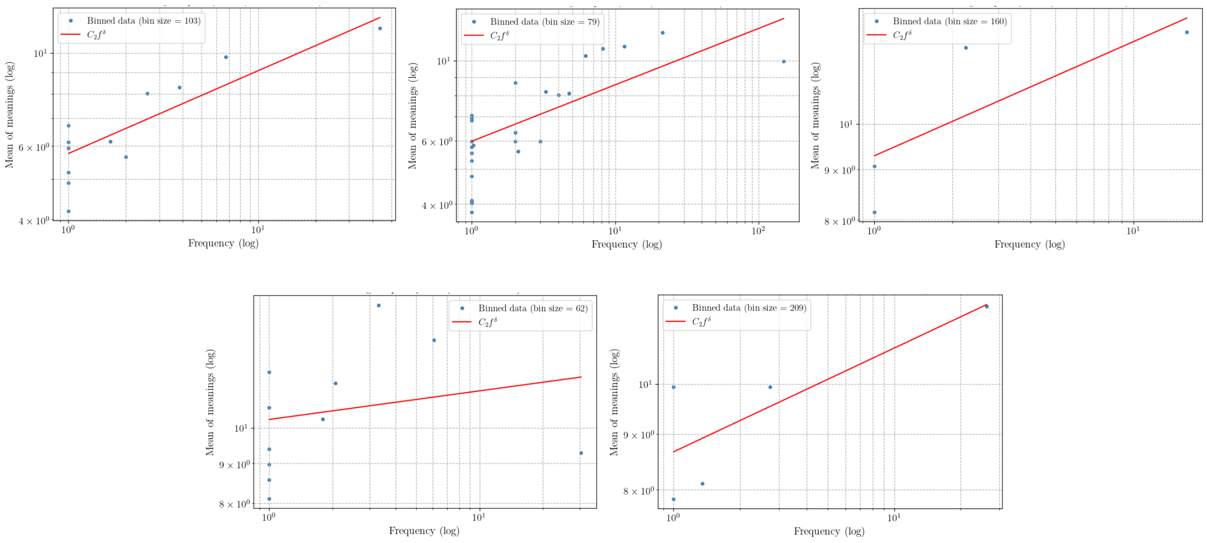

**Figure 5:** Zipf's meaning-frequency law: Average number of word meanings as a function of frequency ($f$), using equal size binning (blue). The red curve represents the best fitting power-law model. Top (from left to right): age groups 2–3, 3–5, and 9–10 years old. Bottom (from left to right): age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

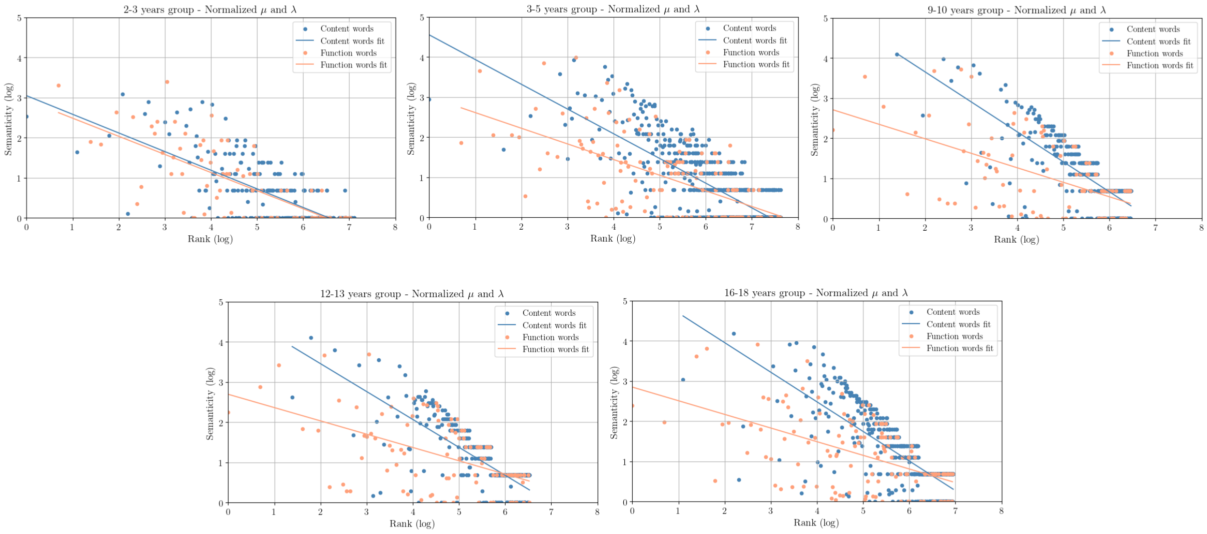

**Figure 6:** Frequency rank vs semanticity with $\mu$ and $\lambda$ normalizations, at $d = 4$. Top (from left to right): age groups 2–3, 3–5, and 9–10 years old. Bottom (from left to right): age groups 12–13 and 16–18 years old.

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 |

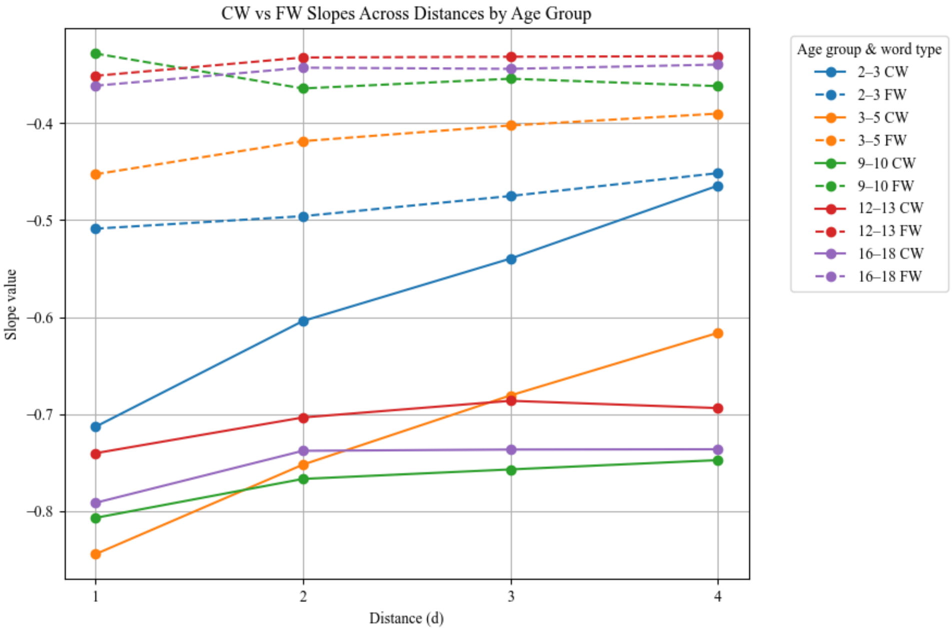

**Figure 7:** Comparison between the slope coefficients of linear regression of frequency versus semanticity for content words (CW) and function words (FW), evaluated across lexical network distances (d) ranging from 1 to 4.

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 $\alpha$ 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 | $C_1$ | $C_2$ | $\gamma$ | $\delta$ |
|---|---|---|---|---|---|---|
| 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 |

# Thematic structure of images in Vladimir Nabokov's lyrics

Vadim Andreev [1*]

[1] HSE University, Moscow
[*] Corresponding author's email: vadim.andreev@ymail.com

DOI: https://doi.org/10.53482/2025_59_429

**ABSTRACT**

Nabokov, known primarily for his prose, was also a remarkable poet. The article aims to examine the semantic (lexical) features of Nabokov's images in his lyrics. These characteristics include 23 semantic (thematic) classes of words that fill two positions of the figurative model – the position of the Target of the metaphoric transfer and the position of its Source. The material includes 4 lyrical collections by Nabokov, published at different stages of his creative path – an early collection (published when he lived in Russia), 2 later collections from the Berlin period (middle stage) and one collection of mature creative activity when Nabokov lived in the USA. The article examines the relationships between the features of the two positions of the image, the distribution of semantic classes of words in the image system, and changes in the frequencies of these classes over time.

**Keywords:** Nabokov lyrics, images, thematic classes, target, source

## 1 Introduction. System of features and research material

V.V. Nabokov's poems are in the shadow of his prose, but it was with poetry that he (under the pseudonym Vladimir Sirin) began his way in literature. The first works written by the beginning author (unsuccessful in the opinion of his contemporaries and later Nabokov himself), appeared in the collection *Poems (Stikhi)*, published in 1916 in Petrograd, when the author was only 17 years old. Then in 1918, new poems followed in a joint collection with A.V. Balashov *Two Ways (Dva puti)*. The next stage of his literary activity was the Berlin period (1922-1937) during which Nabokov not only wrote his first novels *Mary*, *The Luzhin Defense*, *The Gift* and other works, but also published collections of poems *A Bunch (Grozd')* (1923), *The Empyrean Path* (*Gorniy put'*) (1923), *The Return of Chorb. Stories and Poems* (*Vozvrashcheniye Chorba. Rasskazy i stikhi*) (1930). Later Nabokov continued to write and publish poems throughout his career. He published *Poems, 1929-1951* (*Stikhotvoreniya, 1929-1951*) (1952) followed by several collections in which he included poems from periodicals. From the above it follows that Nabokov's poetry should in no way be underestimated or overlooked in his creative life (Boyd 1999; De Vries 1991; Morris 2010).

The purpose of this article is to define the main features of Nabokov’s poetic image system and to identify the tendencies of its development. In our study we use the scheme of images proposed for Russian poetry (Pavlovich 1995, 1999), which is based on the approach to the study of conceptual metaphor in cognitive linguistics.

According to this approach, an image is a fragment of text in which a metaphoric model is realized. The model includes Target domain (the recipient of new properties) and the Source of the properties transferred to the Target, Target and Source being not identical from the point of view of the scientific picture of the world.

Thus, in *Ti, yelochka ustala?* (Are you tired, Christmas-tree?) the author transfers the properties of a human being (ability to be tired, ability to be an interlocutor) to a plant. In the verse line *Prolyutsa shepchuschiye zvuki* (Whispering sounds will pour out) the properties of liquids are transferred to sounds.

To describe the Target and the Source, a number of thematic classes (ThC) are used, a list of which is given below. First, the name of the thematic class is given, then, if applicable, explanations in brackets, and after the colon follow examples from Nabokov's images.

*Area* (including in buildings): countryside, field, land, plain, desert, roof.

*Auditory perception*: sound, voice, music.

*Clothes* (including cloth): linen, dress, kerchief, robe, stockings, velvet

*Container*: container, box, cup, goblet.

*Darkness* (including dark colours): black, dark, darken, shadow, twilight.

*Existential phenomenon*: birth, death, die, life, live.

*Fire:* burn, burning, fire, fiery.

*Food:* anise-tasted, bread, lemon.

*Household items*: cupboard, wardrobe, chest of draws, bed.

*Information*: word, phrase, fairy-tale, tale.

*Instrument* (working or military tools): hammer, needle, scissors, shield, sword.

*Jewelry*: amber, emerald, gold, ruby, silver, pearl.

*Light*: candle, glow, lamp, light, star, sun, ray.

*Liquid*: drop, flow, pour, river, sea, water.

*Living being* (people and animals): man, mother, son, reader, soldier, beetle, bird, to be tired, to feel, butterfly, fish, grasshopper.

*Mental phenomenon* (emotions and mental processes): anger, fear, feeling, hatred, hope, idea, love, sadness, thought, memory.

*Natural phenomenon*: frost, snow-storm, storm, rain, thunder, wind.

*Part of body* (human or animal): hand, hair, heart, eye, foot, lips, tongue, wing.

*Plant*: plant, birch, flower, fur-tree, grass, oak, pine.

*Social phenomenon*: equity, freedom, peace, power, war

*Substance*: glass, dissolve, dust, murky, steel, wax

*Time period*: hour, minute, moment, day, the past.

*Transport*: airplane, boat, car, cart, train.

This scheme of features was compiled empirically during the analysis of the figurative system of Russian poetry, as well as the experience of its application to the poems of American poets (Pavlovich 1995; Andreev 2012).

Words of the above mentioned thematic classes can be found in the model both in the Source and Target positions. In the phrase *A wave of tall grass* the word *wave* (Liquids thematic class) is the source of new properties for the *grass* (Plant thematic class). On the other hand, in the example *screaming sea* it is the representative of the Liquid thematic class *sea* that takes on new properties transferred from the Living being theme, thus acting as Target.

Research material compiles 4 poetry collections that mark different stages in Nabokov's poetic life[1].

*Two Ways* (TW) (1918) – the collection published by a young author unknown to the general public.

*A Bunch* (BN) (1923) – the first collection of poems published during Berlin period.

*The Return of Chorb* (RTB) (1930). The last collection with poems during Berlin period published when the author was already relatively famous thanks to his first novels.

*Poems, 1929-1951* (PM) (1952). This collection was published when Nabokov had firmly established himself among famous writers. Poems written since 1939 were taken from this collection for this study.

## 2 Results

As a result of the analysis, the following data on the frequency of the thematic classes in both positions of the model were obtained (Table 1). Counts were made separately for the Source and Target themes.

[1] **Nabokov, V.V.** (2002). *Stihotvoreniya*. Akademicheskij proekt, Saint Petersburg, Russia.

**Table 1.** Frequencies of thematic classes in 4 books.

| Thematic class | Target domain |  |  |  | Source domain |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | TW | BN | RTB | PM | TW | BN | RTB | PM |
| Area | 8.14 | 4.12 | 2.83 | 2.46 | 5.81 | 3.61 | 1.94 | 2.90 |
| Auditory perception | 3.49 | 2.58 | 2.24 | 1.56 | 1.74 | 0.00 | 0.60 | 0.22 |
| Clothes | 0.00 | 1.03 | 0.30 | 0.45 | 2.91 | 2.58 | 1.04 | 1.12 |
| Container | 0.00 | 0.52 | 0.00 | 0.00 | 2.33 | 0.52 | 0.60 | 0.67 |
| Darkness | 0.00 | 4.64 | 0.75 | 0.45 | 0.58 | 1.55 | 0.15 | 0.67 |
| Existential phenomenon | 0.58 | 1.55 | 1.04 | 3.79 | 0.00 | 0.00 | 0.00 | 0.00 |
| Fire | 0 | 0 | 0 | 0 | 1.74 | 0.52 | 1.34 | 0.89 |
| Food | 0.00 | 1.55 | 0.00 | 0.00 | 0.00 | 1.03 | 0.30 | 0.45 |
| Household items | 0.58 | 1.03 | 3.13 | 0.67 | 3.49 | 1.55 | 0.45 | 0.22 |
| Information | 1.16 | 2.06 | 0.45 | 3.57 | 0.58 | 2.06 | 0.00 | 1.12 |
| Instrument | 2.91 | 0.52 | 0.89 | 0.22 | 0.00 | 0.00 | 0.75 | 0.45 |
| Jewelry | 0.58 | 0.00 | 0.00 | 0.22 | 2.91 | 4.12 | 0.45 | 0.67 |
| Light | 5.23 | 7.22 | 2.38 | 0.89 | 4.07 | 7.22 | 2.53 | 1.12 |
| Liquid | 2.33 | 2.58 | 0.45 | 0.89 | 2.33 | 2.06 | 1.04 | 0.45 |
| Living being | 1.74 | 10.31 | 1.04 | 5.13 | 30.23 | 24.23 | 11.18 | 11.61 |
| Mental phenomenon | 5.23 | 2.06 | 2.38 | 1.79 | 1.74 | 4.12 | 0.89 | 1.56 |
| Natural phenomenon | 4.07 | 0.52 | 0.30 | 0.00 | 0.00 | 0.00 | 0.30 | 0.22 |
| Plant | 18.60 | 3.61 | 1.49 | 1.12 | 0.58 | 2.58 | 0.30 | 1.79 |
| Part of body | 6.40 | 9.28 | 2.53 | 1.34 | 0.58 | 2.06 | 1.04 | 0.89 |
| Social phenomenon | 0.58 | 0.00 | 0.45 | 4.02 | 0.58 | 0.00 | 0.30 | 0.22 |
| Substance | 1.74 | 2.06 | 0.89 | 0.89 | 4.07 | 5.15 | 0.75 | 2.23 |
| Time period | 3.49 | 3.09 | 1.64 | 1.12 | 1.16 | 0 | 0 | 0.22 |
| Transport | 0.58 | 4.64 | 1.49 | 0.00 | 0.00 | 0.00 | 0.75 | 0.89 |

Notes: the frequencies are normalized to 100 lines.

The relationship between the thematic organization of Source and Target and its stability over time was established using the rank correlation coefficient. For each of the four collections, the Spearman rank correlation coefficient was calculated between the thematic classes of Source and Target.

All the obtained coefficients (for STW = 0.10, SBN = 0.37, SRTB = 0.21, SPM = 0.17) turned out to be statistically insignificant. Thus, the spectra of the ThCs that Nabokov seeks to rethink (Target) and the ThCs that he intuitively considers understandable (Source) are fundamentally different.

At the same time, comparison of the ranks of the Target themes from different collections yields significantly different results – here statistically significant correlations for $p < 0.05$ are observed (Table 2). The same applies to the Source themes (see also Table 3).

**Table 2:** Rank correlation coefficients between thematic classes of Target domains across collections.

| | **TW** | **BN** | **RTB** | **PM** |
|---|---|---|---|---|
| **TW** | x | | | |
| **BN** | 0.48 | x | | |
| **RTB** | 0.66 | 0.63 | x | |
| **PM** | 0.44 | 0.38* | 0.51 | x |

**Table 3**: Rank correlation coefficients between thematic classes of Source domains across collections.

| | **TW** | **BN** | **RTB** | **PM** |
|---|---|---|---|---|
| **TW** | x | | | |
| **BN** | 0.72 | x | | |
| **RTB** | 0.60 | 0.53 | x | |
| **PM** | 0.50 | 0.80 | 0.58 | x |

As can be seen from the data in this table, in only one case is the coefficient statistically insignificant, but at the same time, almost all their values are moderate in strength. This can be interpreted as follows. On the one hand, this means that the hierarchy of classes is more or less preserved and in the sphere of metaphorical rethinking the poetic world is relatively stable over time, but on the other hand, a number of significant changes are taking place.

The distribution of Source and Target themes frequencies, ranked in descending order, was fitted using an exponential function plus 1, which has been successfully applied in a number of studies (Mistecky, Altmann 2019; Kelih 2024):

$$y = 1 + a * e^{-b*x},$$

where *a* and *b* – parameters.

In our study this formula was used in all cases except one (in collection RTB for Target ThC) where the fitting was not successful. But applying the exponential function without added 1 improved the result. Thus in all cases the fitting yielded good results. Generally speaking, somewhat better results can be obtained using other functions, for example the Zipf-Alekseev function (Rácová et al. 2019), but it contains more parameters (three) while we tried to minimize their number.

The results of the fitting are presented in Table 4 and in Fig. 1-8. More detailed data about the fitting can be found in the appendix. Classes with zero frequency values were skipped.

**Table 4:** Fitting the exponential function and the exponential function plus one to the ranked distribution of relative frequencies of thematic classes.

| **TW: Target** | **BN: Target** | **RTB: Target** | **PM: Target** |
|---|---|---|---|
| r2 = 0.91<br>a = 23.33<br>b = 0.41 | r2=0.97<br>a=11.98<br>b=0.23 | r2=0.97<br>a=3.70<br>b=0.11 | r2=0.90<br>a=5.00<br>b=0.31 |
| **TW: Source** | **BN: Source** | **RTB: Source** | **PM: Source** |
| r2=0.96<br>a=140.03<br>b=1.57 | r2=0.93<br>a=59.04<br>b=0.96 | r2=0.96<br>a=57.93<br>b=1.74 | r2=0.95<br>a=47.55<br>b=1.74 |

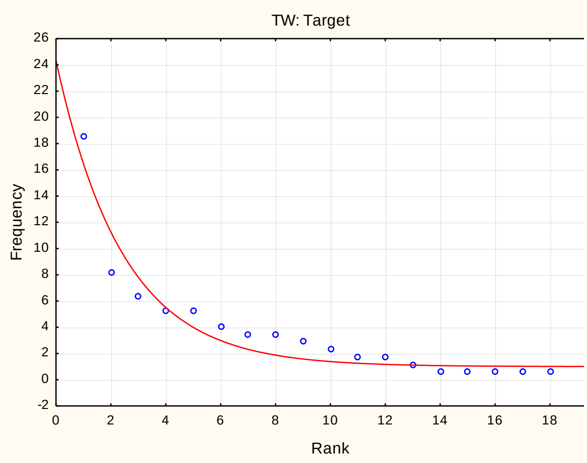

**Figure 1**

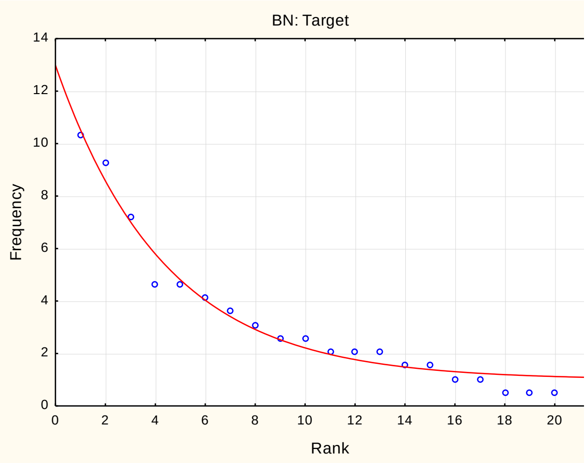

**Figure 2**

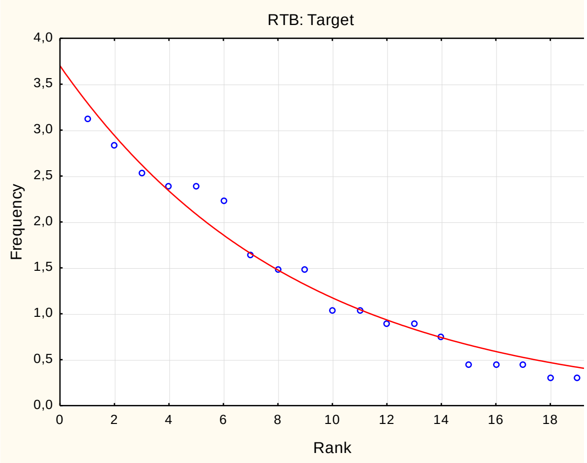

**Figure 3**

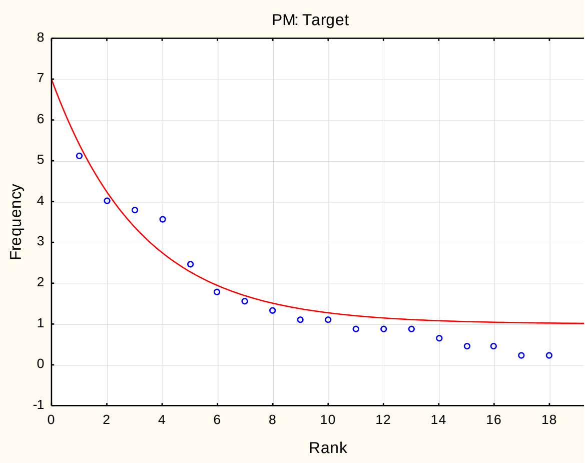

**Figure 4**

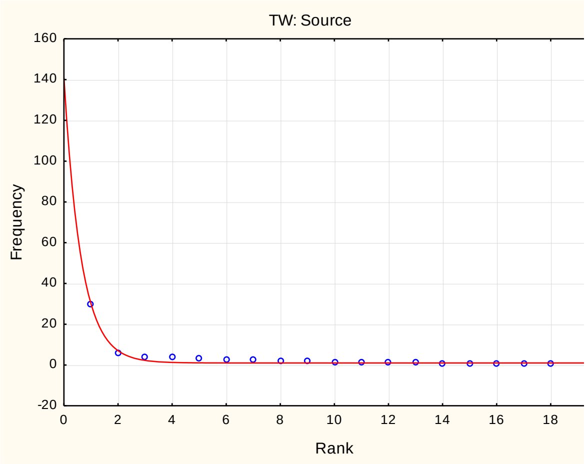

**Figure 5**

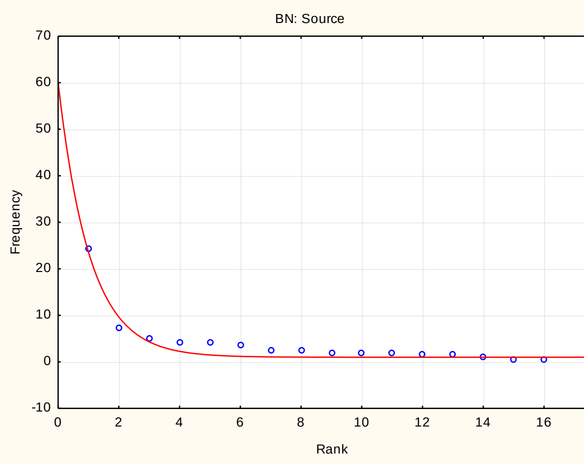

**Figure 6**

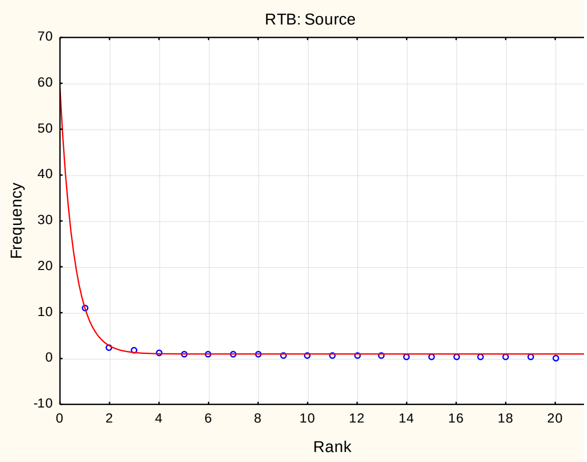

**Figure 7**

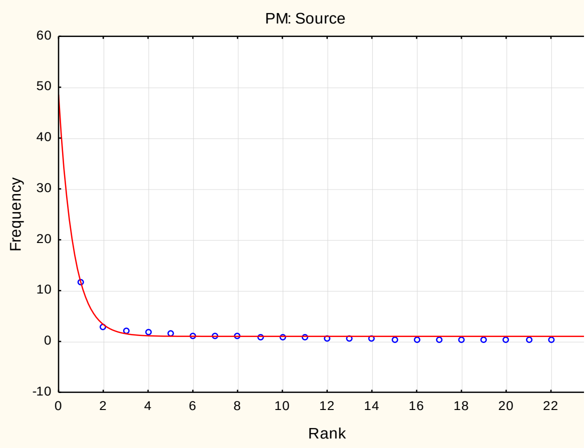

**Figure 8**

As seen from the data in Table 3, the parameter b, which reflects the rate of change of the function, is very different for the Target thematic classes, on the one hand, and for the Source domain, on the other hand. In the first case, b is significantly lower than 1 and reflects the tendency for a relatively slow (in Fig.3 very slow) decrease in the function. For the Source, the parameter b is greater than 1 in most cases (for BN it is equal to 1), reflecting a faster decay of the function. This is clearly visible in the figures which show the low and fast decay of the curve for the Source and the Target, respectively.

It seems appropriate to consider how these distributions affect the composition of thematic cores and their characteristics. In order to determine the core (the upper tail of the frequency curve) in the frequency list of thematic classes we use the method based on the Hirsh point (*h*) which was proposed for the purpose of determining the lexical nucleus in linguistic studies in I-I. Popescu (Popescu 2007) and has been used in a number of studies (Kubát, Čech 2016; Popescu, Altmann 2006; Popescu et al. 2007). According to this approach, the rank that coincides with the frequency of the given unit is chosen as the threshold. In cases where there is no complete coincidence of the values of rank and frequency, the following criterion proposed for such cases can be applied (Čech, Kubát 2016, p. 8; Kubát, Čech 2016, p. 152):

$$h = \frac{f(r_i)r_{i+1} - f(r_{i+1})r_i}{r_{i+1} - r_i + f(r_i) - f(r_{i+1})},$$

where *h* is the Hirsch point, *r* is the rank, *f(r)* is the frequency of a given rank ThC, $r_i$ is the highest number for which $r_i < f(r_i)$ and $r_{i+1}$ is the lowest number for which $r_{i+1} > f(r_{i+1})$.

For each core an analysis of thematic concentration (TC) was performed applying the formula (Čech, Kubát 2016, p. 8; Kubát, Čech 2016, p. 152; Popescu 2007):

$$TC = 2 * \sum_{r'=1}^{m} \frac{(h - r') * f(r')}{h(h-1) * f(1)},$$

where *r'* are ranks smaller than *h*, *f(r')* are frequencies corresponding to these ranks, *f*(1) is the highest frequency, *h* is the value of the Hirsch point, and *m* is the number of ThCs with *r* < *h*.

The thematic concentration analysis allows one to quantify the ratio of frequencies of the core elements, necessary for revealing the author's systematic preferences in constructing their poetic worldview.

It should be noted that this concentration indicator is usually used for words while in this paper we use it to analyze more general entities, lexical (thematic) classes of words.

In our case thematic concentration serves as a measure of homogeneity among thematic classes in the frequency core.

As a result, the following data presented in Table 5 were obtained.

**Table 5:** H-point and concentration index in the core.

| | **Target** | | **Source** | |
|---|---|---|---|---|
| | **H-point** | **TC** | **H-point** | **TC** |
| **TW** | 6.6 | 0.55 | 5.6 | 0.46 |
| **BN** | 7 | 0.76 | 6.43 | 0.47 |
| **RTB** | 9 | 0.82 | 7 | 0.40 |
| **PM** | 7 | 0.77 | 6 | 0.47 |

At the final stage of the analysis, we considered the degree of similarity of the collections in the space of the full list of parameters, using the Euclidean distance:

$$d_{(p,q)}= \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2},$$

where $p$ and $q$ are points in n-dimensional space.

The distances between adjacent collections are graphically reflected by the histogram in Fig. 9. Judging by the height of the histogram columns reflecting the distances between collections, one can draw conclusions about the degree of changes in the representation of thematic classes across collections.

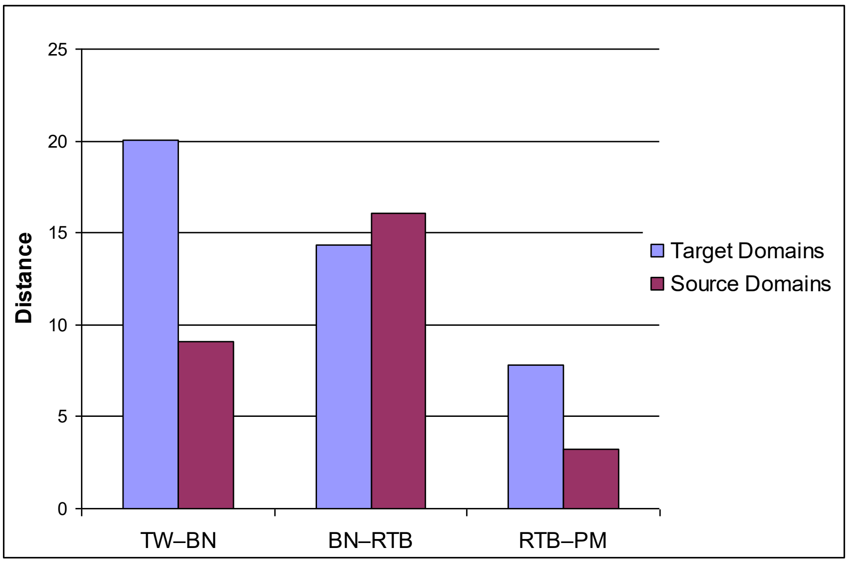

**Figure 9:** Distances between collections adjacent in time.

## 3 Discussion

The analysis reveals significant divergence between Target and Source domain core structures: Target domains exhibit greater frequency homogeneity among thematic classes, while Source domains show marked frequency disparities, particularly between the dominant Living being class and other core elements. While Living being metaphors constitute the most frequent cross-author pattern, the magnitude of intra-core frequency variation differs substantially. We interpret this variability as an objective signature of individual authors' figurative systems.

The disparity is particularly pronounced in RTB, where the Target domain's concentration index (0.82) doubles that of the Source domain (0.40), indicating fundamentally different organizational principles in their core structures. In the youth collection (TW) this tendency is somewhat less expressed. If we consider the concentration indicators of themes for the Target and the Source separately, we can see that in the Source domain the concentration of themes in the core initially increases, and in the later collection decreases to approximately the average values of the Berlin period.

In early lyrics Nabokov's core of the Target domain reflects his interest in the metaphorical interpretation of nature (Plant, Space, Light, Substance), where the typical locus is a sunlit meadow with sparse trees or a grove on a summer day.

Human being is represented by his material and mental features (Part of body, Mental phenomenon). In the Source domain Living being tops the list of the most frequent thematic classes. The remaining thematic classes of the core (Area, Light, Substance) perform the functions of both the Target and the Source.

The beginning of the Berlin period is associated with a shift in the author's focus on human being. The number of representatives of the human world grows, including Transport.

In the Source domain, the concentration is stable, with the exception of RTB. The themes that are represented here to the greatest extent are those covering the human micro-world (Part of body, Instrument, Living being, Transport, etc.).

The Source core remains sufficiently stable due to the ThCs of the Living being, Light and Area. The concentration of the Source fluctuates within a narrow range of values.

Changes in the frequencies of thematic classes in the Source and Target domains are quite pronounced, but manifest themselves differently at different stages.

The cores of thematic classes, identified using the Hirsch point, include mostly 6-7 elements. Two different trends are observed in the concentration of thematic classes in the core. The concentration of the Target increases, reaching a peak by the end of the 1920s. The maximum changes in the frequencies of Target domain ThCs are observed between TW, the early collection, and BN, the first poetry collection of the Berlin period. During the Berlin period, changes continue but are less pronounced, and after the Berlin period are significantly reduced. The Source is characterized by a more stable level of concentration.

Among the Source domain thematic classes, the peak of changes occurred during the Berlin period itself (BN–RTB). The changes between the first and second collections under study are less than this peak value (16.07) in the Berlin period and are clearly inferior to the changes recorded in the Target Sphere between TW and BN. Furthermore, between the final collection of the Berlin period and the subsequent collection, the changes in the Source domain are very minor. Thus, here both domains of thematic features behave identically showing minimal changes.

It is quite obvious that Nabokov's poetic style was formed towards the end of his Berlin period and his worldview changed little thereafter. The development of the Target domain was more intensive and began earlier, the development of the Source domain became more active towards the last poetry collection of the Berlin period already after the appearance of prose works.

# 4 Conclusion

Analysis of the figurative system of Nabokov's lyrics has shown a number of tendencies in the structure of images and their development over time. Thematic classes filling the positions of Target and Source in the image model in each collection differ significantly in their hierarchical structure.

Small values of the correlation between thematic classes in the Target and Source domains shows that the description of the world in Nabokov's poetry required an asymmetry in the hierarchical organization. At the same time, correlation of the thematic classes of the Target and of the Source across different collections reveal certain continuity in Nabokov’s poetic images over time. On the other hand the frequencies of the Target and Source thematic classes vary across collections.

The study of the similarity of images in different collections according to the frequencies of thematic classes in the Target and Source domains made it possible to identify the main stages of changes. The greatest changes for the Target ThCs occurred between the earliest collection under study and the collection of the beginning of the Berlin period. Among the Source themes, the peak of change occurred during the Berlin period itself (BN–RTB). Nabokov's style changed until the end of the Berlin period, moving away from the model established in his youth. Berlin was at that time one of the centers of the Russian émigré creative intelligentsia and it was precisely during the Berlin period, by the end of the 1920s, Nabokov’s poetic worldview was formed.

The distribution of thematic classes in both positions is well fitted by the exponential function plus 1 and the exponential function without the added 1. For the Target themes the graph declines much more smoothly than for the Source.

The research has contributed to a better understanding of the individual style of Nabokov as a poet. The findings of this research show the importance of investigating poetry written by the famous prose writer which can provide additional material for the study of periodization of Nabokov’s creative activity, the dynamics of style not only in poetry but also in prose, and will help reveal the main trends in his style alteration.

The findings, however, remain intermediate. Future research of semantic classes frequency in prose compared to the presented data may provide better prospects for research into Nabokov’s style.

# References

**Andreev, V.** (2012). Analysis of Imagery System Development in the Individual Style of E.A. Poe. In: *Language, Narrative and the New Media. Book of Abstracts of the Poetics and Linguistics Association International Conference (PALA 2012), Msida, 2012*, p. 4. University of Malta.

**Boyd, B.** (1999). *Nabokov's Pale fire: the magic of artistic discovery*. Princeton, NJ: Princeton University Press.

**Čech, R., Kubát, M.** (2016). Text length and the thematic concentration of text. *Mathematical Linguistics*, 1(2), pp. 5–13.

**De Vries, G.** (1991). Some Remarks on Nabokovs Pale Fire. *Russian Literature Triquarterly*, 24, pp. 239–267.

**Kelih, E.** (2024). Modelling the frequency of loanwords in different semantic fields in core vocabularies (based on WOLD data). *Glottometrics*, 56, pp. 59–77.

**Kubát, M., Čech, R.** (2016). Thematic concentration and vocabulary richness. In: Kelih, E., Knight, R., Mačutek, J., Wilson, A. (Eds.). *Issues in quantitative linguistics*, 4, pp, 150–159. Lüdenscheid: RAM-Verlag.

**Místecký, M., Altmann, G.** (2019). Tense and person in English: modelling attempts. *Glottometrics*, 46, pp. 98–104.

**Morris, P.D.** (2010). *Vladimir Nabokov: poetry and the lyric voice*. Toronto; Buffalo: University of Toronto.

**Pavlovich, N.V.** (1999). *Slovar' poeticheskikh obrazov: Na materiale russkoy khudozhestvennoy literatury XVIII–XX vekov* [Dictionary of poetic images: On the material of Russian literature of the XVIII–XX centuries]. Moscow: Editorial URSS.

**Pavlovich, N.V.** (2004). *Yazyk obrazov: paradigmy obrazov v russkom poeticheskom yazyke. Izd 2.* [The Language of Images: Paradigms of Images in Russian Poetic Language. 2nd edition]. Moscow: Azbukovnik.

**Popescu, I.-I.** (2007). Text ranking by the weight of highly frequent words. In: Köhler, R., Grzybek, P (Eds.). *Exact methods in the study of language and text,* pp. 555–565. Berlin/New-York: Mouton de Gruyter.

**Popescu, I.-I., Altmann, G.** (2006). Some aspects of word frequencies. Glottometrics 13, 23–46.

**Popescu, I.-I., Best, K.-H., Altmann, G.** (2007). On the dynamics of word classes in text. *Glottometrics*, 14, pp. 58–71.

**Rácová, A., Zöring, P., Altmann, G.** (2019). Syllable Structure in Romani: A Statistical Investigation. *Glottotheory*, 46, pp. 41–60.

# Appendix

Fitting the exponential function plus one to the ranked distribution of thematic relative frequencies of thematic classes

| **TW: Target** | | | **BN: Target** | | | **RTB: Target** | | | **PM: Target** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Plant | 18.60 | 16.48 | Liv-B | 10.31 | 10.53 | H-hold | 3.13 | 3.30 | Liv-B | 5.13 | 5.41 |
| Area | 8.14 | 11.27 | PoB | 9.28 | 8.58 | Area | 2.83 | 2.94 | Social | 4.02 | 4.24 |
| PoB | 6.40 | 7.81 | Light | 7.22 | 7.03 | PoB | 2.53 | 2.62 | Exist | 3.79 | 3.38 |
| Ment | 5.23 | 5.52 | Tr-port | 4.64 | 5.79 | Ment | 2.38 | 2.34 | Inf | 3.57 | 2.75 |
| Light | 5.23 | 4.00 | D-ness | 4.64 | 4.81 | Light | 2.38 | 2.08 | Area | 2.46 | 2.28 |
| Nat | 4.07 | 2.99 | Area | 4.12 | 4.03 | Audit | 2.24 | 1.86 | Ment | 1.79 | 1.94 |
| Time | 3.49 | 2.32 | Plant | 3.61 | 3.41 | Time | 1.64 | 1.66 | Audit | 1.56 | 1.69 |
| Audit | 3.49 | 1.87 | Time | 3.09 | 2.92 | Plant | 1.49 | 1.48 | PoB | 1.34 | 1.51 |
| Instr | 2.91 | 1.58 | Liq-ds | 2.58 | 2.53 | Tr-port | 1.49 | 1.32 | Time | 1.12 | 1.37 |
| Liq-ds | 2.33 | 1.38 | Audit | 2.58 | 2.21 | Liv-B | 1.04 | 1.17 | Plant | 1.12 | 1.28 |
| Sbst | 1.74 | 1.26 | Sbst | 2.06 | 1.96 | Exist | 1.04 | 1.05 | Sbst | 0.89 | 1.20 |
| Liv-B | 1.74 | 1.17 | Inf | 2.06 | 1.77 | Sbst | 0.89 | 0.93 | Liq-ds | 0.89 | 1.15 |
| Inf | 1.16 | 1.11 | Ment | 2.06 | 1.61 | Instr | 0.89 | 0.83 | Light | 0.89 | 1.11 |
| J-ry | 0.58 | 1.07 | Food | 1.55 | 1.49 | D-ness | 0.75 | 0.74 | H-hold | 0.67 | 1.08 |
| H-hold | 0.58 | 1.05 | Exist | 1.55 | 1.39 | Liq-ds | 0.45 | 0.66 | Cls | 0.45 | 1.06 |
| Social | 0.58 | 1.03 | H-hold | 1.03 | 1.31 | Inf | 0.45 | 0.59 | D-ness | 0.45 | 1.04 |
| Tr-port | 0.58 | 1.02 | Cls | 1.03 | 1.24 | Social | 0.45 | 0.53 | J-ry | 0.22 | 1.03 |
| Exist | 0.58 | 1.01 | Cont | 0.52 | 1.19 | Nat | 0.30 | 0.47 | Instr | 0.22 | 1.02 |
| Food | 0.00 | - | Instr | 0.52 | 1.15 | Cls | 0.30 | 0.42 | Food | 0.00 | - |
| Cont | 0.00 | - | Nat | 0.52 | 1.12 | J-ry | 0.00 | - | Cont | 0.00 | - |
| Cls | 0.00 | - | J-ry | 0.00 | - | Food | 0.00 | - | Nat | 0.00 | - |
| D-ness | 0.00 | - | Social | 0.00 | - | Cont | 0.00 | - | Tr-port | 0.00 | - |
| Fire | 0.00 | - | Fire | 0.00 | - | Fire | 0.00 | - | Fire | 0.00 | - |
| r2 = 0.91<br>a = 23.33<br>b = 0.41 | | | r2=0.97<br>a=11.98<br>b=0.23 | | | r2=0.97<br>a=3.70<br>b=0.11 | | | r2=0.90<br>a=5.00<br>b=0.31 | | |

| **TW: Source** | | | **BN: Source** | | | **RTB: Source** | | | **PM: Source** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Liv-B | 30.23 | 30.09 | Liv-B | 24.23 | 23.60 | Liv-B | 11.18 | 11.15 | Liv-B | 11.61 | 11.55 |
| Area | 5.81 | 7.04 | Light | 7.22 | 9.65 | Light | 2.53 | 2.78 | Area | 2.90 | 3.34 |
| Sbst | 4.07 | 2.25 | Sbst | 5.15 | 4.31 | Area | 1.94 | 1.31 | Sbst | 2.23 | 1.52 |
| Light | 4.07 | 1.26 | J-ry | 4.12 | 2.27 | Fire | 1.34 | 1.05 | Plant | 1.79 | 1.12 |
| H-hold | 3.49 | 1.05 | Ment | 4.12 | 1.49 | Liq-ds | 1.04 | 1.01 | Ment | 1.56 | 1.03 |
| J-ry | 2.91 | 1.01 | Area | 3.61 | 1.19 | PoB | 1.04 | 1.00 | Inf | 1.12 | 1.01 |
| Cls | 2.91 | 1.00 | Plant | 2.58 | 1.07 | Cls | 1.04 | 1.00 | Light | 1.12 | 1.00 |
| Liq-ds | 2.33 | 1.00 | Cls | 2.58 | 1.03 | Ment | 0.89 | 1.00 | Cls | 1.12 | 1.00 |
| Cont | 2.33 | 1.00 | Liq-ds | 2.06 | 1.01 | Sbst | 0.75 | 1.00 | Fire | 0.89 | 1.00 |
| Audit | 1.74 | 1.00 | Inf | 2.06 | 1.00 | Instr | 0.75 | 1.00 | PoB | 0.89 | 1.00 |
| Fire | 1.74 | 1.00 | PoB | 2.06 | 1.00 | Tr-port | 0.75 | 1.00 | Tr-port | 0.89 | 1.00 |
| Ment | 1.74 | 1.00 | H-hold | 1.55 | 1.00 | Audit | 0.60 | 1.00 | J-ry | 0.67 | 1.00 |
| Time | 1.16 | 1.00 | D-ness | 1.55 | 1.00 | Cont | 0.60 | 1.00 | Cont | 0.67 | 1.00 |
| Inf | 0.58 | 1.00 | Food | 1.03 | 1.00 | J-ry | 0.45 | 1.00 | D-ness | 0.67 | 1.00 |
| PoB | 0.58 | 1.00 | Cont | 0.52 | 1.00 | H-hold | 0.45 | 1.00 | Liq-ds | 0.45 | 1.00 |
| Plant | 0.58 | 1.00 | Fire | 0.52 | 1.00 | Food | 0.30 | 1.00 | Food | 0.45 | 1.00 |
| Social | 0.58 | 1.00 | Time | 0.00 | - | Plant | 0.30 | 1.00 | Instr | 0.45 | 1.00 |
| D-ness | 0.58 | 1.00 | Audit | 0.00 | - | Nat | 0.30 | 1.00 | Time | 0.22 | 1.00 |
| Food | 0.00 | - | Instr | 0.00 | - | Social | 0.30 | 1.00 | Audit | 0.22 | 1.00 |
| Instr | 0.00 | - | Nat | 0.00 | - | D-ness | 0.15 | 1.00 | H-hold | 0.22 | 1.00 |
| Nat | 0.00 | - | Social | 0.00 | - | Time | 0.00 | - | Nat | 0.22 | 1.00 |
| Tr-port | 0.00 | - | Tr-port | 0.00 | - | Inf | 0.00 | - | Social | 0.22 | 1.00 |
| Exist | 0.00 | - | Exist | 0.00 | - | Exist | 0.00 | - | Exist | 0.00 | - |
| r2=0.96<br>a=140.03<br>b=1.57 | | | r2=0.93<br>a=59.04<br>b=0.96 | | | r2=0.96<br>a=57.93<br>b=1.74 | | | r2=0.95<br>a=47.55<br>b=1.74 | | |