UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE POLITICHE,
ECONOMICHE E SOCIALI

Master's Degree Course in Data Science and Economics

THE DISTRIBUTION OF DEPENDENCY DISTANCES

Thesis by: Sonia Petrini

Supervisor: Prof. Nicolò Cesa-Bianchi

External Supervisor: Prof. Ramon Ferrer-i-Cancho

Academic Year: 2022/2023

"The limits of my language mean the limits of my world."

– Ludwig Wittgenstein,
Tractatus logigo-philosphicus,
1922

Acknowledgments

This work marks the finish line of an incredible journey. A journey through myself, through knowledge, discovery, science, through friendship, through Europe. There are so many levels on which the past two years have influenced and shaped me. So many people to thank. So many things I have understood about myself and about the world around me. My heart is full of emotions while writing these lines, and I want to express them all to do justice to each single wonderful person that walked this journey with me.
Firstly and fundamentally, I want to thank my family. Thank you mom and dad for believing in me, and for all the love and support that you did and that you endlessly give to me. You provided me with the means to study what I love, to do what I love and to live in a place I love, and I am profoundly grateful for this. If I feel free it's because I always have you behind my back.
Thank you Linda, for being an inexhaustible source of inspiration. I believe you don't fully realize how much admiration, love, and respect I have for you, and for how you are dwelling in your own journey. You are amazing, and I have faith in you. Also, thank you for 'lending' me the shirt with the zebras, which I have been wearing to every oral test, from the admission interview to the last exam.
Then, I want to thank who was on the same boat with me, sailing the sea of Data Science. Even during the pandemics we had the luck to be able to attend lectures in person (thank you University of Milan!). In this context, something magical happened. A curiously assorted group of people with the most diverse backgrounds put together all their knowledge of economics, mathematics, informatics, communication, political science, international relations science, and so on, creating a fun and interesting environment in which to learn from one another, and with each other. Thank you Matteo for sharing with all of us your knowledge, for always being there for me, for answering all my stupid or smart questions, for being a great friend. Thank you Elisa for all the adventures, the laughs, the train rides, the mornings after a party in which you would wake up with the ticking of my fingers on the keyboard. Thank you Luca for being critical, for all the animated and interesting discussions, for suggesting me a book that deeply changed my perspectives. Thank you Rob for being who you are, for always making me feel at home, for all the dancing and the late night talks about life. Thank you Gas, Liz, Ruben, Sara, Mathias, Lorenzo, Margherita, Andrea...
In this last year I have also sailed with another crew, over the seas of Barcelona. One day I left and I did not come back, finding myself in yet another loving and inspiring environment. First, I want to thank both University of Milan and Universitat Politécnica de Catalunya for giving me this opportunity. In particular, thank you Ramon for being a wonderful professor and human being, for introducing me to the magical world of quantitative linguistics, for inspiring and guiding me, for opening to me the doors of research, for welcoming me in your team, and for the positive energy that you always bring with you. Thank you Blanche, for taking care of me in my endless nights and days of studying, always providing me with support and sweetness (and chocolate croissants). Thank you Lennart for all your smart comments and insights, for having offered me a space for confrontation and sharing of ideas, for

having listened to all my stressed talks. Thank you Gio for looking after me, for all the music, the talks, the positive vibes. Thank you Vincenzo for the skate rides, the encouragement, the adventures, for being my fan and my friend. Thank you Eloy for the walks around Barcelona, for being funny and sweet, for being my friend. Thank you LLuís for being a mentor to me, for your help, your availability, your kindness, your patience (and for your computational power). Thank you Rui and Alexandra for the fun lunches together. Thank you Kevin, for being so busy and yet finding time to care about other people. Thank you Alessia, for being part of why I feel at home here, and for being a great person. Thank you Barcelona for offering me everything I feel like I need right now, for being so colorful and warm, for the skate rides to the sea and the amount of diversity that you attract.
Finally, I want to thank who has always been on my side. Thank you Mattia, Laetitia, Alessandro, and Giulia for being my pillars. Even while in different countries never have I ever felt like we were really distant, your incessant support always has been and still is tangible. Thank your making me feel loved, and for being happy of my happiness and of my success like I am of yours.
Actually, I feel like there is someone else who deserves to be mentioned here. I want to thank myself. For all the efforts, the enthusiasm, the passion, the hard work, the times in which I could go out but I preferred to stay home with my computer, the struggles. In fact, this journey has not been easy. Many times I have felt frustrated, overwhelmed, not good enough, lost, or in the wrong place. But I made it and I feel proud of it, proud of myself. And I feel like I am exactly in the place where I should be, where I want to be.

ABSTRACT
The syntactic structure of a sentence can be represented as a graph where vertices are words and edges indicate syntactic dependencies between them. In this setting, the distance between two syntactically linked words can be defined as the difference between their positions. Here we want to contribute to the characterization of the actual distribution of syntactic dependency distances, and unveil its relationship with short-term memory limitations. We propose a new double-exponential model in which the decay in probability is allowed to change after a break-point. This transition could mirror the transition from the processing of words chunks to higher-level structures. We find that a two-regime model – where the first regime follows either an exponential or a power-law decay – is the most likely one in all 20 languages we considered, independently of sentence length and annotation style. Moreover, the break-point is fairly stable across languages and averages values of 4-5 words, suggesting that the amount of words that can be simultaneously processed abstracts from the specific language to a high degree. Finally, we give an account of the relation between the best estimated model and the closeness of syntactic dependencies, as measured by a recently introduced optimality score.

Contents

1. Introduction

Language is one of the most complex and fascinating expressions of humans as social animals, stemming from the antithesis between our urge for communication and physical and cognitive limitations existing on memory. The interaction between these two forces inevitably shapes language in all its aspects (Liu et al., 2017). Among them we here focus on syntax, namely the way in which words in a sentence compose into larger hierarchical structures, creating a parallel dimension to their plain linear arrangement. The hierarchical structure arises from the relations between words and their heads, modeled by means of a directed edge in the one-dimensional network of a sentence. We call the resulting structure a syntactic dependency tree: each vertex is a word, and each word – besides the root – depends syntactically from its head, to which it is connected by an edge. We define ***d*** as the absolute value of the difference between the position of two syntactically related words. For instance, in figure 1 'John' and 'gave' are at distance 1, 'gave' and 'painting' are at distance 3, and so on.

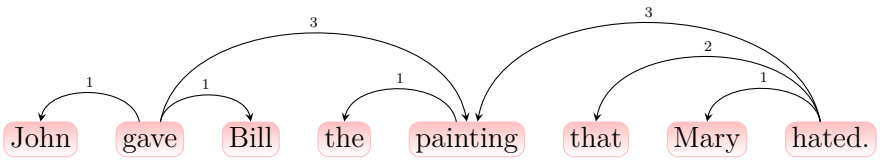


Figure 1. Example of syntactic dependency tree. Edges are labelled with the value of the syntactic dependency distance between the words they connect.

A well-established principle of Dependency Distance minimization (DDm) has been consistently found in languages, implying the preference for short dependencies (Ferrer-i-Cancho, 2004; Futrell et al., 2015; Gildea and Temperley, 2007; Liu, 2007). This phenomenon has first been related to the cognitive realm with Yngve's depth hypothesis (Yngve, 1960), and the relation between optimized syntactic structures and short-term memory (STM) limitations has been object of study since then. Moreover, the classical study by Miller has set the grounds for investigations on such STM limitation, in particular concerning its quantification and universality (Miller, 1956). Let us first consider the relevant literature on the distribution of dependency distances, to then move to existing views on STM constraints and their coping mechanisms. Finally, we will discuss how the importance of the contribution of this work emerges from the interaction between the two.

1.1. On the distribution of dependency distances

The large body of evidence in favor of DDm suggests that there are universal patterns underlying language structure, which abstract from the specific language, and are rather related to the functioning of the brain. Understanding the probability distribution of dependency distances, being it an echo of our brains functioning, could shed light on these patterns (Liu et al., 2017). Ferrer-i-Cancho – moving from an

entropy maximization argument – described a dependency's probability as an exponentially decaying function of distance for sentences of fixed length in Czech and Romanian (Ferrer-i-Cancho, 2004). He also made an interesting observation concerning a change in the speed of the decay: the probability of observing a dependency at distance 4-5 or more is higher than expected, in the sense that the decay slows down, which is apparently in contrast with the minimization principle itself. Later on, Liu proposed a power-law behaviour to describe the distribution of dependency distances in a Chinese treebank, considering sentences of mixed length (Liu, 2007). A later cross-linguistical study covering 30 languages identified a power-law distribution for long sentences, and an exponential trend in short ones (Haitao, 2016), conveying the complexity of the analysed problem. Nevertheless, all these distributions have a similar shape, characterized by the dominance of very short distances and a long tail (Jiang and Liu, 2015). The observed differences could hence derive from systematic discrepancies in sentence lengths, context, and annotation style, which all influence syntactic dependency distances (Jiang and Liu, 2015). Moreover, power-laws can emerge from mixing other distributions (Stumpf and Porter, 2012), for instance from differently parameterized exponentials (Ferrer-i Cancho and Liu, 2014). Hence the need – expressed in various studies (Ferrer-i-Cancho, 2004; Ferrer-i Cancho and Liu, 2014; Jiang and Liu, 2015) – to find the common ground of these results, analyzing the distribution of dependency distances while accounting for all these factors: considering both mixed and fixed sentence lengths in a large enough parallel corpus, while also controlling for annotation style. In particular, in the attempt to identify the real shape of such distribution, we build on the peculiar phenomenon observed in (Ferrer-i-Cancho, 2004) to propose a double-exponential model, in which the speed of exponential decay changes after a break-point. The theoretical motivation for this model shall be explained below.

1.2. Short-term memory limitations

The pioneering work by Miller has set the grounds for research on a possible absolute constraint on the amount of information that can be temporarily stored in memory, and on the mechanisms enacted to cope with it (Miller, 1956). The questions arising are multiple: does such limit exist? If it does, what is its span? Is it absolute or task-specific? Where does it originate from? Is it a storage-based limit, or is it due to other factors? Despite the different views existing on each of these questions, there is common agreement that a constraint actually exists (Christiansen and Chater, 2015; Cowan, 2001; Henderson, 1972; Lewis and Vasishth, 2005; Miller, 1956), and that it is understood in terms of information chunks rather than single units. This latter point could partly explain the inconsistency between the estimated values of this maximum span, from $\mathbf{7 \pm 2}$ (Miller, 1956), to $\mathbf{2 - 3}$ (Lewis and Vasishth, 2005), to $\mathbf{4}$ (Cowan, 2001). In fact, while it is true that different limitations could arise from different contexts and tasks, one cannot exclude that the observed limitations all underlie the same cognitive mechanisms, which manifest differently based on the analysed stimulus, due to different ways of organizing information. Indeed, in experimental settings it is likely to be difficult to distinguish the real maximum span from the observed one: the latter could be an artifact of binding chunks of information together, thus

resulting in the ability to recall larger amount of units, while the amount of chunks is the same (Cowan, 2001). These considerations seem to be particularly relevant in the scope of linguistic communication: communicating requires constantly receiving and processing new inputs, without losing reference to the previous ones. Let an open dependency be one in which only one of the two elements that compose it has already appeared, and a close dependency one in which both the head and the dependent have already been encountered. Then, in the context of dependency structure the success of communication depends on the ability to keep track of an open dependency while opening new ones, and without knowing a priori when it is going to be closed (Liu et al., 2017). Notice that dependencies clearly represent relations between words, which are necessary for the speaker to convey a complex message building it from smaller units (encoding), and for the listener to recover such message by understanding the underpinning structure of the sequence of words (decoding). Thus, in some form syntactic structure really reveals the way in which humans deal with physical limitations to be able to produce and process a possibly infinite amount of words. Christiansen and Chater provided an integrated framework to describe both the perceptual constraints affecting STM in language processing – what they call the "now-or-never bottleneck" – and the chunking strategy enacted to cope with them, which they refer to as "chunk-and-pass" mechanism (Christiansen and Chater, 2015). They collected a wide set of empirical results, describing the bottleneck as mainly arising from our short memory for auditory signals, the speed of new incoming linguistic input, and from memory limitations on sequence recalling tasks. According to the authors, to deal with these constraints the human cognitive system relies on a series of strategies. That is, as we receive new linguistic input, we eagerly process it by grouping units into chunks, and passing them at a more abstract level of representation; once a chunk has been integrated in the existing knowledge hierarchy, a new one can be processed and again passed at higher representation levels. This model entails that chunking is required to permanently store information: while a single word would be an easily forgotten piece of de-contextualized information, grouping words together produces a meaningful abstract image, which can be related to the following incoming concept. This mechanism would thus guarantee effective and efficient communication, profoundly shaping the structure of language itself.

1.3. Aim and structure

The primary aim of this work is to test the hypothesis that dependency distances in real languages are distributed following two exponential regimes, modeled by means of a two-regime geometric distribution (both with and without a right-truncation). The proposal of two-regimes is motivated both empirically and theoretically. On one hand, it builds on the observations by Ferrer-i-Cancho concerning a change in probabilistic decay (Ferrer-i-Cancho, 2004). On the other hand, the existence of two different regimes would be consistent with the widely accepted idea that words are being chunked in order to be processed (Christiansen and Chater, 2015). Indeed, in a commentary on the work by Christiansen and Chater Ferrer-i-Cancho had suggested a link between his empirical observation and their processing framework (Ferrer-i-Cancho, 2017). Verifying this hypothesis opens the path for a deeper understanding

of both previous results on the distribution of dependency distances, and on the existence and universality of memory constraints. Concerning the first point, we believe our work will contribute to the existing literature on the distribution of dependency distances, finding a common ground to previous results by accounting for each of sentence length, context, and annotation style. In fact, we consider both the syntactical structure of sentences with a specific length, and of various sentence lengths jointly, performing the analysis on a parallel corpus following two alternative syntactic dependency annotation schemes. The second point is related to one of the free parameters of our models, namely the break-point between the two regimes. If the change in probability is a mirror of the chunking mechanism enacted in language processing, the break-point may approximate the distance after which the burden of a new dependency is not offset by its contribution to the understanding of the chunk. Put differently, the chunks we divide a sentence in to overcome STM limitations still need to contain enough information so as to acquire an incremental understanding. Thus, on one hand we would like to simultaneously process the largest possible amount of words, on the other hand, our fleeting memory and the interference created by intermediate words constrain the chunk size. In our view, the change in probability decay may represent the balance found between these two forces (understanding and processing difficulty). Therefore, looking at the stability of the estimated break-point values could shed light on common patterns related to STM limitations in languages. Formally, we aim at verifying the following two-fold hypothesis:

- $\boldsymbol{H_1}$: Dependency distances are distributed following two exponential regimes.
- $\boldsymbol{H_2}$: The break-point between the two regimes is stable across languages.

Additionally, we present an analysis of the relation between the best estimated model and the optimality of dependency distances, quantified through the recently introduced optimality score $\boldsymbol{\Omega}$ (Ferrer-i Cancho et al., 2022). Combining these two sources of information allows us on one hand to verify the results of model selection, on the other to reproduce patterns concerning the different strength of DDm depending on sentence length, which have already been observed in a previous study (Ferrer-i Cancho and Gómez-Rodríguez, 2021). In fact, the pressure for minimization of syntactic dependency distances has been found to be weaker in short sentences, in which the burden of long dependencies is lighter, so that DDm could be surpassed by other word order principles. The paper is structured as follows: section 2 provides the definitions of the models for the distribution of dependency distances, including the formalization of the proposed two-regime models. In section 3, we report the results from model selection, and from a complementary analysis on the relation between the best model and the optimality of dependency distances. In section 4, we discuss the obtained results, first focusing on the verification of our hypotheses and on other emerged global patterns, then accounting for the observed cross-linguistic variability. Section 5 summarises the main findings of the paper. Finally, in section 6 we provide a description of the data, while in section 7 we report on the employed methodology, concerning the process of artificial samples generation, the implementation of the model selection procedure, and the computation of the optimality score $\boldsymbol{\Omega}$. The code for this work was written both in `R` and `python`, and is available here.

2. Models

In order to test $\boldsymbol{H_1}$ we compare the proposed two-regime model against an ensemble of models, carrying out a model selection procedure as described in section 7.2. Let us introduce the models included in the ensemble. As mentioned above, $\boldsymbol{d}$ is the distance between linked words. Here we borrow the convention that consecutive words are at distance $\mathbf{1}$, so that words separated by an intermediate word are at distance $\mathbf{2}$ and so on (Ferrer-i-Cancho, 2004). Hence $\boldsymbol{d} \in [\mathbf{1}, \boldsymbol{n}-\mathbf{1}]$, where $\boldsymbol{n}$ is the number of words in a sentence. We use $\boldsymbol{p(d)}$ to refer to the probability that two linked words are at distance $\boldsymbol{d}$. Table 1 summarises the ensemble of models, while figure 2 displays the shape of the models against an artificial random sample of their probability distributions. The methodology employed to generate the samples is described in section 7.
The first model that we consider is Model 0, the null model obtained when a real sentence is shuffled at random or, equivalently, when there is no word order constraint (and all the $\boldsymbol{n!}$ word orderings are equally likely). Then

$$\boldsymbol{p(d) = \frac{1}{\binom{n}{2}}(n-d).} \tag{1}$$

The parametrization of Model 0 is explained at the end of this section, as it requires further elaboration.
Model 1 is the displaced geometric distribution (with $\boldsymbol{p(0) = 0}$), defined as

$$\boldsymbol{p(d) = q(1-q)^{d-1},} \tag{2}$$

where $\boldsymbol{q}$ is the only parameter. The displaced geometric assumes that $\boldsymbol{d} \in [\mathbf{1}, \boldsymbol{\infty})$, but in real sentences $\boldsymbol{p(d) = 0}$ for $\boldsymbol{d \geq n}$. For this reason, we also consider Model 2, that is a right-truncated version in which probability mass is restricted to $\boldsymbol{d} \in [\mathbf{1}, \boldsymbol{d_{max}}]$, i.e.

$$\boldsymbol{p(d) = \frac{q(1-q)^{d-1}}{1-(1-q)^{d_{max}}}.}$$

The formula derivation can be found in Appendix A.
Given that distances are discrete, an exponential decay can be modeled with a geometric curve. Thus, Model 3 is a generalization of Model 1, obtained by splitting such model into two regimes, one for $\boldsymbol{d} \in [\mathbf{1}, \boldsymbol{d^*}]$ and another for $\boldsymbol{d} \in [\boldsymbol{d^*}, \boldsymbol{\infty}]$, as

$$\boldsymbol{p(d)} = \begin{cases} \boldsymbol{c_1(1-q_1)^{d-1}} & \text{if } \boldsymbol{d \leq d^*} \\ \boldsymbol{c_2(1-q_2)^{d-1}} & \text{if } \boldsymbol{d \geq d^*} \end{cases}$$

where $\boldsymbol{c_1}$ and $\boldsymbol{c_2}$ are normalization factors related by the following equation:

$$\boldsymbol{c_2 = \tau c_1} \tag{3}$$

and τ is such that the two regimes have the same value in d^*. That is, given $p'(d)$ and $p''(d)$, the probability distributions in the first and in the second regime respectively:

$$\begin{aligned} p'(d^*) &= p''(d^*) \\ c_1(1-q_1)^{d^*-1} &= c_2(1-q_2)^{d^*-1} \\ c_2 &= \frac{(1-q_1)^{d^*-1}}{(1-q_2)^{d^*-1}} c_1 \end{aligned}$$

from which it follows:

$$\tau = \frac{(1-q_1)^{d^*-1}}{(1-q_2)^{d^*-1}}. \tag{4}$$

The only free parameters are q_1, q_2 and d^*. It can be shown (Appendix A) that

$$c_1 = \frac{q_1 q_2}{q_2 + (1-q_1)^{d^*-1}(q_1 - q_2)}. \tag{5}$$

Model 4 is a generalization of Model 3 by right truncation, namely one regime for $d \in [1, d^*]$ and another one for $d \in [d^*, d_{max}]$, as

$$p(d) = \begin{cases} c_1(1-q_1)^{d-1} & \text{if } d \leq d^* \\ c_2(1-q_2)^{d-1} & \text{if } d^* \leq d \leq d_{max} \end{cases}$$

where c_1 and c_2 are normalization factors related by equation (3), where τ is computed as in equation (4). The only free parameters are q_1, q_2, d^* and d_{max}. For this model (see Appendix A) :

$$c_1 = \frac{q_1 q_2}{q2 + (1-q_1)^{d^*-1}(q_1 - q_2 - q_1(1-q_2)^{d_{max}-d^*+1})}. \tag{6}$$

Next, following research on syntactic dependency distances (Liu, 2007), we also consider Model 5, a power-law model that is a right-truncated Zeta distribution with parameters γ and d_{max}, as follows:

$$p(d) = \frac{d^{-\gamma}}{H(d_{max}, \gamma)}$$

where

$$H(d_{max}, \gamma) = \sum_{k=1}^{d_{max}} \frac{1}{k^\gamma}$$

is the generalized harmonic number of order γ of d_{max}.
Finally, we introduce Models 6 and 7: these are also composed of two regimes, the first one distributed as a power-law and the second one as an geometric. Model 6 is defined as:

$$p(d) = \begin{cases} c_1 d^{-\gamma} & \text{if } d \leq d^* \\ c_2 (1-q)^{d-1} & \text{if } d \geq d^* \end{cases}$$

where c_1 and c_2 are again related by equation (3), and τ is obtained as:

$$\begin{aligned} p'(d^*) &= p''(d^*) \\ c_1 {d^*}^{-\gamma} &= c_2 (1-q)^{d^*-1} \\ c_2 &= \frac{{d^*}^{-\gamma}}{(1-q)^{d^*-1}} c_1 \end{aligned} \tag{7}$$

from which it follows:

$$\tau = \frac{{d^*}^{-\gamma}}{(1-q)^{d^*-1}}. \tag{8}$$

Then, c_1 can be computed as (Appendix A):

$$c_1 = \frac{q}{qH(d^*, \gamma) + {d^*}^{-\gamma}(1-q)}.$$

Model 7, the right-truncated version of this double-regime distribution, is defined as:

$$p(d) = \begin{cases} c_1 d^{-\gamma} & \text{if } d \leq d^* \\ c_2 (1-q)^{d-1} & \text{if } d^* \leq d \leq d_{max} \end{cases}$$

where c_1 and c_2 are related by equation (7), τ is computed as in equation (8), and c_1 is obtained as (Appendix A):

$$c_1 = \frac{q}{qH(d^*, \gamma) + {d^*}^{-\gamma}(1-q-(1-q)^{d_{max}-d^*+1})}. \tag{9}$$

Let us go back to Model 0. Its formulation assumes that sentence length is known and unique, two assumptions which are not necessarily satisfied. In fact, on one hand model selection on real languages is performed both by fixing sentence length, and by considering jointly all sentences of any length for a given language. On the other

hand, in artificial datasets sentence length is set by us, and it is thus known. Then, in the context of model selection on fixed sentence lengths and on artificial data sentence length is unique, while in the case of model selection on sentences of various lengths it is not. Even where sentence length is unique, we give the model the freedom to select the value of $\boldsymbol{n}$ that maximises the likelihood by means of $\boldsymbol{d_{max}+1}$, where $\boldsymbol{d_{max}}$ is the free parameter we estimate. This is motivated by the fact that $\boldsymbol{n-1}$ is the theoretical maximum value of $\boldsymbol{d}$, but distances with this value might not be present in the samples. As shown in tables B2, B7, and B11 the estimated $\boldsymbol{d_{max}}$ always corresponds with the real maximum distance value observed in a sample. Then, for fitting purposes we distinguish between two specifications of Model 0. In the first one, Model 0.0, we relax the first assumption and estimate $\boldsymbol{n}$ by means of $\boldsymbol{d_{max}+1}$:

$$\boldsymbol{p(d) = \frac{1}{\binom{d_{max}+1}{2}}(d_{max}+1-d).}$$

where $\boldsymbol{d_{max}}$ is the only free parameter. This model is included in model selection on artificial data and on real sentences of fixed lengths.
Then, we derive an extended version of the model for model selection on sentences of mixed lengths, namely Model 0.1. Here we relax the assumption on uniqueness accounting for the presence of sentences of various lengths, such that:

$$\boldsymbol{p(d) = \sum_{n=min(n)}^{max(n)} p(d|n)\, p(n)}$$

where $\boldsymbol{p(d|n)}$ is $\boldsymbol{p(d)}$ in a sentence of length $\boldsymbol{n}$, $\boldsymbol{p(n)}$ is the proportion of sentences having length $\boldsymbol{n}$, and $\boldsymbol{min(n)}$ and $\boldsymbol{max(n)}$ are the minimum and maximum observed values of $\boldsymbol{n}$ in the sample.

Table 1. Models for the distribution of syntactic dependency distances. $\boldsymbol{K}$ is the number of free parameters. Refer to section 2 and appendix A for the derivation of the equations.

model	function	$\boldsymbol{K}$	equation
0.0	Null model	1	$\frac{1}{\binom{d_{max}+1}{2}}(d_{max}+1-d)$
0.1	Extended Null model	0	$\sum_{n=min(n)}^{max(n)} p(d\|n)\,p(n)$
1	Geometric	1	$q(1-q)^{d-1}$
2	Right-truncated geometric	2	$\frac{q(1-q)^{d-1}}{1-(1-q)^{d_{max}}}$
3	Two-regime geometric	3	$\begin{cases} c_1(1-q_1)^{d-1} & \text{if } d \leq d^* \\ c_2(1-q_2)^{d-1} & \text{if } d \geq d^* \end{cases}$
4	Two-regime right-truncated geometric	4	$\begin{cases} c_1(1-q_1)^{d-1} & \text{if } d \leq d^* \\ c_2(1-q_2)^{d-1} & \text{if } d^* \leq d \leq d_{max} \end{cases}$
5	Right-truncated Zeta distribution	2	$\frac{d^{-\gamma}}{H(d_{max},\gamma)}$
6	Two-regime zeta-geometric	3	$\begin{cases} c_1 d^{-\gamma} & \text{if } d \leq d^* \\ c_2(1-q)^{d-1} & \text{if } d \geq d^* \end{cases}$
7	Two-regime right-truncated zeta-geometric	4	$\begin{cases} c_1 d^{-\gamma} & \text{if } d \leq d^* \\ c_2(1-q)^{d-1} & \text{if } d^* \leq d \leq d_{max} \end{cases}$

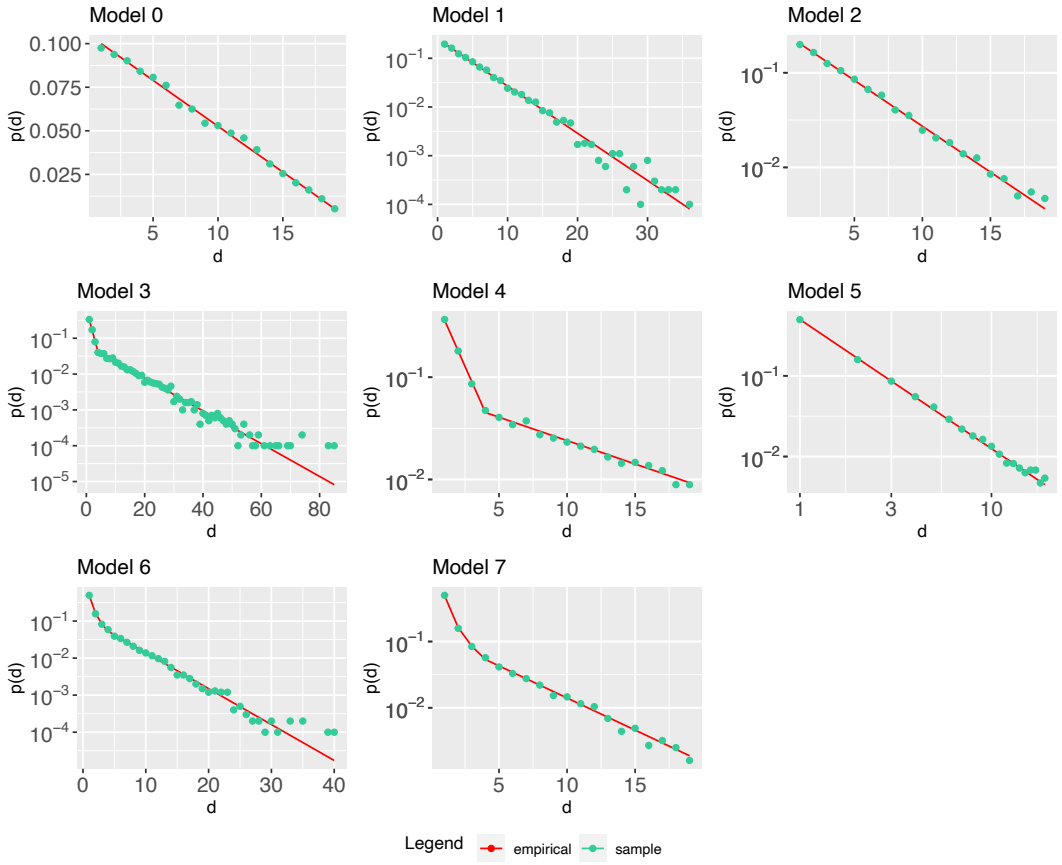


Figure 2.$\boldsymbol{p(d)}$, the probability of $\boldsymbol{d}$ in a model versus a random sample of itself. The random sample has size $\mathbf{10^4}$. $\boldsymbol{n = 20}$ ($\boldsymbol{d_{max} = 19}$) for the right-truncated models. $\boldsymbol{d^* = 4}$ for the two-regime models. For the equations of the models refer to table 1, while for the complete list of parameter values refer to table B3.

3. Results

We fit the models introduced in the previous section to a parallel collection of texts from 20 languages, that has been annotated with syntactic dependencies as in figure 1. To control for annotation style we consider two variants, PUD with the original annotation style (Nivre et al., 2017) and PSUD, that follows the alternative SUD annotation style (Gerdes et al., 2018). Refer to section 6 for further details on the data, and to section 7.2 for a complete description of the model selection procedure. This section is organized as follows: first, we report on the estimated best models, break-points ($\boldsymbol{d^*}$) and slope parameters ($\boldsymbol{q_1}$ and $\boldsymbol{q_2}$) for each language, both by considering fixed and mixed sentence lengths. We also include the result of a systematic analysis to check the robustness of the drawn conclusions to constraints on sample representativeness. Detailed tables of the estimated parameters, AIC scores, and AIC differences for both collections can be found in Appendix C. Then, we link a syntactic dependency closeness score (refer to section 7.4 for details) to the results of model selection on fixed sentence lengths. Notice that we often refer jointly to Models 3 and 4 (6 and 7) as '3-4' ('6-7'), given that they model the same probability distribution with or without a truncation point.

3.1. Model selection

Table 2 shows that the best model to describe syntactic dependency distances independent of sentence length is composed of two regimes in every language and collection. Models 3 and 4 dominate over the Models 6 and 7, with 13/20 languages in PUD and 11/20 in PSUD having Model 3 or 4 as the best one. We find overall agreement between the two annotation styles, both in terms of best model and in terms of right-truncation. The exceptions to this agreement are Indonesian and Japanese – for which PUD predicts an exponential decay in first regime, while PSUD identifies a power-law one – and Chinese, English, and Italian, where the best model in PUD and PSUD differs by right truncation. In figure 3, we show how the best estimated models in PUD are able to accurately capture the bulk of the distribution, with some variability left in the tail. The same figures can be found in Appendix C for PSUD.

The most voted best model for sentences of fixed lengths shows some variability for short and long sentences (see figure 4). Nevertheless, table 3 shows that a double regime model is the most frequent one in 17/20 languages in PUD (including a tie between Model 5 and Models 6-7 in Chinese), and in 14/20 languages in PSUD. Out of the languages for which a two-regime model is the best one, Models 3-4 win in 13/17 languages in PUD, and in 9/14 in PSUD. Once again, we find high consistency between annotation styles, with the exceptions of Indonesian, Polish, and Russian, for which PSUD predicts Model 5 as the most frequent best one (while PUD predicts models 3-4), and Japanese, for which PUD and PSUD differ in the type of two-regime model. Finally, in Arabic, Chinese, and Thai Model 5 is the most frequent one in both collections. However, we show in figure 5 how the most voted best model ceases to be Model 5 in some instances of both PUD and PSUD when the representativeness of a sentence length is taken into account. Refer to section 4.1 for details on this issue.

The only languages in which Model 5 is consistently the most frequent one even after imposing an arbitrary high threshold are Thai, Indonesian, and Arabic in PSUD. Arabic also shows a border-line behaviour in PUD, with Model 5 being consistently the most voted only up to a certain threshold value.

Table 2.Best model for the syntactic dependency distance distribution of sentences of mixed lengths for every language and collection. Models 3-4 are marked with pink and Models 6-7 with blue to ease visualization.

language	PUD	PSUD
Arabic	7	7
Chinese	6	7
Czech	3	3
English	3	4
Finnish	6	6
French	4	4
German	3	3
Hindi	7	7
Icelandic	3	3
Indonesian	3	7

language	PUD	PSUD
Italian	4	3
Japanese	4	7
Korean	7	7
Polish	3	3
Portuguese	3	3
Russian	3	3
Spanish	4	4
Swedish	3	3
Thai	6	6
Turkish	7	7

Table 3.Most voted best model for the syntactic dependency distance distribution of sentences of fixed lengths for every language and collection. The most voted best model is computed aggregating models by type, thus counting together the occurrences in which Models 3-4 (Models 6-7) are the best. Models 3-4 are marked with pink, Model 5 with yellow, and Models 6-7 with blue to ease visualization.

language	PUD	PSUD
Arabic	5	5
Chinese	5	5
Czech	3-4	3-4
English	3-4	3-4
Finnish	6-7	6-7
French	3-4	3-4
German	3-4	3-4
Hindi	6-7	6-7
Icelandic	3-4	3-4
Indonesian	3-4	5

language	PUD	PSUD
Italian	3-4	3-4
Japanese	3-4	6-7
Korean	6-7	6-7
Polish	3-4	5
Portuguese	3-4	3-4
Russian	3-4	5
Spanish	3-4	3-4
Swedish	3-4	3-4
Thai	5	5
Turkish	6-7	6-7

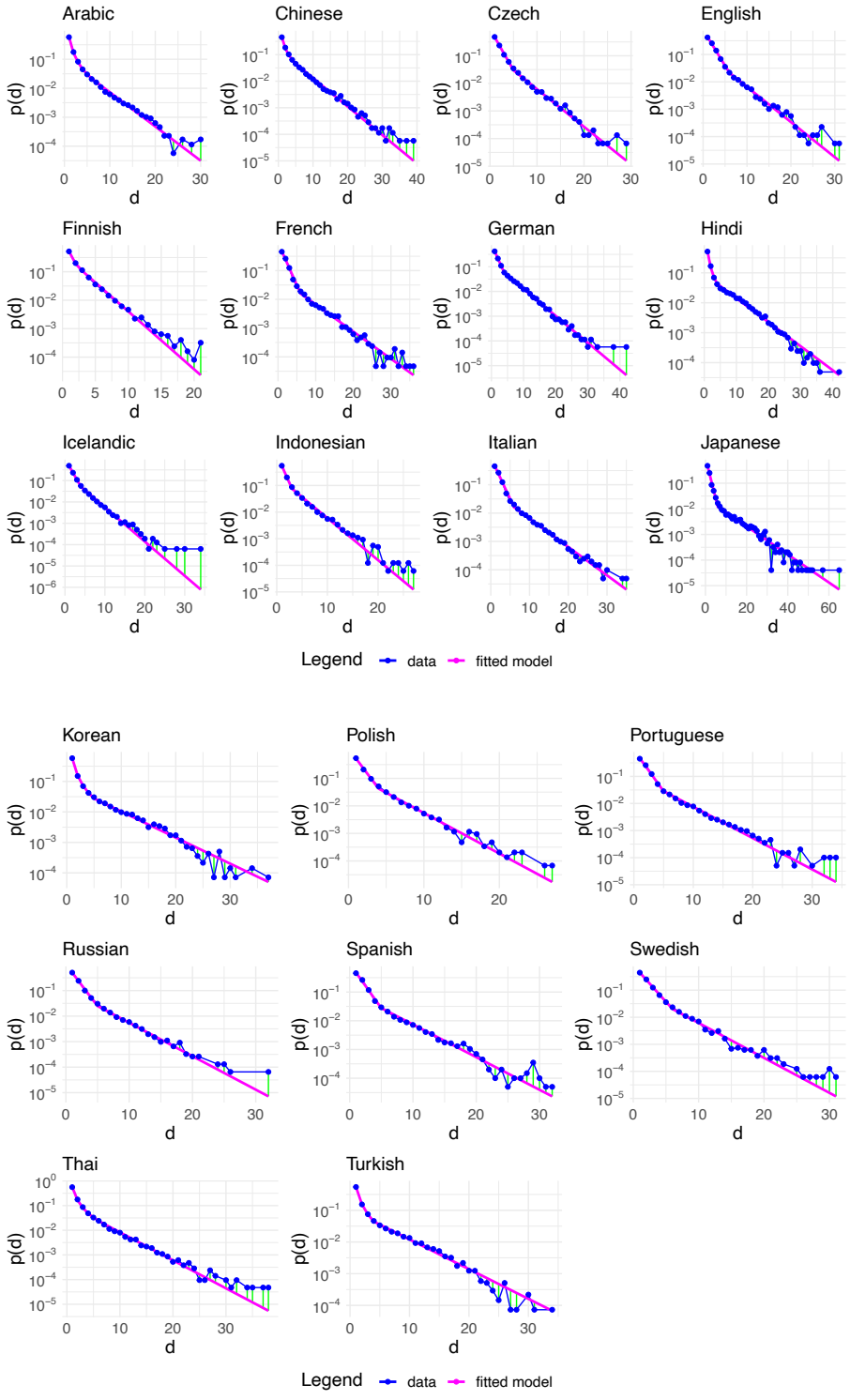
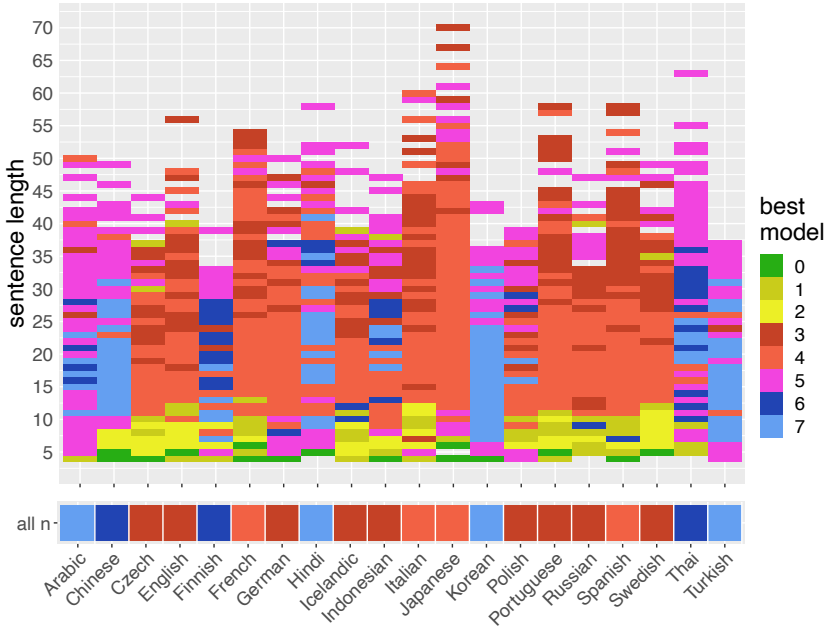
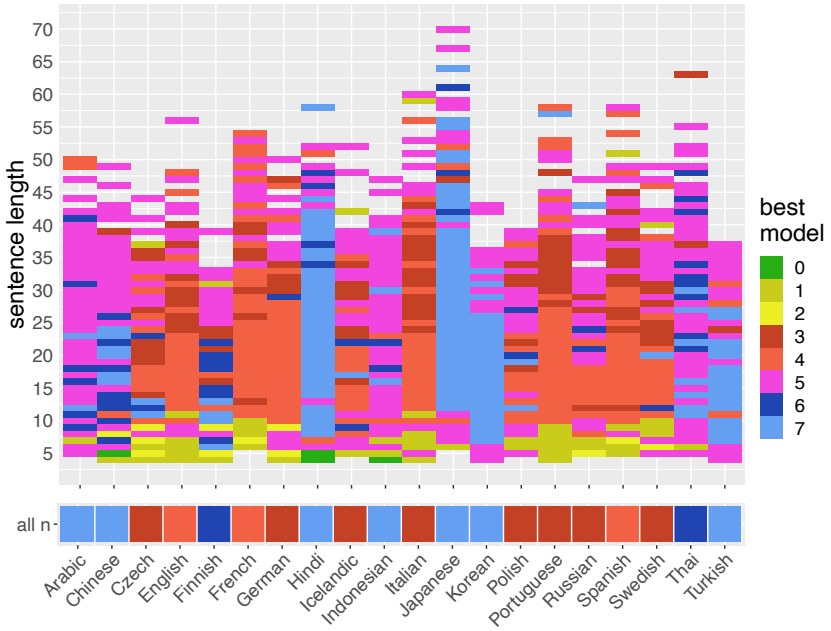


Figure 3.$\boldsymbol{p(d)}$, the probability that a dependency link is formed between words at distance $\boldsymbol{d}$ according to the best model for every language in PUD.



(a) PUD collection



(b) PSUD collection

Figure 4.Distribution of best model for each sentence length on top, with reference to the best model on mixed sentence lengths at the bottom. (a) PUD collection. (b) PSUD collection. In both (a) and (b) the empty tiles mark lengths for which no sentence was observed, or on which model selection was not performed given the minimum requirement to fit a double-regime model, described in section 7.2.4.

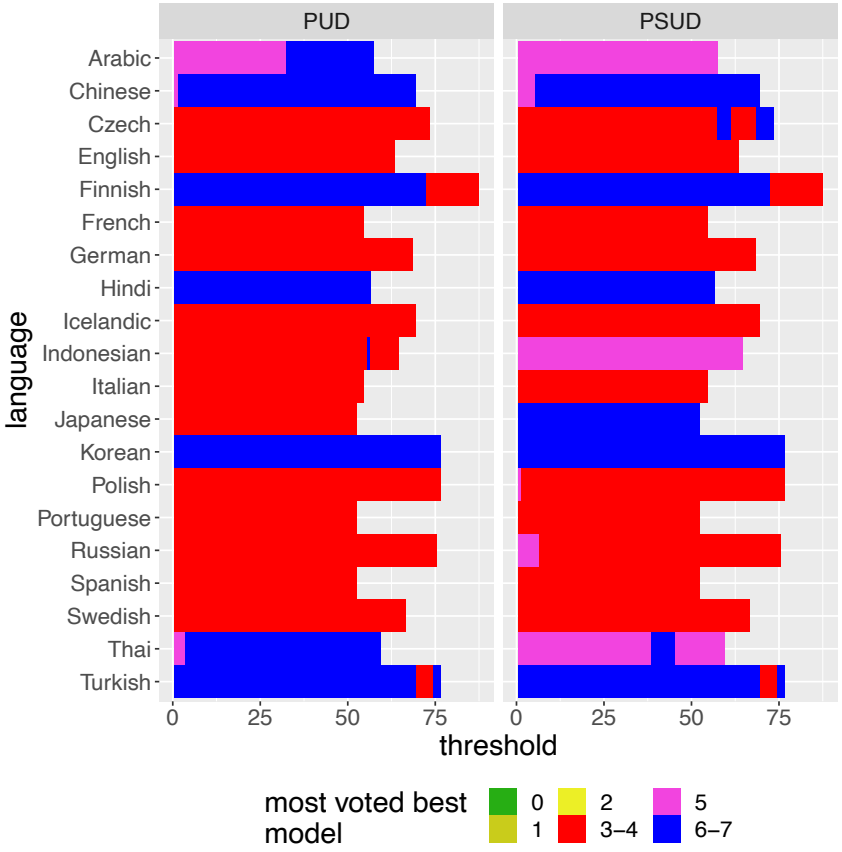


Figure 5. Most voted best model type across sentence lengths for increasing threshold: the threshold is the minimum number of distinct sentences with a certain length for such length to be included in model selection. For instance, when no threshold is set (1 minimum sentence) we get the scenario displayed in figure 4. Ties are counted in favour of models without two-regimes.

3.2. The break-point

As shown in figure 6, when looking at languages globally, meaning considering jointly all sentences of any length, we find that the break-point $\boldsymbol{d^*}$ always takes small values – ranging between 2 and 7 – and has a quite small standard deviation (see table 4), meaning that its value is stable across languages. This is especially true for Models 3-4 and the PSUD collection: out of 11 languages having one of Models 3-4 as the best one in this collection, 9 have an estimated break-point at $\boldsymbol{d^* = 4}$. In PUD these models have an average $\boldsymbol{d^*}$ value of 5, but with some more variability across languages. In both types of two regime models median and mean values are virtually the same, independently of annotation style, providing additional evidence of the stability of $\boldsymbol{d^*}$. Checking the distribution of $\boldsymbol{d^*}$ within a language allows us to verify whether global values (found when mixing sentence lengths), namely the bars in figure 6, are good approximations of the break-points actually observed in real sentences of any fixed length. We display the distribution of $\boldsymbol{d^*}$ for each language in the same figure. Once again the median is very close to the mean in almost every combination of two-regimes model and annotation style – with the exception of Models 6-7 in PSUD – further supporting $\boldsymbol{H_2}$. Then, notice that where Models 3-4 are the best we observe relatively narrow distributions, skewed towards low values and showing one or few modes. In particular, the global value of $\boldsymbol{d^*}$ is virtually always found in correspondence of one of these modal values, confirming its representativeness for the whole language. As we mentioned in section 1, the break-point may be understood as the balance in the trade-off between comprehension and processing difficulty, so that it is expected to show some variation depending on the length of the sentence. However, considering that sentences can reach up to a minimum of 37 (Turkish) and a maximum of 70 words (Japanese) (see table B4) the observed variation ranges in Models 3-4 are quite small, with values going up to roughly $\boldsymbol{d^* = 13}$. On the other hand, within languages for which Models 6-7 are the best when mixing sentence lengths, the distribution of $\boldsymbol{d^*}$ across different sentence lengths is generally flatter, especially in PSUD. Even where values are centered around a mode, this does not correspond with the break-point estimated globally, with the exception of Hindi. Thus, it appears like the global break-points estimated in Models 3-4 are good approximations of the values observed within the language, while estimates of $\boldsymbol{d^*}$ in Models 6-7 are less reliable as representations of the actual break-point.

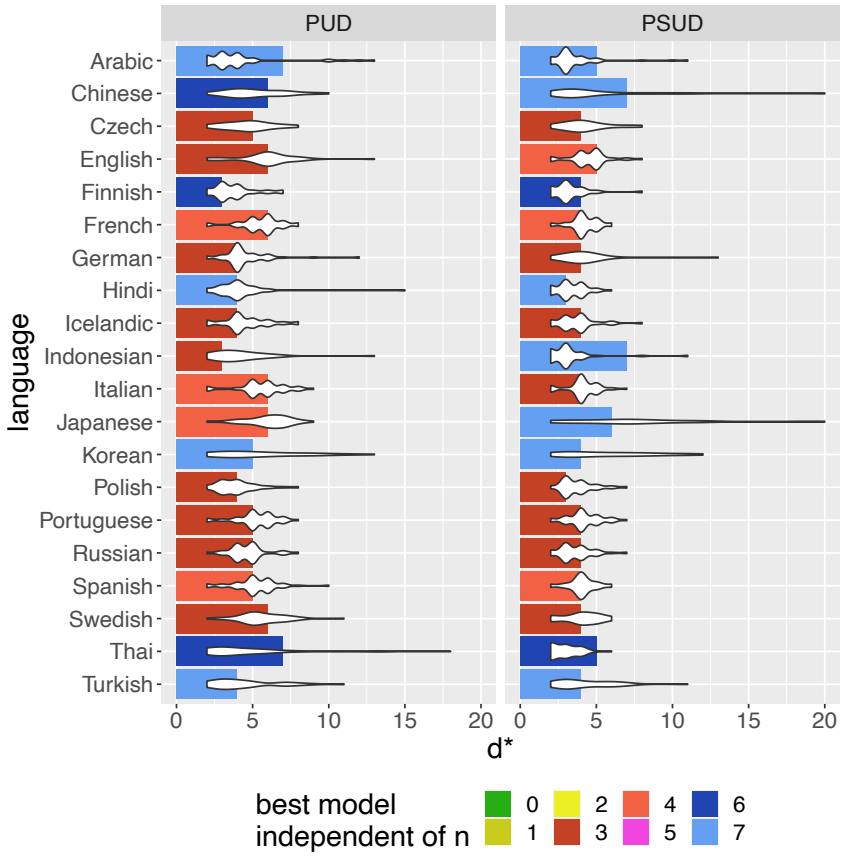


Figure 6. Value of d^* for mixed sentence lengths (bars) in each language and collection, and its distribution across fixed sentence lengths, color-coded by best model independent of sentence length (namely the best model estimated on sentences of mixed lengths).

Table 4.Summary statistics of d^* parameter, by annotation style and type of two-regime model, estimated from model selection on sentences of mixed lengths. The summary is computed over languages where Models 3-4 are the best, where Models 6-7 are the best, and over all languages where a double-regime model is the best (Models 3,4,6,7). Thus, sample size is expressed in number of languages. 'sd' stands for 'standard deviation'.

	models	sample size	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
PUD	3-4	13	3.00	4.00	5.00	5.00	6.00	6.00	1.00
	6-7	7	3.00	4.00	5.00	5.14	6.50	7.00	1.57
	3-4-6-7	20	3.00	4.00	5.00	5.05	6.00	7.00	1.19
PSUD	3-4	11	3.00	4.00	4.00	4.00	4.00	5.00	0.45
	6-7	9	3.00	4.00	5.00	5.00	6.00	7.00	1.41
	3-4-6-7	20	3.00	4.00	4.00	4.45	5.00	7.00	1.10

Table 5.Summary statistics of d^* parameter, by collection and type of two-regime model, estimated from model selection on sentences of fixed lengths. The summary is computed over sentence lengths and languages where Models 3-4 are the best, where Models 6-7 are the best, and over all languages and sentence lengths where a double-regime model is the best (Models 3,4,6,7). Thus, sample size is expressed in number of distinct sentence lengths mixing all languages. 'sd' stands for 'standard deviation'.

	models	sample size	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
PUD	3-4	431	2.00	4.00	5.00	5.37	6.00	13.00	1.43
	6-7	134	2.00	4.00	6.00	6.28	7.00	18.00	3.03
	3-4-6-7	565	2.00	4.00	5.00	5.59	6.00	18.00	1.97
PSUD	3-4	297	3.00	4.00	4.00	4.32	5.00	13.00	1.11
	6-7	190	2.00	3.00	5.00	6.21	8.00	20.00	3.73
	3-4-6-7	487	2.00	4.00	4.00	5.06	5.00	20.00	2.65

3.3. Speed of decay

The speed of decay in a geometric distribution can be computed from its q parameter as $\log q(1-q)$, that is the slope of the curve in log-linear scale (refer to section 7.3 for the full derivation). Figure 7 shows that such slope becomes steeper (more negative) as q increases, and we thus refer to q as the 'slope parameter' of a geometric curve, quantifying the speed of probability decay. Thus, q_1 and q_2 are the slope parameters of Models 3-4. Concerning Models 6-7, the slope of the first power-law regime is approximated through a geometric curve, as explained in section 7.3. For each language in which a two-regime model is the best, we consider q_1, q_2, and their ratio $\frac{q_1}{q_2}$, where the latter quantity is computed to establish which slope is steeper. When we refer to a slope, we refer to its absolute value.
Where the best model has two regimes, the slope parameters estimated are fairly stable across languages and show a clear pattern, with probability in the first regime consistently decaying faster compared to the second one. We observe this in figure 8 and table 6, noticing that the ratio $\frac{q_1}{q_2}$ is larger than 1 for every language and annotation style, and that q_1 and q_2 have a quite small standard deviation. Standard deviation values are the same for the two parameters, but q_2 takes much lower values, meaning that it is relatively more variable than q_1. Moreover – as in the case of the break-point parameter – median and mean values are virtually the same, for both q_1 and q_2 and independently of annotation style. Figure 8 (a) also shows that the slope estimated in the first regime on PUD is significantly lower than the one estimated in PSUD. In figure 9 we can observe how the pattern observed in the relation between the slopes holds also for the overwhelming majority of sentence lengths within a language, with few exception found for very short sentence length.

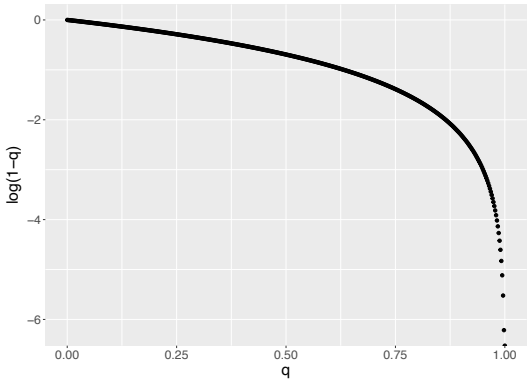
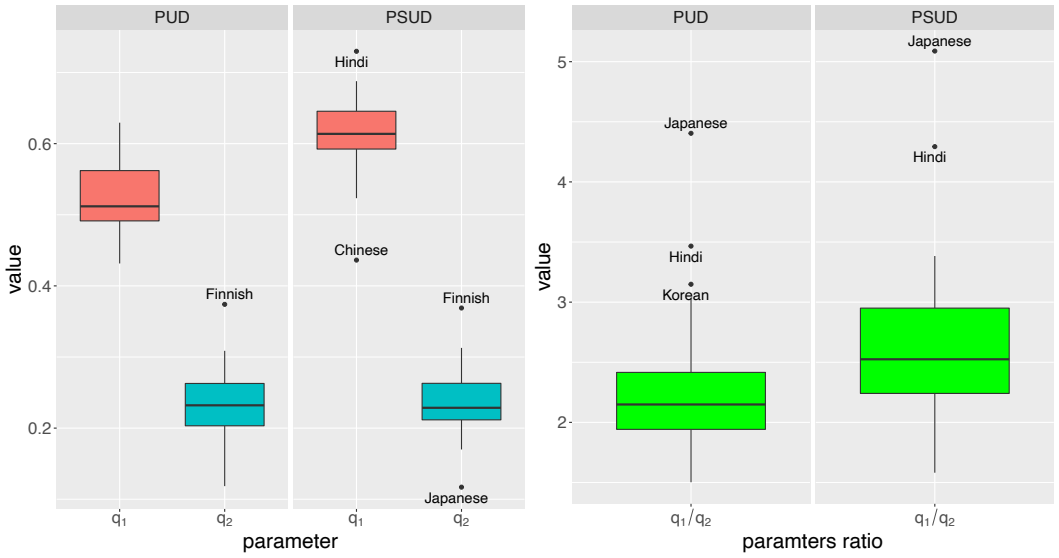


Figure 7.Slope of a geometric curve in log-linear scale as a function of its parameter $\boldsymbol{q}$.



(a) $\boldsymbol{q_1}$ and $\boldsymbol{q_2}$ parameters distribution

(b) Ratio $\frac{\boldsymbol{q_1}}{\boldsymbol{q_2}}$ distribution

Figure 8.Distribution of slope parameters $\boldsymbol{q_1}$ and $\boldsymbol{q_2}$ and their ratio. Isolated points are labelled with the corresponding language.

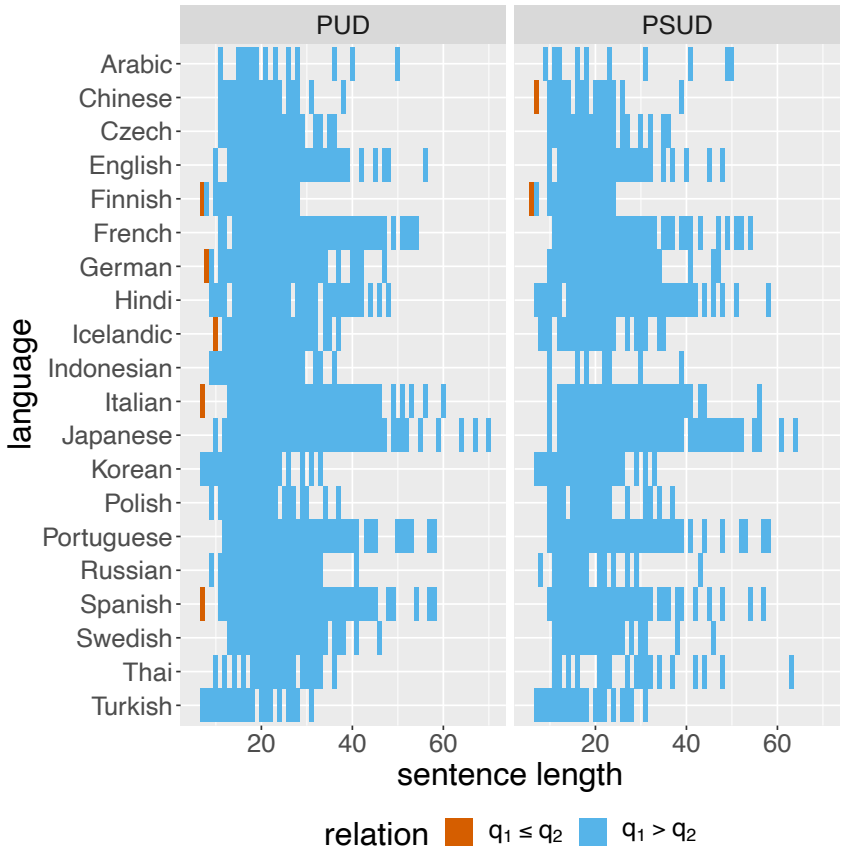


Figure 9. Relation between slope parameters $\boldsymbol{q_1}$ and $\boldsymbol{q_2}$ estimated from model selection on fixed sentence lengths. Lengths for which $\boldsymbol{q_1} \leq \boldsymbol{q_2}$ are colored in red, while those for which $\boldsymbol{q_1} > \boldsymbol{q_2}$ are colored in blue. Where the best model was 6 (7), the first slope was approximated by fitting Model 3 (4) with the original value of $\boldsymbol{d^*}$. The empty tiles indicate lengths for which no sentence was observed, a two-regime model was not the best one, or on which model selection was not performed given the minimum requirement on the number of observed distance values to fit a double-regime model, described in section 7.2.4.

Table 6.Summary statistics of q_1 and q_2 parameters and their ratio for model selection on sentences of mixed lengths, by annotation style (referred to as 'collection'). Statistics are computed over all sentence lengths and languages for which a double-regime model is the best. "sd" stands for "standard deviation".

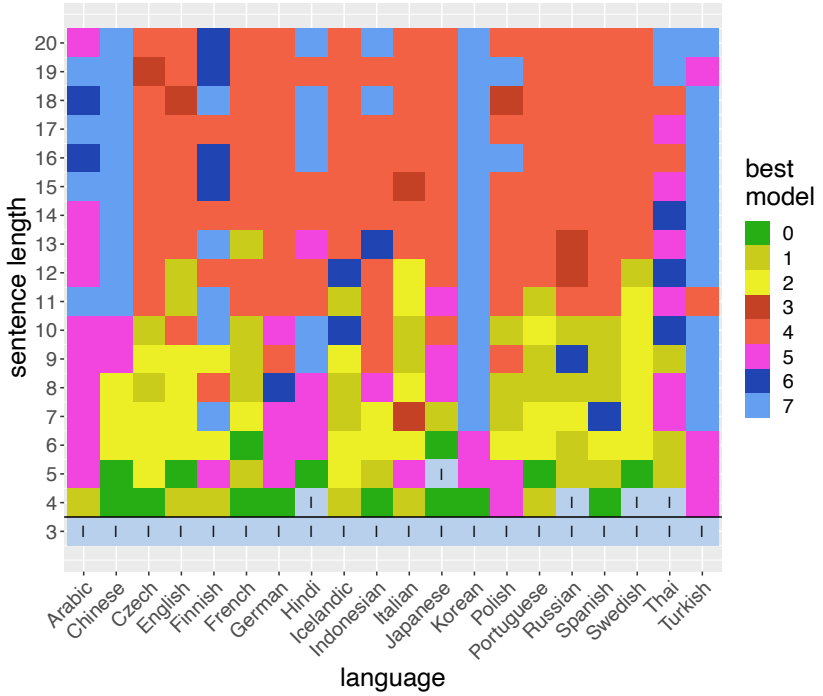
	collection	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
q_1	PUD	0.43	0.49	0.51	0.52	0.56	0.63	0.05
	PSUD	0.44	0.59	0.61	0.61	0.65	0.73	0.06
q_2	PUD	0.12	0.20	0.23	0.24	0.26	0.37	0.06
	PSUD	0.12	0.21	0.23	0.24	0.26	0.37	0.05
$\frac{q_1}{q_2}$	PUD	1.50	1.94	2.15	2.32	2.42	4.40	0.70
	PSUD	1.58	2.24	2.52	2.75	2.95	5.09	0.79

3.4. Optimality and best model

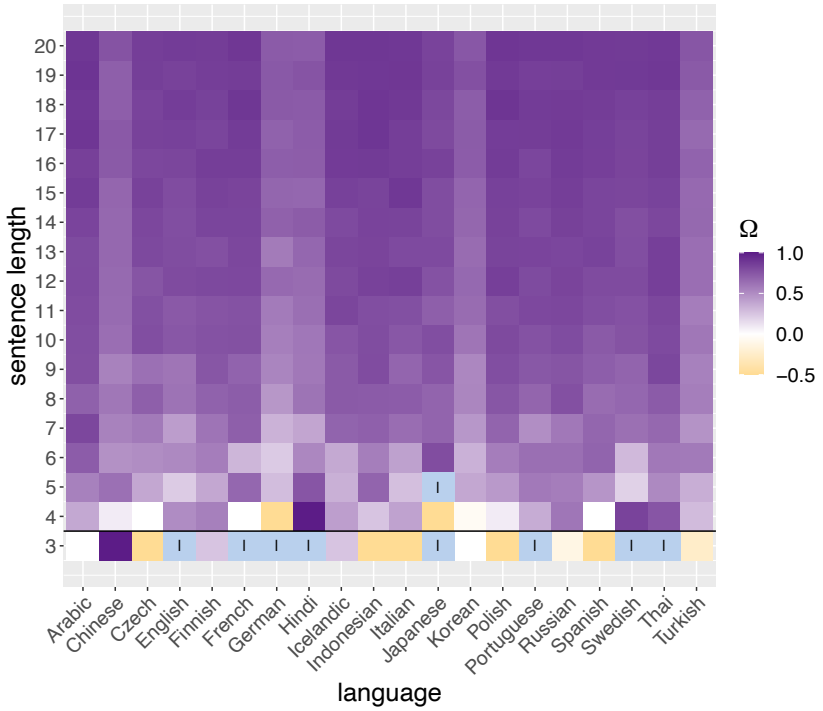
$\mathbf{\Omega}$ is a new closeness scores for syntactic dependency distances. The higher its value, the closer the syntactically related words. Refer to section 7.4 for further details on its properties and computation. The score takes positive values when syntactic dependency distances are minimized, negative values when they go against minimization, and values around 0 when there is no pressure in either direction (Ferrer-i Cancho et al., 2022). Let $\langle \mathbf{\Omega} \rangle$ be the average value of $\mathbf{\Omega}$ over all sentences with a given length in a language. We here focus on the scenarios in which $\langle \mathbf{\Omega} \rangle \approx \mathbf{0}$ or sentence length is small, and see how these are mirrored on the best estimated model. Table 7 shows how the best model in languages and sentence lengths for which $|\langle \mathbf{\Omega} \rangle - \mathbf{0.1}| \leq \mathbf{0}$ is always 0, with the exceptions of Korean in PSUD and Polish in PUD, for which the best model is Model 5. This can be also visualized in figure 10 and figure 11, by looking at the correspondence between green tiles in (a), indicating Model 0, and white tiles in (b), marking $\langle \mathbf{\Omega} \rangle \approx \mathbf{0}$. Concerning short sentences, figures 10 (b) and 11 (b) show how for sentences of 3-4 words we find the coexistence of three states, with syntactic structures being shaped by either minimization, maximization, or neither.

Table 7.Estimated best model on fixed sentence in collections, languages, and sentence lengths for which $|\langle \mathbf{\Omega} \rangle - \boldsymbol{\epsilon}| \leq \mathbf{0}$, with $\boldsymbol{\epsilon} = \mathbf{0.1}$. $\langle \mathbf{\Omega} \rangle$ is the average value of $\mathbf{\Omega}$ over all sentences with a given length in a language.

collection	language	n	$\langle \mathbf{\Omega} \rangle$	best model
PUD	Korean	4	-0.05	0
PUD	Czech	4	0.00	0
PUD	French	4	0.00	0
PUD	Spanish	4	0.00	0
PUD	Polish	4	0.08	5
PUD	Chinese	4	0.08	0
PSUD	Korean	4	-0.10	5
PSUD	Hindi	4	0.00	0

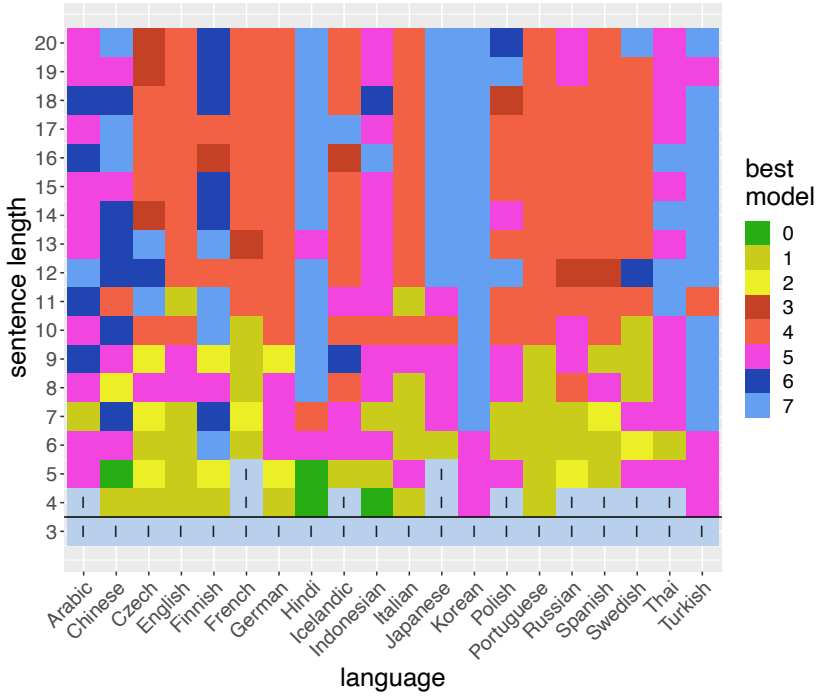


(a) Best model for every language and sentence length.

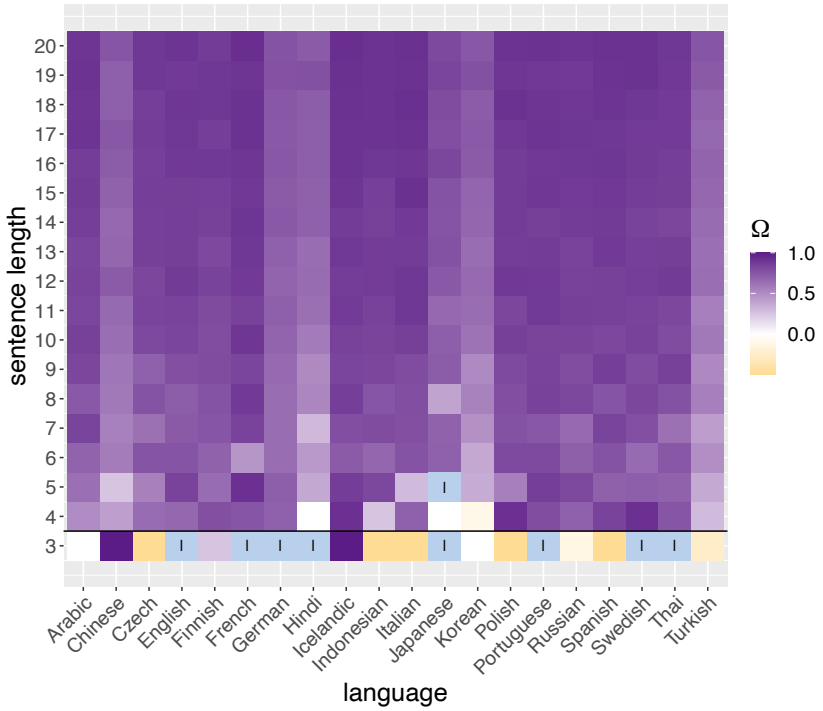


(b) $\langle \mathbf{\Omega} \rangle$ for each language and sentence length: orange signals negative values, white signals values around 0, and purple signals positive values.

Figure 10.Relation between $\mathbf{\Omega}$ score and best model in PUD. The barred grey cells indicate the sentence lengths which have not been observed, or that where excluded from model selection given the minimum requirement on the number of observed distance values to fit a double-regime model, described in section 7.2.4. Sentence lengths are cut at $\boldsymbol{n = 20}$ to ease visualization.



(a) Best model for every language and sentence length.



(b) $\langle \boldsymbol{\Omega} \rangle$ for each language and sentence length: orange signals negative values, white signals values around 0, and purple signals positive values.

Figure 11.Relation between $\boldsymbol{\Omega}$ score and best model in PSUD. The barred grey cells indicate the sentence lengths which have not been observed, or that where excluded from model selection given the minimum requirement on the number of observed distance values to fit a double-regime model, described in section 7.2.4. Sentence lengths are cut at $\boldsymbol{n = 20}$ to ease visualization.

4. Discussion

Now we review in detail the results presented in section 3. First, we focus on the two hypotheses object of study, namely that dependency distances are distributed following two exponential regimes ($\boldsymbol{H_1}$), and that the break-point is stable across languages ($\boldsymbol{H_2}$). Our results provide strong evidence for both hypotheses in a large group of languages, mainly Indo-European, consistently across annotation styles. Second, we make considerations on regularities in probability decay emerged from the analysis, concerning the greater steepness of the first regime with respect to the second one, and the stability of the estimated slopes across languages. Then, after having examined common patterns across languages, we focus on dissimilarities, addressing the origins of the observed variability. Finally, we discuss the relation between the best estimated model and the closeness of syntactic dependencies as captured by the optimality score $\boldsymbol{\Omega}$ (Ferrer-i Cancho et al., 2022), and make a summary of the effect of annotation style on the presented results.

4.1. The reality of two regimes

As it is often the case, the path to the truth seems to lie in the middle. We could neither generalize to all languages hypothesis $\boldsymbol{H_1}$ (supported by 13/20 languages in PUD and by 11/20 in PSUD), first advanced in (Ferrer-i-Cancho, 2004), nor fully corroborate the finding that dependency distances are power-law distributed in Chinese (Liu, 2007). However, we provided evidence for a possible explanation integrating both: a two-regime model in which the first regime is either exponential or power-law distributed, and the second one follows an exponential decay. Tables 2 and 3 show how this explanation is supported by all languages when considering sentences of mixed lengths, and is still robustly found for the majority of them (17/20 languages in PUD and 14/20 in PSUD) when specific sentence lengths are considered, independently of annotation style. Notice that these proportions are lower bounds: the reported figures and tables include every value of $\boldsymbol{n}$, without considering the numerosity of its sample, thus including in the statistics sentence lengths for which, for instance, only one sentence had been observed. This raises two concerns: on one hand, underrepresentation is strictly related to rareness, so that these sentence lengths should have a lower weight in the process of understanding the main distribution of syntactic dependencies in a language. On the other hand, from a statistical point of view a small sample is not enough to stand for a whole category. Nevertheless, we decided not to set a threshold in the reported results, as this would have been arbitrary, and it could have mistakenly hidden important aspects of the analysis. In fact, the second above mentioned concern does not have the same impact for every sentence length: consider a very long sentence, composed of 50 words, and a very short one, of only 4 words. While, keeping fixed the syntactic structure, the first one could appear with $\boldsymbol{50!}$ different re-orderings, the second one could only be written in $\boldsymbol{4!}$ possible ways. Thus, a single sentence observed for $\boldsymbol{n = 4}$ is much more representative (as the expected variability in dependency distance is lower) for the whole length category than a single one observed for $\boldsymbol{n = 50}$. Figure 5 shows how by only removing few underrepresented sentence lengths from the count the overall stability increases,

together with the languages supporting the presence of two regimes. Moreover, as we will discuss in more detail later on, the best model also seems to be related to $\boldsymbol{n}$, with the two regimes mainly appearing for long enough sentences (see figure 4).
The languages supporting $\boldsymbol{H_1}$, for which two exponential regimes are the most likely distribution, both fixing and mixing sentence lengths are: Swedish, Spanish, Russian, Portuguese, Polish, Japanese, Italian, Indonesian, Icelandic, German, French, English, and Czech in PUD, and a subset of them excluding Russian, Polish, Japanese and Indonesian in PSUD. Figure 5 shows how imposing even a low threshold on sentence length representativeness leads to the inclusion of Russian and Polish in the PSUD group. Concerning the remaining languages, a two-regime model is still found to be the best one independent of $\boldsymbol{n}$, but allowing for an initial power-law decay. When looking at the stability of these results across sentence lengths we observe different behaviours. As figure 5 shows, Turkish, Korean, Hindi, and Finnish in PUD, with the addition of Japanese in PSUD all have one of Models 6-7 as the most frequent best one, robustly to the imposed threshold. Then, while imposing a small threshold leads to include Chinese in the above mentioned group for both collections, for Thai this is true only in PUD. In fact, in PSUD a power-law is consistently the most frequent best one for Thai, Indonesian, and Arabic, even for arbitrary high thresholds. A limit behaviour is also observed in Arabic in PUD, for which the best model is consistently Model 5 up to a certain threshold value, and consistently one of Models 6-7 after it.
Nevertheless, in the overwhelming majority of languages the probability of observing a dependency follows a different pattern before and after a break-point, both when sentence lengths are considered jointly (see figure 6) and in isolation (see figures 4 and 5). These are mainly Indo-European languages. In particular, the shape of the distribution in the first regime is found to be either a power-law or an exponential, depending on the language, while the second regime follows an exponential decay. Moreover, in a few languages the two regimes disappear when looking at fixed sentence lengths, in a Simpson's paradox like fashion, leaving a power-law (Model 5) as the best one. Thus, our results suggest a unification of existing findings – with both exponential and power-law distributions shaping dependency distances – by providing evidence for a model encompassing both. While, as discussed later, for short sentences a single regime can occur, for long enough sentences we observe the emergence of the combination of two different probability regimes. Although we cannot exclude that other explanations for this structural phenomenon are possible, given the consistent body of literature supporting the chunking paradigm we find this to be the most likely explanation for the presence of the break-point. While within a chunk DDm pressures towards an exponential or a power-law decline in probability, when the constraints of STM become too burdensome a new chunk needs to be compressed and passed to higher levels of representation, and a new dependency is created to connect sequential chunks. Thus, after the break-point it is more likely than expected to find a new dependency. Notice that this has implications on the slopes of probability decay in the two regimes, which will be discussed below. Next, to verify the universality of this cognitive constraint we move to the analysis of the stability of the break-point.

4.2. The stability of the break-point

We were able to verify the stability of the break-point ($\boldsymbol{H_2}$) to a large degree, especially across those languages for which $\boldsymbol{H_1}$ holds, and when dependencies are annotated according to SUD. In particular, the estimated break-point values are consistent with the literature on capacity limitations of short term memory: in no language $\boldsymbol{d^*}$ exceeds the 'magical number' 7 (Miller, 1956), and the bulk of the values is centered at $\mathbf{4 \pm 1}$, which is generally recognized to be the working memory limitation on a wide range of tasks (Cowan, 2001). Notice that an implicit assumption of $\boldsymbol{H_2}$ is that the break-point estimated for a language is stable within the language itself. In fact, the goal of looking at the consistency of $\boldsymbol{d^*}$ across languages is providing support for the two regimes being a manifestation of a stable constraint on human memory, not depending on the specific language. However, this implies that the break-point value estimated for a given language is a reliable approximation of the constraint acting upon it at the sentence level. Thus, before comparing the value of $\boldsymbol{d^*}$ between languages, we analyzed the shape of its distribution within each language.

4.2.1. Fixed sentence lengths

In figure 6 we identify two main patterns, which seem to be related to the type of double-regime model found to be the best one on sentences of mixed lengths. All of the languages having a two-regime exponential (Models 3-4) as the best model have a $\boldsymbol{d^*}$ distribution skewed toward small values, covering a small portion of the domain, and showing one or few modal break-point values. In particular, in these languages the value of $\boldsymbol{d^*}$ estimated globally corresponds to the mode – or one of the modes – of its distribution across $\boldsymbol{n}$. On the other hand, the majority of languages having Model 6 or 7 as the best one show flat distributions of $\boldsymbol{d^*}$ that do not settle around any specific value, with the noticeable examples of Thai in PUD, and Japanese and Chinese in PSUD (see figure 6). Even where the distribution is skewed towards a certain value, this is almost always shifted with respect to the break-point estimated on mixed lengths, with the exceptions of Hindi and Finnish. Thus, where $\boldsymbol{H_1}$ holds the break-point values estimated on fixed sentence lengths are consistent with the value estimated globally for a language, and vary in a small range. This suggests that for languages in which Models 3-4 are the best ones, the break-point values estimated globally are robust approximations of the constraint shaping such languages at the level of sentences. On the other hand, $\boldsymbol{d^*}$ values estimated in Models 6-7 are less stable within sentences of fixed lengths. Thus, in the following discussion we consider the two types of double-regime distributions separately.

4.2.2. Mixed sentence lengths

Break-point values in Models 3-4 not only show greater stability within languages, but also between them. In figure 6 and table 4, we can see how in both collections all languages having Models 3-4 as the best ones have similar values of $\boldsymbol{d^*}$. In particular, when SUD annotation is used the number $\mathbf{4}$ appears in virtually every language. Given the robustness shown when considering both mixed and fixed sentence lengths, this phenomenon seems to be a genuine manifestation of a widely-spread pattern,

suggesting that STM limitations act in the same way upon a consistent group of languages.

4.3. Patterns in probability decay across regimes

While, as seen above, the discussion about the break-point is articulate and needs to be contextualized with respect to the best model and the annotation style, clear and general regularities arise in the relation between probability decay in the first and in the second regime.

4.3.1. Speed of decay

First, probability in the first regime decays faster compared to the second one, and this is true in all languages when mixing all sentence lengths (figure 8), as well as for the overwhelming majority of sentence lengths within each language (figure 9), independently of annotation style. This provides additional support for the "chunk-and-pass" processing being related to the initial stronger pressure of DDm, followed by its weakening when the limit of words that can be simultaneously processed in chunk is reached. At that point, a new longer connection becomes more likely, due to the need of linking the forthcoming chunk. The few exceptions are all related to sentence lengths up to $\boldsymbol{n = 10}$, for which – as we will discuss below – the shortness of the sentence can allow for it to be processed as a single chunk. This phenomenon is still consistent with $\boldsymbol{H_1}$: when the shortness of the sentence is such that the 'chunk-and-pass' mechanism is not at play yet, it is reasonable that dependencies longer than the cognitive constraint are even more of a burden, and are thus even less likely.

4.3.2. Stability of the slopes

Figure 8 shows how the slopes observed in the various languages are quite narrowly distributed around the same values. The range of variation of $\boldsymbol{q_1}$ and $\boldsymbol{q_2}$ is also similar, both between them and between collections. It is interesting to notice that, while the main difference related to annotation style is observed in the first slope, which is significantly larger in PSUD, $\boldsymbol{q_2}$ takes virtually the same values in the two collections. The identified outliers are very similar, allowing to draw some conclusions on the respective distance between languages: Finnish robustly shows a particularly high $\boldsymbol{q_2}$ value (implying a smaller difference between the two regimes), while Japanese and Hindi are largely over the average in terms of $\frac{\boldsymbol{q_1}}{\boldsymbol{q_2}}$ ratio, especially in PSUD. The greater difference in the slopes of probability decay for these languages is affine with the within-collection consistency of the best predicted model for fixed sentence lengths (figure 4). Indeed, clearer distinction between the two decays leads to increased likelihood of unveiling a two-regime distribution where present. Hindi in PUD does not show such great consistency, but it's also not particularly far from the bulk of the distribution.

4.4. Explaining variability

The search for universals is guided by the idea that languages show great diversity, but are also united by general principles shaping them at a higher level. In the above sections we discussed the general patterns object of our tested hypotheses. Now, we provide an account of the intrinsic linguistic diversity observed in our results.

4.4.1. Tail behaviour

First, plotting the best predicted model against the real data allows to visually assess the quality of its fit to the data. Figures 3 and B3 show the resulting plots for PUD and PSUD respectively. The best predicted models are able to capture very well the shape of the bulk of the distribution and the initial bending in all languages. However, they are not always able to fully capture the variability along the tail of the distribution. To begin with, noise naturally emerges for longer distances, which belong to rare long sentences. As we explained above, there are lengths for which only one sentence is observed, and which are thus under-represented. Taken this into account, for some languages the deviation from the best model could suggest the possible presence of an unveiled pattern. We hypothesise the existence of more than one break-point, implying a more structured view of the "chunk-and-pass" mechanism. However, introducing more regimes would greatly increase both the complexity of estimation (maximum likelihood estimation already requires to put particular care in the estimation of 3/4 parameters, see section 7), and the risk of overfitting the data. Thus, a thorough and rigorous methodology would need to be employed for such modelling.

4.4.2. DDm and sentence length

Then, we want to make some important and highly intertwined considerations on the relation between Dependency Distance minimization and sentence length. First, it is well known that a small sample size decreases reliability, and moving towards lower or greater sentence lengths the number of distinct sentences goes down considerably. In fact, only one sentence is observed for some sentence lengths, which is arguably not enough to make statements about the general distribution of a whole category. As previously anticipated, and as shown in figure 5, setting a constraint on the minimum number of sentences required to perform model selection leads to more stable results, together with a reduction in the number of languages for which Model 5 is the most frequent one. However, one should be careful when imposing such a constraint, as the strictness of the requirement would need to be related to length itself. Second, expanding on the last statement, DDm is not independent from the length of a sentence (Ferrer-i Cancho and Gómez-Rodríguez, 2021; Ferrer-i Cancho et al., 2022): as discussed in the previous section, pressure towards optimization tends to weaken in very short sequences, yielding either random or anti-dependency distance minimization effects. Third, given the previous point, it is necessary to consider the sentence lengths under examination in order to compare results. As a matter of fact, the mean sentence length in the Chinese corpus used in Haitao Liu's analysis was of 25 words (Liu, 2007), while in PUD and PSUD this value lowers

to 18.6, and this relevant difference could partly explain the inconsistency in the obtained results for such language. In the light of these considerations, it is easier to understand the variability observed in the best predicted model when controlling for sentence length. By looking at figure 4 it is clear that a double regime model predominates for middle range lengths, while other trends are predicted at the two extremes of the domain. Precisely, the sentence length domain can be the split in the following four (potentially overlapping) regions:

- *Random ordering:* the green instances (best model is 0) appearing for $\boldsymbol{n \leq 6}$ provide additional evidence for the already observed phenomenon concerning the weakening of DDm in sequences of few elements (Ferrer-i Cancho and Gómez-Rodríguez, 2021; Ferrer-i Cancho et al., 2022). In short sentences DDm is likely to be surpassed by other word order principles and rules, also due to the lower workload on memory.
- *Single chunk:* up to roughly 13 words the best predicted model is either 1, 2, or 5 in most languages. This possibly indicates that the sentence can be processed as a whole chunk when the number of words is short enough, and dependencies must be highly local to allow for this.
- *Emergence of two regimes:* for sentences in the middle length range we see the explosion (highly consistent in some languages) of the two-regime models. At these lengths our limited working memory requires the sentence to be broken down to be processed, so that after a very steep decrease in probability, a long dependency becomes more likely in order to link a chunk to the previous one.
- *Variability of long sentences:* for long (and rare) sentences no clear pattern appears, as the scarcity of examples for large sentence lengths introduces variability in the estimation of the best model.

4.5. Optimality of dependency distances

The main well-established principle underlying the present analysis is that dependency distances are minimized. So far, we have been addressing this topic from the point of view of the best model to describe their distribution, in which probability is a very rapidly decreasing function of distance. However, by identifying the best model we cannot quantify the degree of optimization of syntactic distances in a language. To achieve this purpose we compute $\boldsymbol{\Omega}$, a recently introduced normalized score for the quantification of the closeness between syntactically related words (Ferrer-i Cancho et al., 2022). Given that there are many sentences with the same length, we obtain $\langle \boldsymbol{\Omega} \rangle$ by averaging over the values of $\boldsymbol{\Omega}$ computed in each sentence for a given length. Let us consider figures 10 and 11. What we shall focus our attention on is not the great amount of sentence lengths showing a positive score, but rather on the instances in which $\langle \boldsymbol{\Omega} \rangle$ is close to 0, and on the values of $\langle \boldsymbol{\Omega} \rangle$ in short sequences.

4.5.1. Random word ordering

Concerning the first point, recall that $\langle \boldsymbol{\Omega} \rangle \approx \boldsymbol{0}$ is obtained when there is no pressure, either towards minimization or against it, and that Model 0 entails a random ordering of words. As figures 10 and 11 show, at $\boldsymbol{n = 4}$ – the first length on which

we could perform model selection based on the minimum requirement described in section 7.2.4 – we see how languages displaying values of $\langle \boldsymbol{\Omega} \rangle$ close to 0 (light tiles) also have Model 0 as the best one. We find this correspondence in almost every instance, with the exception of Korean in PSUD and Polish in PUD, for which the best model is Model 5 (table 7). This not only confirms the results of model selection, but also provides additional evidence for the finding that the pressure for syntactic dependency minimization is lower in short sequences (Ferrer-i Cancho et al., 2022). Notice that we cannot expect the complementary implication to be true, namely that $\langle \boldsymbol{\Omega} \rangle \approx \mathbf{0}$ where Model 0 is the best model. In fact, the latter only implies that Model 0 is the model best describing the distribution among the tested ones, but it might not be the absolute best one. Thus, the condition on $\langle \boldsymbol{\Omega} \rangle$ is stronger than the one on the best model, and the latter does not need to imply the former.

4.5.2. $\langle \Omega \rangle$ in short sequences

As a second point, while the level of optimization of distances increases with sentence length, in short sequences the picture is diversified. Figures 10 and 11 show how for $\boldsymbol{n = 3}$ and $\boldsymbol{n = 4}$ we see the coexistence of the three possible systems: anti-DDm (orange tiles), no bias (white tiles), and pro-DDm (purple tiles). This supports the finding that DDm can be surpassed by other word order principles in short sequences (Ferrer-i Cancho et al., 2022). Thus, in both PUD and PSUD we find languages in which $\langle \boldsymbol{\Omega} \rangle$ reaches -0.5, but also languages such as Arabic and Korean, in which $\langle \boldsymbol{\Omega} \rangle$ is virtually 0, and Chinese, for which sentences of 3 words are arranged so that distances are fully minimized (and $\langle \boldsymbol{\Omega} \rangle = \mathbf{1}$). The great consistency between the two collections for $\boldsymbol{n = 3}$ supports the robustness of these results.

4.6. The effect of annotation style

In the previous sections we presented the obtained results in relation to both PUD and PSUD, highlighting differences and commonalities. We now make a summary of these points, aiming to characterize the effect of annotation style on the considered hypotheses. Overall, the qualitative results are robust to annotation style suggesting the solidity of the observed patterns, but some differences emerge. The discussion on the origins of such differences is open, and is connected to the fundamental question of whether an annotation style is a more accurate representation of our brain's functioning than the other, or whether different styles simply mirror different aspects of this functioning. While providing a rather descriptive account of such differences, we partly attempt to address this question.

4.6.1. Best model

The first main point concerns the very high consistency in the best estimated models across annotation styles. Figure 6 shows that there are however a few exceptions, which we classified in two types: between-collection differences in right truncation, and in type of two-regime model. The latter is clearly of greater interest and it concerns two languages, Japanese and Indonesian, both having Models 3-4 as the

best one in PUD and Model 7 in PSUD, but showing a very different behaviour. Fro Japanese, the best model estimated on fixed sentence lengths is highly consistent with the one found when mixing sentence lengths in PUD and PSUD respectively, and in both collections the break-point value is $\boldsymbol{d^* = 6}$. This suggests a real difference in probability decay within a chunk depending on the chosen annotation guidelines, but also conveys the concreteness of the quantified limit on memory for such language. On the other hand, for Indonesian we find mixed evidence, both in terms of estimated break-point, which goes from $\boldsymbol{d^* = 3}$ in PUD to $\boldsymbol{d^* = 7}$ in PSUD, and in terms of best model for fixed sentence lengths (which is consistently a one-regime power-law in PSUD). In fact, this takes us to one of the main differences between annotation styles, which we can visualize in figure 5: while in PUD the only language showing some evidence for a single power-law regime for fixed sentence lengths is Arabic, in PSUD we have three languages strongly supporting the reality of such distribution. For Arabic, Indonesian, and Thai, the two regimes observed for mixed sentence lengths contradict what is found when sentence lengths are analysed in isolation, which seems to reflect Simpson's paradox, a phenomenon according to which a statistical trend disappears when single groups are considered.

4.6.2. The break-point

We have seen in figure 6 how the break-points estimated in both collections cover the same portion of domain, ranging from 3 to 7, with average values close to 5 in PUD and to 4 in PSUD. Thus, both annotation styles are indicate a memory constraint in line with the values reported by the relevant literature (Cowan, 2001; Miller, 1956). However, while in PUD there is no settling around a particular value, in PSUD $\boldsymbol{d^*}$ is nearly uniform at $\boldsymbol{d^* = 4}$, especially within Models 3-4. This raises the following questions: is this regularity given by chance? Or does it mirror a better ability of SUD to capture syntactic relations as formed by our minds? Given that – besides individual differences – the overall structure of the brain is the same for all humans, the constraint on memory is expected to be uniform across languages (hence the motivation for $\boldsymbol{H_2}$). Thus, one could speculate that SUD annotation style is actually more capable of unveiling this uniformity, that is assumed to exist.

4.6.3. Dependency distance optimization

SUD guidelines have been found to adhere more closely to the principle of Dependency Distance minimization (Ferrer-i Cancho et al., 2022), and this is confirmed by our findings. In fact, despite predicting a power-law decay in the first regime for two more languages compared to PUD, $\boldsymbol{q_1}$ is significantly higher in PSUD (figure 8). This entails a faster decay in probability within the chunk, related to the predominance of short local dependencies in PSUD. Moreover, we can observe from figures 10 (b) and 11 (b) how the values of $\boldsymbol{\Omega}$ computed in the PSUD collection are generally larger (darker tiles), suggesting the stronger degree of optimization of dependency distances in the SUD setting.

5. Conclusions

18 years after the publication of "Euclidean distance between syntactically linked words" (Ferrer-i-Cancho, 2004), some light has been shed on the particular shape observed in the distribution of syntactic dependency distances. As a crucial finding of the present research, in each analyzed language the probability of observing a dependency – independently of the length of the sentence it belongs to – is best described by a double-regime model. Support for the presence of two probability regimes also comes from a systematic analysis performed at a finer-grained level, distinctively considering each sentence length. In this setting, for the great majority of languages a double-regime model is the most frequent one, while the few remaining languages show a power-law decay as the most frequent, partly in accordance with what has been found concerning a Chinese treebank, where however sentences of mixed lengths were analysed (Liu, 2007). Furthermore, the break-point between the two regimes estimated globally for each language varies in a small range ($\mathbf{3} \leq \boldsymbol{d^*} \leq \mathbf{7}$), which becomes even narrower when only languages in which $\boldsymbol{H_1}$ holds are considered. In fact, $\boldsymbol{H_2}$ seems to be related to the probability distribution observed in the first regime, leading to the identification of a group of languages where probability follows a double-exponential decay ($\boldsymbol{H_1}$), and within which the break-point is very similar ($\boldsymbol{H_2}$). This group is mainly populated by Indo-European languages. However, languages from this family are over-represented in our sample, and other interesting patterns could emerge if a larger group of languages from other families where analysed. These considerations hold independently of annotation style, but it has not escaped our attention that in PSUD values of $\boldsymbol{d^*}$ for such group are almost uniform at 4, a widely accepted quantification of the constraint on short term memory (Cowan, 2001). This could, in our opinion, reflect a higher sensitivity of SUD annotation style to the way in which our minds create and process language, bringing to light a 'universal' constraint which is not language dependent. Another general pattern emerged in the relation between the speeds of the decays, whereas probability in the first regime is always faster than in the second one. Few exceptions concern very short sequences, and are still consistent with the 'chunk-and-pass' mechanism. In the framework of language processing, these findings provide strong support for the "chunk-and-pass" mechanism (Christiansen and Chater, 2015). In fact, the presence of these two different regimes could actually mirror the two different speeds at which probability decays within a chunk and outside of it. In our view, this appears to be the most reasonable and pertinent explanation for the observed systematic decrease in the strength of DDm, but we do not exclude that other explanations could as well be plausible. Future work could further investigate the distribution in the second regime, exploring different combinations of exponential and power-law decay. Then, the possible presence of more than one break-point could be explored, yet taking the due precautions to avoid overfitting. Importantly, to understand the extent to which the observed phenomena can be considered universal, the same analysis shall be performed on a wider set of languages, within the bounds of the employment of parallel corpora to ensure comparability.

6. Materials

6.1. Real languages

We run our analyses on a parallel subset of 20 languages from the Universal Dependencies collection (Nivre et al., 2017). Table 8 shows languages with the linguistic family they belong to, and their writing systems. This subset is parallel in the sense that it contains the same sentences translated in every language. We use version 2.6, available here. Parallelism is a crucial factor to account for when aiming at cross-linguistic comparisons, as context can largely influence various aspects of language, including dependency structure. Another factor that shall be considered is annotation style, as there is no univocal way to generate syntactic dependency trees starting from a sentence. For this reason, we compare two different annotation styles: Universal Dependencies (Nivre et al., 2017) and the alternative Surface Syntactic Universal Dependencies (Gerdes et al., 2018). We refer to the two resulting versions of the collection as PUD and PSUD. Tables B4 and B8 summarize PUD and PSUD respectively. As these tables show, mean distance values ($\bar{d}$) are smaller in PSUD.

Table 8. Languages included in the analysis, their linguistic family, and writing system.

language	family	writing system
Arabic	Afro-Asiatic	Arabic
Chinese	Sino-Tibetan	Han
Czech	Indo-European	Latin
English	Indo-European	Latin
Finnish	Uralic	Latin
French	Indo-European	Latin
German	Indo-European	Latin
Hindi	Indo-European	Devanagari
Icelandic	Indo-European	Latin
Indonesian	Austronesian	Latin
Italian	Indo-European	Latin
Japanese	Japonic	Japanese
Korean	Koreanic	Hangul
Polish	Indo-European	Latin
Portuguese	Indo-European	Latin
Russian	Indo-European	Cyrillic
Spanish	Indo-European	Latin
Swedish	Indo-European	Latin
Thai	Kra-Dai	Thai
Turkish	Turkic	Latin

6.2. Artificial data

With the purpose of validating our model selection procedure we generate a random artificial sample of each model. The methodology to generate random deviates from each distribution is described in section 7.

7. Methods

We here report on technical details of the employed methods. This section has the following structure. First, we explain the methodology to generate artificial samples from a distribution. Second, we describe model selection procedure, log-likelihood function derivation for each model, and parameter estimation through maximum likelihood. Then, we characterize the relation between parameter q of a geometric distribution and its slope in log-linear scale, which is relevant for the analysis on the speed of decay in the two regimes. Finally, we expand on the optimality score used to quantify the pressure for minimization of syntactic dependency distances, and on its computation.

7.1. Artificial data generation

In the following, let $p_x(d)$ be the probability of d according to Model x. The parameter values used to generate each model are reported in table B3, while sample size is $N = 10^4$ for each model. For right-truncated models sentence length is set to $n = 20$, so that $d_{max} = 19$.

7.1.1. Von Newman simple rejection method

We generate a sample of the first specification of Model 0, namely Model 0.0, with parameter $d_{max} = 19$. To do so, we leverage Von Newman's simple rejection method, which approximates a distribution by only including in the sample values underneath its curve. A random sample of Model 0 is created in the following way:

- generate a random uniform deviate d_r in $[1, d_{max}]$.
- generate a random uniform deviate u in $[0, 1]$.
- if $p_0(d_r) \leq u$, distance d_r is added to the sample.

7.1.2. Geometric distribution: random deviate generation

For the geometric distribution and its right-truncated version, namely Model 1 and Model 2, we use the following equation (Dagpunar et al., 1988; Devroye, 1986):

$$L = 1 + \left\lfloor \frac{\log x}{\lambda} \right\rfloor$$

where $\lambda = \log(1 - q)$, and $q = 0.2$ is the parameter of the desired geometric distribution. This formula allows for the generation of a random geometric deviate L from a random uniform deviate x, saving computational time by pre-computing λ.

7.1.3. Zeta distribution: random deviate generation

For Model 5, we employed the algorithm proposed by Devroye to generate efficiently a random deviate from a power-law distribution (Devroye, 1986), adapting it to allow for right-truncation. Two random uniform deviates U and V in [0,1] are generated,

and two quantities $\boldsymbol{X}$ and $\boldsymbol{T}$ are computed from them, according to the following equations:

$$\boldsymbol{X} = \left\lfloor \boldsymbol{U}^{-\frac{1}{\gamma-1}} \right\rfloor$$
$$\boldsymbol{T} = \left(\mathbf{1} + \frac{\mathbf{1}}{\boldsymbol{X}}\right)^{\boldsymbol{\gamma-1}}$$

Then, $\boldsymbol{X}$ is added to the sample if the two following conditions are simultaneously satisfied:

$$\boldsymbol{V X} \frac{\boldsymbol{T-1}}{\boldsymbol{b-1}} \leq \frac{\boldsymbol{T}}{\boldsymbol{b}}$$
$$\boldsymbol{X} \leq \boldsymbol{d_{max}}$$

where $\boldsymbol{b = 2^{\gamma-1}}$, and $\boldsymbol{\gamma}$ is the desired exponent for the distribution. The first condition is required by the original algorithm, while the second is imposed to obtain a right truncation in the probability distribution. The chosen value of $\boldsymbol{\gamma = 1.6}$ is the value obtained from fitting a right-truncated Zeta distribution to a Chinese treebank (Liu, 2007).

7.1.4. Simulating two regimes

Random samples of the two-regime models, namely Models 3, 4, 6, and 7, are generated using the tabular inversion method (Muller, 1958). This method generates artificial distances in a pre-specified range, namely $[\mathbf{1}, \boldsymbol{d_{max}}]$ in our context. Thus, in order to simulate Models 3 and 6 – which do not have a right-truncation – we set $\boldsymbol{d_{max} = 10^6}$, while for Models 4 and 7 $\boldsymbol{d_{max} = 19}$.
For the simulation of Model $\boldsymbol{x}$:

- Compute vector $\boldsymbol{p_x(d)}$ with $\boldsymbol{d} \in [\mathbf{1}, \boldsymbol{d_{max}}]$.
- Compute vector $\boldsymbol{P_x(d)}$, the backward cumulative of $\boldsymbol{p_x(d)}$, as

$$\boldsymbol{P_x(d)} = \sum_{\boldsymbol{d'=1}}^{\boldsymbol{d}} \boldsymbol{p_x(d')}.$$

- Generate a random uniform deviate $\boldsymbol{u}$ in $\mathbf{(0,1)}$.
- Locate the first element of vector $\boldsymbol{P_x(d)}$ such that $\boldsymbol{P_x(d) \geq u}$.
- Add the index of such element to the sample.

7.2. Model selection

We here describe the model selection procedure implemented to test $\boldsymbol{H_1}$. Optimal parameters for each model are estimated by maximum likelihood. Then, the best model

is selected according to Information Criteria (Anderson and Burnham, 2004). Before proceeding with fitting the real data, in order to validate the employed methodology we test our model selection procedure on the artificial random samples of each model in the ensemble, generated as described in the above section. In real languages, models are compared through Akaike Information Criterion (AIC). However, since in the artificial samples the true data generating process is known, the best model is selected through Bayes Information Criterion (BIC), which relies on the assumption that the real distribution is among the tested ones (Wagenmakers and Farrell, 2004). With respect to AIC, the criterion proposed by Schwarz (BIC) applies a stronger penalty to the number of parameters. Given that both AIC and BIC are measures of information loss, the best model for a sample is the one minimizing this quantity. We aim to find the best model in a sample of $\boldsymbol{N}$ distances $\boldsymbol{\{d_1, d_2, ..., d_i, ..., d_N\}}$, where $\boldsymbol{min(d)}$ and $\boldsymbol{max(d)}$ are, respectively, the minimum and maximum observed distances, and $\boldsymbol{f(d)}$ is the frequency of distance $\boldsymbol{d}$ in the sample.

7.2.1. Log-likelihood functions derivation

This section is devoted to the description of the derivation of the log-likelihood functions of the tested models, summarized in Table 7. Following the standard procedure, we derive the log-likelihoods of the models by multiplying the predicted probabilities for every observed distance, and taking the logarithm. In this way, we can rewrite the formula as a sum of the log probabilities over the sample size, $\boldsymbol{N}$.

$$\mathcal{L} = \sum_{i=1}^{N} \log p(d_i) = \sum_{d=1}^{max(d)} f(d) \log p(d)$$

where

$$\sum_{d=1}^{max(d)} f(d) = N.$$

For the first specification of the *Null model*, namely Model 0.0, in which $\boldsymbol{d_{max}}$ is the only free parameter:

$$\begin{aligned}
\mathcal{L} &= \sum_{d=1}^{max(d)} f(d) \log \left(\frac{2(d_{max} + 1 - d)}{d_{max}(d_{max} + 1)} \right) \\
&= \sum_{d=1}^{max(d)} f(d) \left[\log \left(\frac{2}{d_{max}(d_{max} + 1)} \right) + \log(d_{max} + 1 - d) \right] \\
&= N \log \left(\frac{2}{d_{max}(d_{max} + 1)} \right) + W
\end{aligned}$$

where

Table 7. The log-likelihood $\mathcal{L}$ for each of the probability mass functions. $\boldsymbol{K}$ is the number of free parameters, $\boldsymbol{N}$ is the sample size, i.e. $\boldsymbol{N = \sum_{i=1}^{max(d)} f(d_i)}$, $\boldsymbol{M}$ is the sum of distances weighted by frequency, i.e. $\boldsymbol{M = \sum_{i=1}^{max(d)} f(d_i)d_i}$, $\boldsymbol{M^*}$ is the same sum up to $\boldsymbol{d^*}$, i.e. $\boldsymbol{M^* = \sum_{i=1}^{d^*} f(d_i)d_i}$, $\boldsymbol{M'}$ is the sum of log distances weighted by frequency, i.e. $\boldsymbol{M' = \sum_{i=1}^{max(d)} f(d_i)log(d_i)}$, and $\boldsymbol{W}$ is such that $\boldsymbol{W = \sum_{i=1}^{max(d)} f(d_i)log(d_{max} + 1 - d_i)}$. $\boldsymbol{N^*}$ is the sum of distance frequencies up to $\boldsymbol{d^*}$, i.e. $\boldsymbol{N^* = \sum_{i=1}^{d^*} f(d_i)}$, and $\boldsymbol{M'^*}$ is $\boldsymbol{M'}$ up to $\boldsymbol{d^*}$, i.e. $\boldsymbol{M'^* = \sum_{i=1}^{d^*} f(d_i)log(d_i)}$. Finally, $\boldsymbol{W_n = \sum_{i=1}^{max(d)} f(d_i)\log(n - d_i))}$ and $\boldsymbol{N_n = \sum_{i=1}^{max(d)} f(d_i)}$ in sentences of length $\boldsymbol{n}$.

Model	Function	$\boldsymbol{K}$	$\mathcal{L}$
0.0	Null model	1	$\boldsymbol{N \log(\frac{2}{d_{max}(d_{max}+1)}) + W}$
0.1	Extended Null model	0	$\boldsymbol{\sum_{n=min(n)}^{max(n)}[N_n \log\left(\frac{2}{n(n-1)}\right) + W_n]}$
1	Geometric	1	$\boldsymbol{N \log q + (M - N)\log(1 - q)}$
2	Right-truncated geometric	2	$\boldsymbol{N \log\left(\frac{q}{1-(1-q)^{d_{max}}}\right) + (M - N)\log(1 - q)}$
3	Two-regime geometric	3	$\boldsymbol{N^* \log c_1 + (N - N^*)\log c_2 + (M^* - N^*)\log\left(\frac{1-q_1}{1-q_2}\right) + (M - N)\log(1 - q_2)}$
4	Two-regime right-truncated geometric	4	$\boldsymbol{N^* \log c_1 + (N - N^*)\log c_2 + (M^* - N^*)\log\left(\frac{1-q_1}{1-q_2}\right) + (M - N)\log(1 - q_2)}$
5	Right-truncated Zeta distribution	2	$\boldsymbol{-\gamma M' - N \log(H(d_{max}, \gamma))}$
6	Two-regime zeta-geometric	3	$\boldsymbol{N^* log(r_1) - \gamma M'^* + (N - N^*)log(r_2) + (M - M^* - N + N^*)log(1 - q)}$
7	Two-regime right-truncated zeta-geometric	4	$\boldsymbol{N^* log(r_1) - \gamma M'^* + (N - N^*)log(r_2) + (M - M^* - N + N^*)log(1 - q)}$

$$N = \sum_{d=1}^{max(d)} f(d)$$
$$W = \sum_{d=1}^{max(d)} f(d)\log(n-d).$$

For the extended version of the Null model, namely Model 0.1 in which the observed sentence lengths are supplied and there is no free parameter:

$$\begin{aligned}\mathcal{L} &= \sum_{n=min(n)}^{max(n)} \sum_{d=1}^{max(d)} f(d)\log\frac{2(n-d)}{n(n-1)} \\ &= \sum_{n=min(n)}^{max(n)} \sum_{d=1}^{max(d)} f(d)\left[\log\frac{2}{n(n-1)} + \log(n-d)\right] \\ &= \sum_{n=min(n)}^{max(n)} \left[N_n \log\frac{2}{n(n-1)} + W_n\right]\end{aligned}$$

where

$$W_n = \sum_{d=1}^{max(d)} f(d)\log(n-d)$$
$$N_n = \sum_{d=1}^{max(d)} f(d)$$

in sentences of length n.
For the *geometric models*, we start from the derivation of the right-truncated version, namely Model 2:

$$\begin{aligned}\mathcal{L} &= \sum_{d=1}^{max(d)} f(d)\log\frac{q(1-q)^{d-1}}{1-(1-q)^{d_{max}}} \\ &= \sum_{d=1}^{max(d)} f(d)\left[\log\frac{q}{1-(1-q)^{d_{max}}} + (d-1)\log(1-q)\right] \\ &= N\log\frac{q}{1-(1-q)^{d_{max}}} + (M-N)\log(1-q)\end{aligned}$$

where $\boldsymbol{M = \sum_{d=1}^{max(d)} f(d)\, d}$ and $\boldsymbol{N}$ is defined as above.
Then, we obtain the log-likelihood function of Model 1 as a particular case of this formula in which $\boldsymbol{d_{max} = \infty}$. Given that $\boldsymbol{q > 0}$ and thus $\boldsymbol{\lim_{d_{max}\to\infty}(1-q)^{d_{max}} = 0}$:

$$\mathcal{L} = N \log q + (M - N) \log(1 - q)$$

where $\boldsymbol{N}$ and $\boldsymbol{M}$ are defined as above.
For the *two-regime geometric models*, we proceed in the same way, starting from the log-likelihood function of Model 4:

$$\begin{aligned}
\mathcal{L} &= \sum_{d=1}^{d^*} f(d) \log\left[c_1(1-q_1)^{d-1}\right] + \sum_{d=d^*+1}^{max(d)} f(d) \log\left[c_2(1-q_2)^{d-1}\right] \\
&= \sum_{d=1}^{d^*} f(d)\left[\log c_1 + (d-1)\log(1-q_1)\right] + \sum_{d=d^*+1}^{max(d)} f(d)\left[\log c_2 + (d-1)\log(1-q_2)\right] \\
&= N^* \log c_1 + (M^* - N^*)\log(1-q_1) + (N - N^*)\log c_2 + \\
&+ (M - M^* - N + N^*)\log(1-q_2) \\
&= N^* \log c_1 + (N - N^*)\log c_2 + (M^* - N^*)\log\frac{1-q_1}{1-q_2} + (M - N)\log(1-q_2)
\end{aligned}$$

where

$$M^* = \sum_{d=1}^{d^*} f(d)\, d$$
$$N^* = \sum_{d=1}^{d^*} f(d),$$

and $\boldsymbol{c_1}$ and $\boldsymbol{c_2}$ are computed from equations (6) and (6). Again, by setting $\boldsymbol{d_{max} = \infty}$ in the computation of $\boldsymbol{c_1}$ given by equation (5), we get the normalization term of Model 3, given by equation (5). Thus, the log-likelihood functions of Model 3 and Model 4 only differ in the computation of $\boldsymbol{c_1}$ and $\boldsymbol{c_2}$.

For the *right truncated Zeta model*, namely Model 5:

$$\begin{aligned}\mathcal{L} &= \sum_{d=1}^{max(d)} f(d) \log \frac{d^{-\gamma}}{H(d_{max}, \gamma)} \\ &= \sum_{d=1}^{max(d)} f(d) \left[-\gamma \log d - (d_{max}, \gamma)\right] \\ &= -\gamma M' - N \log H(d_{max}, \gamma)\end{aligned}$$

where $M' = \sum_{d=1}^{max(d)} f(d) \log(d)$ and N is defined as above.

Finally, for Models 6 and 7, we start again from the derivation of the log-likelihood of the right-truncated version:

$$\begin{aligned}\mathcal{L} &= \sum_{d=1}^{d^*} f(d) \log(r_1 d^{-\gamma}) + \sum_{d=d^*+1}^{max(d)} f(d) \log \left[r_2 (1-q)^{d-1}\right] \\ &= \sum_{d=1}^{d^*} f(d) \left[\log r_1 - \gamma \log(d)\right] + \sum_{d=d^*+1}^{max(d)} f(d) \left[\log r_2 + (d-1) \log(1-q)\right] \\ &= N^* \log r_1 - \gamma M'^* + (N - N^*) \log r_2 + (M - M^* - N + N^*) \log(1-q)\end{aligned}$$

where r_1 and r_2 are computed as in equations (9) and (7). Then, the formula for Model 6 can be obtained by setting $d_{max} = \infty$ in equation (9).

7.2.2. Parameter estimation

It is well-known that maximum likelihood estimation is highly sensitive to a good choice of the starting values, as it may incur in local optima when minimizing the minus log-likelihood function (Myung, 2003). We here describe the criteria used to select the initial and bound values for the parameters, which are summarised in tables 8 and 9 respectively. Let x_{init} be the initial value of parameter x. Then:

- d_{max}: the maximum observed distance is both the starting point and the lowest admissible value, while there is no upper bound.
- q: the initial value for q in the geometric models is the maximum likelihood estimator, i.e. the inverse of the mean observed distance $1/\bar{d}$. The bounds are set so that q is always strictly included between 0 and 1 to avoid values out of the domain of the log-likelihood function. In Models 6 and 7, the initial value for q is computed by only considering distances greater than d^*, and again taking the maximum likelihood estimator $1/\bar{d}$.
- q_1 and q_2: these two parameters are both initialized by first running a linear

regression of $\log p(d)$ over d, for $d \leq d^*$ in the case of q_{1init}, and for $d \geq d^*$ in the case of q_{2init}. Then, given the two estimated slopes β_1 and β_2, the initial values can be computed as $q_{1init} = 1 - e^{\beta_1}$ and $q_{2init} = 1 - e^{\beta_2}$. Notice that, since the tail of the distribution is noisy, for values of d^* very close to $max(d)$ the estimated slope sometimes results in a 0 or even a positive value: in these situations, the corresponding q_{2init} was set to its lower bound. As in q, the bounds are set so that q is always strictly included between 0 and 1.

- d^*: the initial value is 5, as estimated by the visual inspection of the plots. The parameter is bounded to vary between 2 and $max(d) - 1$, based on the minimum requirement on the size of the two regimes (refer to section 7.2.4). In fact, by setting d^* either to 1 or to $max(d)$, one of the two regimes would only be composed by one isolated observation, from which no trend can be inferred.
- γ: the initial value is computed through a formula to extract the exponent of a power-law (Newman, 2005):

$$\gamma_{init} = 1 + N \left[\sum_{i=1}^{N} \frac{d_i}{min(d)} \right]^{-1}.$$

For Models 6 and 7 this formula is adapted to only consider distances up to d^*.

Table 8. Initial values of parameters for maximum likelihood estimation.

Model	d_{max}	q	q_1	q_2	d^*	γ
0	$max(d)$	-	-	-	-	-
1	-	$1/\bar{d}$	-	-	-	-
2	$max(d)$	$1/\bar{d}$	-	-	-	-
3	-	-	q_{1init}	q_{2init}	5	-
4	$max(d)$	-	q_{1init}	q_{2init}	5	-
5	$max(d)$	-	-	-	-	γ_{init}
6	-	$1/\bar{d}$	-	-	5	γ_{init}
7	$max(d)$	q_{init}	-	-	5	γ_{init}

Table 9. Lower and upper bounds of parameters for maximum likelihood estimation. $\epsilon = 10^{-3}$.

	d_{max}		q		q_1		q_2		d^*		γ	
Model	low	up	low	up	low	up	low	up	low	up	low	up
0	$max(d)$	∞	-	-	-	-	-	-	-	-	-	-
1	-	-	ϵ	$1-\epsilon$	-	-	-	-	-	-	-	-
2	$max(d)$	∞	ϵ	$1-\epsilon$	-	-	-	-	-	-	-	-
3	-	-	-	-	ϵ	$1-\epsilon$	ϵ	$1-\epsilon$	2	$max(d)-1$	-	-
4	$max(d)$	∞	-	-	ϵ	$1-\epsilon$	ϵ	$1-\epsilon$	2	$max(d)-1$	-	-
5	$max(d)$	∞	-	-	-	-	-	-	-	-	0	∞
6	-	-	ϵ	$1-\epsilon$	-	-	-	-	2	$max(d)-1$	0	∞
7	$max(d)$	∞	ϵ	$1-\epsilon$	-	-	-	-	2	$max(d)-1$	0	∞

7.2.3. Optimization

Optimization of the derived log-likelihood functions was performed considering two optimization frameworks in `R`: `mle()` from `stats2` and `mle2()` from the `bbmle` package. In fact, the base `R` implementation, `mle()`, may explore values out of the specified bounds, thus resulting in errors. For this reason, when this happens we resort to the enhanced, more robust version of the optimizer, `mle2()`, which is able to return a result even when convergence is not reached. Concerning the estimation of parameters d^* and d_{max} in the double-regime models, given that `mle2()` optimizes on a continuous space, we moved to integer estimation by exhaustively exploring all values of included in their theoretical bounds. In this way, we also decrease optimization complexity by reducing the number of parameters to be optimized through the call to `mle2()`. Thus, for each value of d^* (and d_{max} in the right-truncated models) we optimized the remaining parameters, and finally selected the parameters combination resulting in the highest log-likelihood.

7.2.4. Requirements for two-regime models

We need to make a technical point on the models with two regimes in the context of model selection. In order to fit a double-regime model to a data sample, we need to observe at least 3 distinct values of d in it. In fact, if we measure the size of a regime in terms of the unique d values that follow its probability distribution, a minimum requirement of size 2 for each of the two regimes implies 3 unique values of d in the sample, as the value assigned to the break-point is common to the two regimes. Figure 12 shows an example of this scenario, displaying the distribution of syntactic dependency distances for sentences of 4 words in Italian, annotated according to SUD. Notice that this requirement directly implies that sentences with $n < 4$ are excluded from the model selection procedure.

7.3. Speed of decay

When plotted in log-linear scale, a geometric curve has a linear shape. The slope of the resulting line can be computed as a function of parameter q in equation (2), as follows:

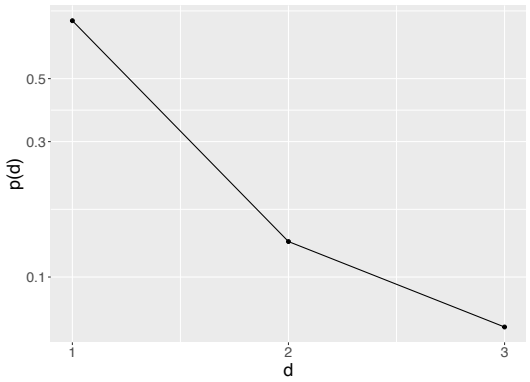


Figure 12.Syntactic dependency distance distribution of sentences with 4 words in Italian, annotated according to SUD. Only three unique values of $\boldsymbol{d}$ have been observed.

$$\begin{aligned}\log p(d) &= \log q(1-q)^{d-1} \\ &= \log q + (d-1)\log(1-q) \\ &= (\log q - \log(1-q)) + d\log(1-q)\end{aligned}$$

Then, $\log(1-q)$ is the slope of the geometric curve in log-linear scale, which gives information on the speed of decay. Figure 7 shows that such slope is a decreasing function of q, meaning that as q increases the slope becomes more negative, and probability decays faster. In light of this fact, we analyse the stability of parameters q_1 and q_2 to account for the speed of exponential decay in the two regimes of Models 3-4, and we refer to them as 'slope parameters' for simplicity. Moreover, we compute the ratio between q_1 and q_2 as a measure of their relative magnitude, namely to understand in which regime probability decays faster. Concerning Models 6-7, recall that the first regime is distributed as a power-law, which only appears as a line in log-log scale. Thus, in order to perform the same analysis for these models, we approximate the speed of decay in the first regime with that of an exponential, for which we can compute the slope. That is, for languages in which Model 6 (7) is the best, we fit Model 3 (4) using their estimated value of d^*, and take the resulting parameter q_1 to compute an estimate of their speed of decay as described above.

7.4. Optimality score Ω

Ω is a recently introduced optimality score for the closeness of syntactic dependency distances, which provides a normalization with respect to the minimum and the random baseline. The score is computed as:

$$\boldsymbol{\Omega} = \frac{\boldsymbol{D_{rla}} - \boldsymbol{D}}{\boldsymbol{D_{rla}} - \boldsymbol{D_{min}}}$$

where $\boldsymbol{D}$ is the observed sum of dependency distances in a sentence, $\boldsymbol{D_{rla}}$ is the expected sum of dependency distances in a uniformly random linear arrangement of words (keeping the relations fixed) (Ferrer-i-Cancho, 2004, 2019), and $\boldsymbol{D_{min}}$ is the sum of dependency distances in a minimum linear arrangement of words (Esteban and Ferrer-i-Cancho, 2017; Shiloach, 1979), again given the network structure. All these tree values are computed using the `python` implementation of the Linear Arrangement Library (Alemany-Puig et al., 2021). Positive values of $\boldsymbol{\Omega}$ indicate that syntactic dependency distances in the sentence are shorter than expected in an ordering picked uniformly at random among all the possible $\boldsymbol{n!}$ orderings, with 1 being the maximum, reached when $\boldsymbol{D}$ is equal to its value in the minimum linear arrangement of the words. Conversely, a negative value indicates that distances are maximized rather than minimized, as they are higher than expected by chance in a random shuffling of words in the sentence. When word order is random, $\boldsymbol{\Omega}$ will take values tending to 0. Given that $\boldsymbol{\Omega}$ is computed for a sentence, we obtain the score for a given sentence length by averaging over all sentences with such length in each language, thus obtaining $\langle\boldsymbol{\Omega}\rangle$.

References

Alemany-Puig, L., Esteban, J., and Ferrer-i-Cancho, R. (2021). The Linear Arrangement Library. A new tool for research on syntactic dependency structures. In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pages 1–16, Sofia, Bulgaria. Association for Computational Linguistics.

Anderson, D. and Burnham, K. (2004). Model selection and multi-model inference. *Second. NY: Springer-Verlag*, 63(2020):10.

Christiansen, M. and Chater, N. (2015). The now-or-never bottleneck: A fundamental constraint on language. *The Behavioral and brain sciences*, pages 1–52.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–114.

Dagpunar, J., Dagpunar, J., et al. (1988). *Principles of random variate generation*. Oxford University Press, USA.

Devroye, L. (1986). *Non-Uniform Random Variate Generation(originally published with.* Springer-Verlag.

Esteban, J. and Ferrer-i-Cancho, R. (2017). A correction on Shiloach's algorithm for minimum linear arrangements of trees. *Society for Industrial and Applied Mathematics*, 46(3):1146–1151.

Ferrer-i-Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E*, 70:056135.

Ferrer-i-Cancho, R. (2017). A commentary on "the now-or-never bottleneck: a fundamental constraint on language", by christiansen and chater (2016).

Ferrer-i-Cancho, R. (2019). The sum of edge lengths in random linear arrangements. *Journal of Statistical Mechanics: Theory and Experiment*, 2019:053401.

Ferrer-i Cancho, R. and Gómez-Rodríguez, C. (2021). Anti dependency distance minimization in short sequences. a graph theoretic approach. *Journal of Quantitative Linguistics*, 28(1):50–76.

Ferrer-i Cancho, R., Gómez-Rodríguez, C., Esteban, J. L., and Alemany-Puig, L. (2022). Optimality of syntactic dependency distances. *Physical Review E*, 105(1):014308.

Ferrer-i Cancho, R. and Liu, H. (2014). The risks of mixing dependency lengths from sequences of different length. *Glottotheory*, 5(2):143–155.

Futrell, R., Mahowald, K., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112:10336 – 10341.

Gerdes, K., Guillaume, B., Kahane, S., and Perrier, G. (2018). SUD or surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium. Association for Computational Linguistics.

Gildea, D. and Temperley, D. (2007). Optimizing grammars for minimum dependency

length. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 184–191, Prague, Czech Republic. Association for Computational Linguistics.

Haitao, L. Q. L. (2016). Does dependency distance distribute regularly. *Journal of Zhejiang University (Humanities and Social Science)*, 2(4).

Henderson, L. (1972). Spatial and verbal codes and the capacity of stm. *Quarterly Journal of Experimental Psychology*, 24(4).

Jiang, J. and 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*, 50:93–104.

Lewis, R. L. and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

Liu, H. (2007). Probability distribution of dependency distance. *Glottometrics*, 15:1–12.

Liu, H., Xu, C., and Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of life reviews*, 21:171–193.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63(2):81–97.

Muller, M. E. (1958). An inverse method for the generation of random normal deviates on large-scale computers. *Mathematics of Computation*, 12(63):167–174.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47(1):90–100.

Newman, M. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):323–351.

Nivre, J., Zeman, D., Ginter, F., and Tyers, F. (2017). Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Shiloach, Y. (1979). A minimum linear arrangement algorithm for undirected trees. Master's thesis.

Stumpf, M. P. H. and Porter, M. A. (2012). Critical truths about power laws. *Science*, 335(6069):665–666.

Wagenmakers, E.-J. and Farrell, S. (2004). Aic model selection using akaike weights. *Psychonomic bulletin & review*, 11(1):192–196.

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

8. Appendices

Appendix A. Models Derivation

In the following section we describe how the mathematical formulation of the models has been derived.

A.1. Model 2

Recall that the cumulative distribution of Model $\boldsymbol{x}$ is

$$\boldsymbol{P_x(N) = \sum_{d=1}^{N} p_x(d).}$$

Model 2 is derived in the following way:

$$\boldsymbol{p_2(d) = \frac{p_1(d)}{P_1(d_{max})}}$$

where

$$\begin{aligned}
\boldsymbol{P_1(d_{max})} &= \boldsymbol{\sum_{d=1}^{d_{max}} q(1-q)^{d-1}} \\
&= \boldsymbol{\sum_{d'=0}^{d_{max}-1} q(1-q)^{d'}} \\
&= \boldsymbol{q\frac{1-(1-q)^{d_{max}}}{q}} \\
&= \boldsymbol{1-(1-q)^{d_{max}}}
\end{aligned}$$

so that:

$$\boldsymbol{p_2(d) = \frac{q(1-q)^{d-1}}{1-(1-q)^{d_{max}}}}$$

A.2. Double-regime models

Let us derive the first normalization factor $\boldsymbol{c_1}$ for Models 3, 4, 6, 7, recalling that $\boldsymbol{c_2}$ can be computed as $\boldsymbol{c_2 = \tau c_1}$, where $\boldsymbol{\tau}$ is computed as in equation (4) for Models 3-4, and as in equation (8) for Models 6-7.

A.2.1. Models 3 and 4

For Model 3, we can write

$$\sum_{d=1}^{\infty} p_3(d) = 1 \iff c_1 (S_1 + \tau S_2) = 1$$

while for Model 4

$$\sum_{d=1}^{d_{max}} p_4(d) = 1 \iff c_1 (S_1 + \tau S_2) = 1$$

meaning that for both models

$$c_1 = \frac{1}{S_1 + \tau S_2}. \tag{A1}$$

where τ can be obtained from equation (4), and, by setting $d' = d - 1$, S_1 is computed as:

$$S_1 = \sum_{d'=0}^{d^*-1} (1 - q_1)^{d'} = \frac{1 - (1 - q_1)^{d^*}}{q_1}$$

while S_2 depends on the truncation point, so that for Model 3 it is computed as:

$$\begin{aligned} S_2 &= \sum_{d'=d^*}^{\infty} (1 - q_2)^{d'} \\ (1 - q_2) S_2 &= S_2 - (1 - q_2)^{d*} + (1 - q_2)^{\infty} \\ S_2 &= \frac{(1 - q_2)^{\infty} - (1 - q_2)^{d*}}{1 - q_2 - 1} \\ S_2 &= \frac{(1 - q_2)^{d^*}}{q_2} \end{aligned}$$

where we used the fact that $q > 0$ and thus $\lim_{d_{max} \to \infty} (1 - q)^{d_{max}} = 0$.
Then, by substituting S_1, S_2 and τ in (A1) the first normalization factor c_1 for Model 3 can be computed as:

$$
\begin{aligned}
c_1 &= \frac{1}{\frac{1-(1-q_1)^{d^*}}{q_1} + \frac{(1-q_1)^{d^*-1}}{(1-q_2)^{d^*-1}}\frac{(1-q_2)^{d^*}}{q_2}} \\
&= \frac{1}{\frac{1-(1-q_1)^{d^*}}{q_1} + \frac{(1-q_1)^{d^*-1}(1-q_2)}{q2}} \\
&= \frac{q_1 q_2}{q_2 - q_2(1-q_1)^{d^*} + q_1(1-q_1)^{d^*-1}(1-q_2)} \\
&= \frac{q_1 q_2}{q_2 + (1-q_1)^{d^*-1}(q_1 - q_2)}
\end{aligned}
$$

On the other hand, for Model 4 the probabilities are restricted up to d_{max}, thus:

$$
S_2 = \sum_{d'=d^*}^{d_{max}-1} (1-q_2)^{d'} = \frac{(1-q_2)^{d*} - (1-q_2)^{d_{max}}}{q_2}
$$

And again S_1, S_2, and τ can be plugged in equation (A1) to compute c_1 for Model 4:

$$
\begin{aligned}
c_1 &= \frac{1}{\frac{1-(1-q_1)^{d^*}}{q_1} + \frac{(1-q_1)^{d^*-1}}{(1-q_2)^{d^*-1}}\frac{(1-q_2)^{d^*} - (1-q_2)^{d_{max}}}{q_2}} \\
&= \frac{q_1 q_2}{q_2 - q_2(1-q_1)^{d^*} + q_1(1-q_1)^{d^*-1}(1-q_2-(1-q_2)^{d_{max}-d^*+1})} \\
&= \frac{q_1 q_2}{q2 + (1-q_1)^{d^*-1}\left[-q_2(1-q_1) + q_1(1-q_2-(1-q_2)^{d_{max}-d^*+1})\right]} \\
&= \frac{q_1 q_2}{q2 + (1-q_1)^{d^*-1}(q_1 - q_2 - q_1(1-q_2)^{d_{max}-d^*+1})}
\end{aligned}
$$

A.2.2. Models 6 and 7

For the second pair of double-regime models, combining a Zeta and a geometric distribution, once again

$$
\sum_{d=1}^{\infty} p_6(d) = 1 \iff c_1 \left(S_1 + \tau S_2\right) = 1
$$

for Model 6, while for Model 7

$$
\sum_{d=1}^{d_{max}} p_7(d) = 1 \iff c_1 \left(S_1 + \tau S_2\right) = 1
$$

meaning that equation (A1) still holds. Then, for both models $\boldsymbol{S_1}$ is computed as

$$\boldsymbol{S_1} = \sum_{\boldsymbol{d=1}}^{\boldsymbol{d^*}} \boldsymbol{d^{-\gamma}} = \boldsymbol{H(d^*, \gamma)}$$

while the second regime is shared with Models 3 and 4, so that $\boldsymbol{S_2}$ corresponds to equations and for Model 6 and Model 7 respectively.
Then, the normalization factors are obtained again through equation (A1), so that for Model 6

$$\begin{aligned} \boldsymbol{c_1} &= \frac{\boldsymbol{1}}{\boldsymbol{H(d^*, \gamma)} + \frac{\boldsymbol{d^*-\gamma}}{\boldsymbol{(1-q)^{d^*-1}}} \frac{\boldsymbol{(1-q)^{d^*}}}{\boldsymbol{q}}} \\ \boldsymbol{c_1} &= \frac{\boldsymbol{q}}{\boldsymbol{qH(d^*, \gamma) + d^{*-\gamma}(1-q)}} \end{aligned}$$

while for Model 7

$$\begin{aligned} \boldsymbol{c_1} &= \frac{\boldsymbol{1}}{\boldsymbol{H(d^*, \gamma)} + \frac{\boldsymbol{d^*-\gamma}}{\boldsymbol{(1-q)^{d^*-1}}} \frac{\boldsymbol{(1-q)^{d^*}-(1-q)^{d_{max}}}}{\boldsymbol{q}}} \\ \boldsymbol{c_1} &= \frac{\boldsymbol{q}}{\boldsymbol{qH(d^*, \gamma) + d^{*-\gamma}(1-q-(1-q)^{d_{max}-d^*+1})}}. \end{aligned}$$

Appendix B. Model selection results

B.1. Artificial datasets - methods validation

As table B1 and figure B1 show, the combination of maximum likelihood estimation and BIC manages to identify the real underlying distribution for every artificial random sample. Figure B1 (b) allows one to quantify the magnitude of the differences in BIC score between the best model and the competing ones for each sample. As we can see, while BIC for Model 0 is always very high (in any other sample), the double-regime models have low differences with respect to the best one, also comparing with those of the one-regime geometric and power-law models. The reason resides in the greater flexibility allowed by the existence of the break-point, which is however compensated by the penalty imposed on the additional parameters. Another concern could rise from the fitting of the random sample of Model 0, in which the BIC score of Model 4 is not that far from the best one. Indeed, two geometric regimes could mimic the linearity of Model 0, but only in the case in which the second regime decays faster than the second. As we discussed in section 4, this is virtually never the case in any language and collection, so that in our analysis we can exclude the risk of falsely identifying a double regime model when distances are randomly distributed. Table B3 displays parameters estimated by maximum likelihood for each of the artificial samples, the real values used to generate the data, and their difference. The error between the real values and the optimal estimated parameters is either 0 or very small. In particular, maximum likelihood estimation seems more likely to underestimate the real value rather than the opposite.

Table B1.BIC score values for model selection on artificial data. Every row is the random sample from the distribution of a model, every column is a fitted model.

sample	Model 0	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
0	55570.65	57527.90	55881.07	55750.98	55615.79	56724.85	55963.35	55697.17
1	60974.42	50037.40	50040.08	50049.99	50056.13	53256.86	50054.30	50057.24
2	51569.46	48995.12	48739.88	48801.18	48755.09	50086.11	48993.61	48757.98
3	76657.57	54995.13	55004.33	51553.49	51561.32	52694.62	51681.76	51689.79
4	51638.78	47122.05	46967.51	45359.06	44595.90	44716.30	44818.89	44685.95
5	49460.37	39609.04	39602.60	37251.07	37076.27	36864.76	36937.93	36881.25
6	61658.20	39436.89	39446.10	37196.76	37204.83	37684.75	37133.38	37141.30
7	48909.08	38217.08	38217.08	36343.48	36242.39	36270.85	36239.71	36175.27

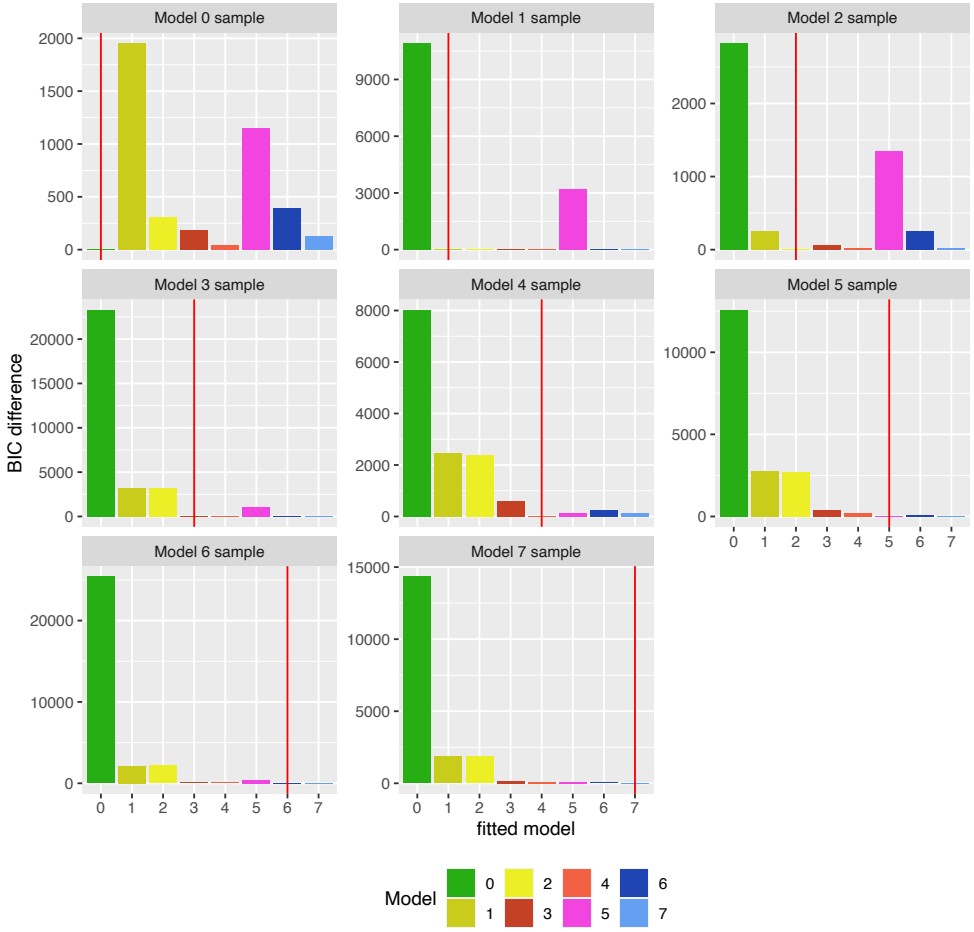


Figure B1. BIC differences for each model in a sample generated by some Model. Red lines show the best model according to BIC.

Table B2. Best parameters estimated in artificial samples by maximum likelihood.

		0	1	2		3			4				5		6			7			
model	$max(d)$	d_{max}	q	q	d_{max}	q_1	q_2	d^*	q_1	q_2	d^*	d_{max}	d_{max}	γ	γ	q	d^*	γ	q	d^*	d_{max}
0	19	19	0.142	0.100	19	0.088	0.643	17	0.071	0.242	13	19	19	0.522	0.302	0.343	13	0.264	0.206	11	19
1	36	36	0.200	0.200	36	0.200	0.543	34	0.191	0.203	5	36	36	1.204	0.274	0.201	2	0.278	0.201	2	36
2	19	19	0.210	0.197	19	0.197	0.622	18	0.199	0.091	17	19	19	0.985	0.418	0.221	4	0.295	0.198	2	19
3	85	85	0.160	0.160	85	0.502	0.101	4	0.502	0.101	4	85	85	1.422	1.373	0.103	5	1.373	0.102	5	85
4	19	19	0.228	0.219	19	0.549	0.166	3	0.503	0.101	4	19	19	1.242	1.234	0.644	18	1.332	0.097	6	19
5	19	19	0.314	0.313	19	0.628	0.201	3	0.641	0.180	3	19	19	1.582	1.578	0.610	18	1.588	0.058	15	19
6	40	40	0.317	0.317	40	0.623	0.204	3	0.624	0.204	3	40	40	1.718	1.613	0.201	4	1.614	0.201	4	40
7	19	19	0.333	0.332	19	0.613	0.221	3	0.622	0.206	3	19	19	1.608	1.541	0.299	12	1.610	0.202	4	19

Table B3. Best estimated parameters, real parameters used to generate the artificial samples, and their difference.

	0	1	2		3			4				5		6			7			
	d_{max}	q	q	d_{max}	q_1	q_2	d^*	q_1	q_2	d^*	d_{max}	d_{max}	γ	γ	q	d^*	γ	q	d^*	d_{max}
estimated	19.000	0.200	0.197	19.000	0.502	0.101	4.000	0.503	0.101	4.000	19.000	19.000	1.582	1.613	0.201	4.000	1.610	0.202	4.000	19.000
real	19.000	0.200	0.200	19.000	0.500	0.100	4.000	0.500	0.100	4.000	19.000	19.000	1.600	1.600	0.200	4.000	1.600	0.200	4.000	19.000
error	0.000	0.000	-0.003	0.000	0.002	0.001	0.000	0.003	0.001	0.000	0.000	0.000	-0.018	0.013	0.001	0.000	0.010	0.002	0.000	0.000

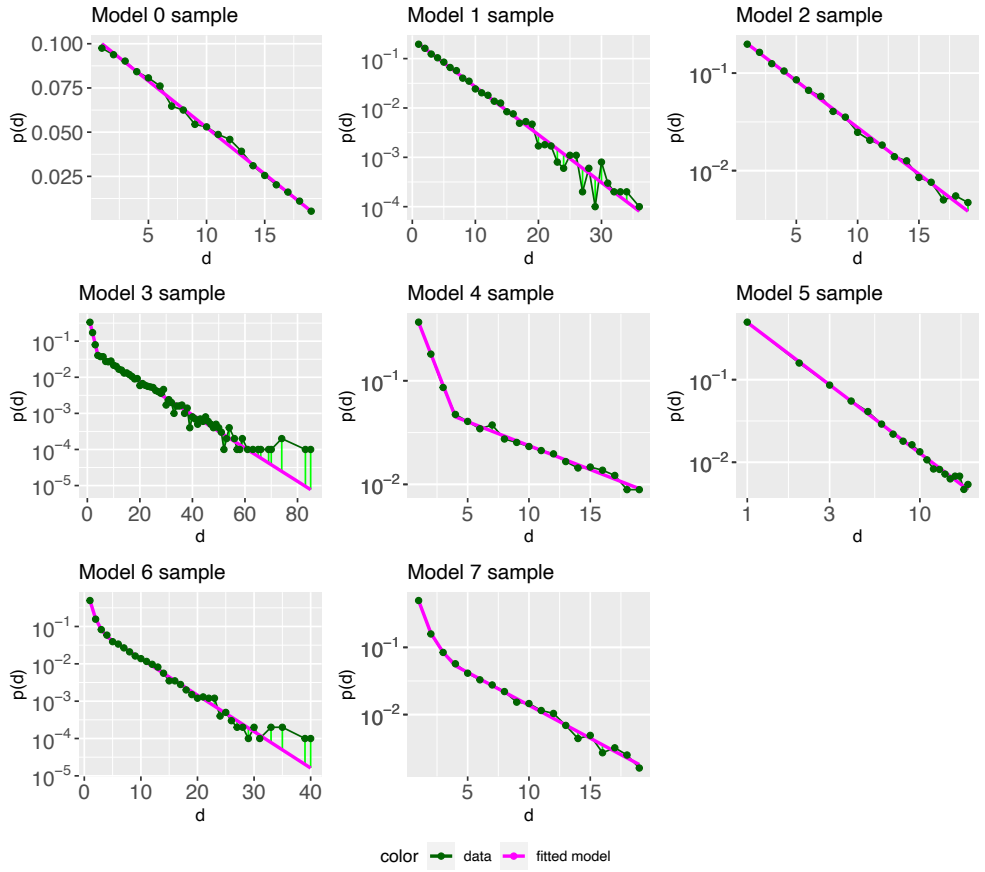


Figure B2. $\boldsymbol{p(d)}$, the probability that a dependency link is formed between words at distance $\boldsymbol{d}$ according to the best model for artificially generated samples (generated by some Model).

B.2. Real languages

We here report the results of model selection on real languages, with all sentence lengths are considered jointly for a given language. Tables B4 and B8 show a summary of PUD and PSUD respectively. Tables B5 and B9 show AIC scores, while tables B6 and B10 show AIC differences, again for PUD and PSUD respectively. Finally, parameters estimated through maximum likelihood are shown in table B7 for PUD and in table B11 for PSUD. Finally, figure B3 shows the best model fitted to the empirical distribution for languages in PSUD.

Table B4. Summary of PUD collection.

language	family	script	no. sentences	no. distances	$min(d)$	$\bar{d}$	$max(d)$	$min(n)$	$\bar{n}$	$max(n)$
Arabic	Afro-Asiatic	Arabic	995	17514	1	2.30	30	3	18.60	50
Czech	Indo-European	Latin	995	14976	1	2.39	29	3	16.05	44
German	Indo-European	Latin	995	17544	1	3.11	42	4	18.63	50
English	Indo-European	Latin	995	17711	1	2.53	31	4	18.80	56
Finnish	Uralic	Latin	995	12465	1	2.24	21	3	13.53	39
French	Indo-European	Latin	995	21165	1	2.52	36	4	22.27	54
Hindi	Indo-European	Devanagari	995	20517	1	3.30	42	4	21.62	58
Indonesian	Austronesian	Latin	995	16311	1	2.26	27	3	17.39	47
Icelandic	Indo-European	Latin	995	15860	1	2.32	34	3	16.94	52
Italian	Indo-European	Latin	995	20413	1	2.48	35	3	21.52	60
Japanese	Japonic	Japanese	995	24703	1	2.97	65	4	25.83	70
Korean	Koreanic	Hangul	995	13978	1	2.75	37	3	15.05	43
Polish	Indo-European	Latin	995	14720	1	2.23	27	3	15.79	39
Portuguese	Indo-European	Latin	995	19808	1	2.53	34	4	20.91	58
Russian	Indo-European	Cyrillic	995	15369	1	2.27	32	3	16.45	47
Spanish	Indo-European	Latin	995	19986	1	2.50	32	3	21.09	58
Swedish	Indo-European	Latin	995	16119	1	2.47	31	4	17.20	49
Thai	Kra-Dai	Thai	995	21034	1	2.44	38	4	22.14	63
Turkish	Turkic	Latin	995	13727	1	2.91	34	3	14.80	37
Chinese	Sino-Tibetan	Han	995	17501	1	3.09	39	3	18.59	49

Table B5. AIC scores in PUD collection, obtained in model selection on mixed lengths.

language	Model 0	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Arabic	84577.28	55188.27	55190.27	52010.89	52011.60	52263.92	51865.76	51864.32
Chinese	86281.29	68120.55	68122.54	65825.64	65826.94	67025.49	65736.87	65737.59
Czech	68423.96	48729.14	48731.14	47871.62	47872.33	49499.07	48211.69	48213.59
English	85820.84	60122.24	60124.24	59401.60	59402.22	62649.37	60054.51	60056.49
Finnish	52892.55	38423.16	38425.06	37955.02	37956.33	38920.99	37927.05	37928.08
French	109197.30	71747.81	71749.81	69419.89	69417.95	72291.26	70944.37	70946.28
German	86626.06	68509.94	68511.94	66698.75	66700.24	68821.36	66955.38	66957.09
Hindi	107388.19	83074.52	83076.51	75832.00	75827.97	76676.12	75788.08	75782.47
Icelandic	73751.52	50241.87	50243.87	49410.74	49412.68	51252.39	49716.26	49718.26
Indonesian	76351.16	50675.51	50677.50	48875.12	48876.20	49596.23	48916.29	48916.99
Italian	104223.33	68312.61	68314.61	66369.64	66368.55	69289.09	67786.29	67788.22
Japanese	135512.12	93745.97	93747.97	85221.98	85221.41	87112.25	86524.36	86524.91
Korean	64172.53	50365.39	50367.39	45473.71	45472.32	45646.54	45336.72	45332.41
Polish	66254.58	45102.63	45104.63	43851.38	43852.11	44718.72	43956.07	43957.64
Portuguese	100042.15	67212.94	67214.94	65360.72	65361.05	68009.88	66557.18	66559.05
Russian	70473.64	47879.18	47881.18	46749.90	46751.27	48290.71	47201.27	47203.22
Spanish	101194.17	67352.99	67354.99	65377.31	65376.25	67933.66	66641.12	66642.84
Swedish	75638.74	53806.67	53808.67	53135.13	53136.01	55622.73	53632.92	53634.89
Thai	108080.67	69552.76	69554.76	65716.52	65718.28	66241.97	65519.47	65520.69
Turkish	62863.72	51438.91	51440.89	47361.98	47357.68	47697.09	47250.44	47244.94

Table B6. AIC differences for each model in the PUD collection, obtained in model selection on mixed lengths. Namely, the difference in AIC between a model and the best one, that is the one with minimum AIC.

language	Model 0	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Arabic	32712.96	3323.95	3325.95	146.57	147.28	399.60	1.44	0.00
Chinese	20544.43	2383.68	2385.68	88.77	90.07	1288.62	0.00	0.72
Czech	20552.34	857.52	859.51	0.00	0.70	1627.44	340.07	341.97
English	26419.25	720.65	722.64	0.00	0.63	3247.77	652.91	654.90
Finnish	14965.50	496.10	498.00	27.96	29.28	993.94	0.00	1.02
French	39779.35	2329.86	2331.86	1.94	0.00	2873.31	1526.42	1528.33
German	19927.31	1811.19	1813.19	0.00	1.49	2122.61	256.63	258.34
Hindi	31605.72	7292.04	7294.03	49.53	45.49	893.65	5.61	0.00
Icelandic	24340.78	831.13	833.13	0.00	1.94	1841.66	305.52	307.52
Indonesian	27476.04	1800.38	1802.38	0.00	1.08	721.10	41.17	41.86
Italian	37854.79	1944.06	1946.06	1.10	0.00	2920.54	1417.74	1419.67
Japanese	50290.71	8524.56	8526.56	0.58	0.00	1890.84	1302.96	1303.51
Korean	18840.12	5032.98	5034.98	141.29	139.91	314.13	4.31	0.00
Polish	22403.20	1251.25	1253.25	0.00	0.73	867.34	104.69	106.27
Portuguese	34681.44	1852.22	1854.22	0.00	0.34	2649.16	1196.46	1198.33
Russian	23723.74	1129.28	1131.28	0.00	1.37	1540.80	451.36	453.32
Spanish	35817.93	1976.75	1978.74	1.07	0.00	2557.41	1264.87	1266.59
Swedish	22503.61	671.54	673.54	0.00	0.88	2487.59	497.78	499.75
Thai	42561.20	4033.29	4035.29	197.05	198.82	722.50	0.00	1.23
Turkish	15618.79	4193.97	4195.96	117.05	112.74	452.16	5.51	0.00

Table B7. Best parameters estimated by maximum likelihood on mixed sentence lengths in PUD collection.

			1	2		3			4				5		6			7			
language	$max(n)$	$max(d)$	q	q	d_{max}	q_1	q_2	d^*	q_1	q_2	d^*	d_{max}	d_{max}	γ	γ	q	d^*	γ	q	d^*	d_{max}
Arabic	50	30	0.434	0.434	30	0.668	0.269	3	0.668	0.269	3	30	30	1.973	1.840	0.240	7	1.842	0.238	7	30
Czech	44	29	0.418	0.418	29	0.499	0.270	5	0.500	0.269	5	29	29	1.837	1.385	0.347	3	1.385	0.347	3	29
German	50	42	0.322	0.322	42	0.485	0.222	4	0.485	0.222	4	42	42	1.675	1.351	0.234	5	1.351	0.234	5	42
English	56	31	0.395	0.395	31	0.453	0.256	6	0.454	0.255	6	31	31	1.759	0.930	0.380	2	0.930	0.380	2	31
Finnish	39	21	0.446	0.446	21	0.639	0.388	2	0.562	0.360	3	21	21	1.855	1.440	0.374	3	1.440	0.374	3	21
French	54	36	0.396	0.396	36	0.491	0.197	6	0.492	0.196	6	36	36	1.826	1.462	0.296	4	1.462	0.296	4	36
Hindi	58	42	0.303	0.303	42	0.671	0.175	3	0.672	0.174	3	42	42	1.735	1.798	0.170	4	1.800	0.169	4	42
Indonesian	47	27	0.442	0.442	27	0.629	0.305	3	0.630	0.304	3	27	27	1.935	1.713	0.295	5	1.714	0.294	5	27
Icelandic	52	34	0.432	0.432	34	0.531	0.309	4	0.531	0.309	4	34	34	1.888	1.412	0.359	3	1.412	0.359	3	34
Italian	60	35	0.403	0.403	35	0.490	0.207	6	0.491	0.206	6	35	35	1.830	1.446	0.307	4	1.446	0.307	4	35
Japanese	70	65	0.337	0.337	65	0.521	0.119	6	0.522	0.118	6	65	65	1.849	1.754	0.130	13	1.755	0.130	13	65
Korean	43	37	0.364	0.364	37	0.700	0.197	3	0.701	0.197	3	37	37	1.886	1.886	0.180	5	1.888	0.179	5	37
Polish	39	27	0.449	0.449	27	0.569	0.288	4	0.569	0.287	4	27	27	1.936	1.653	0.324	4	1.653	0.324	4	27
Portuguese	58	34	0.396	0.396	34	0.504	0.234	5	0.504	0.233	5	34	34	1.812	1.436	0.301	4	1.436	0.301	4	34
Russian	47	32	0.440	0.440	32	0.529	0.261	5	0.529	0.260	5	32	32	1.916	1.564	0.332	4	1.564	0.332	4	32
Spanish	58	32	0.399	0.399	32	0.510	0.231	5	0.510	0.230	5	32	32	1.816	1.460	0.300	4	1.460	0.300	4	32
Swedish	49	31	0.404	0.404	31	0.462	0.257	6	0.462	0.257	6	31	31	1.791	1.214	0.358	3	1.214	0.358	3	31
Thai	63	38	0.409	0.409	38	0.653	0.258	3	0.653	0.258	3	38	38	1.933	1.770	0.230	7	1.770	0.230	7	38
Turkish	37	34	0.343	0.343	34	0.670	0.201	3	0.671	0.200	3	34	34	1.797	1.812	0.195	4	1.815	0.194	4	34
Chinese	49	39	0.323	0.323	39	0.569	0.233	3	0.569	0.232	3	39	39	1.694	1.439	0.219	6	1.440	0.219	6	39

Table B8. Summary of PSUD collection.

language	family	script	no. sentences	no. distances	d_{min}	$\bar{d}$	d_{max}	n_{min}	$\bar{n}$	n_{max}
Arabic	Afro-Asiatic	Arabic	995	17514	1	2.05	30	3	18.60	50
Czech	Indo-European	Latin	995	14976	1	2.11	29	3	16.05	44
German	Indo-European	Latin	995	17544	1	2.82	38	4	18.63	50
English	Indo-European	Latin	995	17711	1	2.12	31	4	18.80	56
Finnish	Uralic	Latin	995	12465	1	2.04	22	3	13.53	39
French	Indo-European	Latin	995	21165	1	2.13	35	4	22.27	54
Hindi	Indo-European	Devanagari	995	20517	1	3.04	38	4	21.62	58
Indonesian	Austronesian	Latin	995	16311	1	2.00	27	3	17.39	47
Icelandic	Indo-European	Latin	995	15860	1	1.92	34	3	16.94	52
Italian	Indo-European	Latin	995	20413	1	2.10	35	3	21.52	60
Japanese	Japonic	Japanese	995	24703	1	2.73	67	4	25.83	70
Korean	Koreanic	Hangul	995	13978	1	2.70	38	3	15.05	43
Polish	Indo-European	Latin	995	14720	1	2.00	27	3	15.79	39
Portuguese	Indo-European	Latin	995	19808	1	2.13	34	4	20.91	58
Russian	Indo-European	Cyrillic	995	15369	1	2.05	32	3	16.45	47
Spanish	Indo-European	Latin	995	19986	1	2.13	31	3	21.09	58
Swedish	Indo-European	Latin	995	16119	1	2.07	31	4	17.20	49
Thai	Kra-Dai	Thai	995	21034	1	2.20	39	4	22.14	63
Turkish	Turkic	Latin	995	13727	1	2.86	33	3	14.80	37
Chinese	Sino-Tibetan	Han	995	17501	1	2.99	39	3	18.59	49

Table B9. AIC scores in PSUD collection, used for model selection on mixed lengths.

language	Model 0	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Arabic	84577.28	55188.27	55190.27	52010.89	52011.60	52263.92	51865.76	51864.32
Chinese	86281.29	68120.55	68122.54	65825.64	65826.94	67025.49	65736.87	65737.59
Czech	68423.96	48729.14	48731.14	47871.62	47872.33	49499.07	48211.69	48213.59
English	85820.84	60122.24	60124.24	59401.60	59402.22	62649.37	60054.51	60056.49
Finnish	52892.55	38423.16	38425.06	37955.02	37956.33	38920.99	37927.05	37928.08
French	109197.30	71747.81	71749.81	69419.89	69417.95	72291.26	70944.37	70946.28
German	86626.06	68509.94	68511.94	66698.75	66700.24	68821.36	66955.38	66957.09
Hindi	107388.19	83074.52	83076.51	75832.00	75827.97	76676.12	75788.08	75782.47
Icelandic	73751.52	50241.87	50243.87	49410.74	49412.68	51252.39	49716.26	49718.26
Indonesian	76351.16	50675.51	50677.50	48875.12	48876.20	49596.23	48916.29	48916.99
Italian	104223.33	68312.61	68314.61	66369.64	66368.55	69289.09	67786.29	67788.22
Japanese	135512.12	93745.97	93747.97	85221.98	85221.41	87112.25	86524.36	86524.91
Korean	64172.53	50365.39	50367.39	45473.71	45472.32	45646.54	45336.72	45332.41
Polish	66254.58	45102.63	45104.63	43851.38	43852.11	44718.72	43956.07	43957.64
Portuguese	100042.15	67212.94	67214.94	65360.72	65361.05	68009.88	66557.18	66559.05
Russian	70473.64	47879.18	47881.18	46749.90	46751.27	48290.71	47201.27	47203.22
Spanish	101194.17	67352.99	67354.99	65377.31	65376.25	67933.66	66641.12	66642.84
Swedish	75638.74	53806.67	53808.67	53135.13	53136.01	55622.73	53632.92	53634.89
Thai	108080.67	69552.76	69554.76	65716.52	65718.28	66241.97	65519.47	65520.69
Turkish	62863.72	51438.91	51440.89	47361.98	47357.68	47697.09	47250.44	47244.94

Table B10. AIC differences for each model in the PSUD collection, obtained in model selection on mixed lengths. Namely, the difference in AIC between a model and the best one, that is the one with minimum AIC.

language	Model 0	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Arabic	32712.96	3323.95	3325.95	146.57	147.28	399.60	1.44	0.00
Chinese	20544.43	2383.68	2385.68	88.77	90.07	1288.62	0.00	0.72
Czech	20552.34	857.52	859.51	0.00	0.70	1627.44	340.07	341.97
English	26419.25	720.65	722.64	0.00	0.63	3247.77	652.91	654.90
Finnish	14965.50	496.10	498.00	27.96	29.28	993.94	0.00	1.02
French	39779.35	2329.86	2331.86	1.94	0.00	2873.31	1526.42	1528.33
German	19927.31	1811.19	1813.19	0.00	1.49	2122.61	256.63	258.34
Hindi	31605.72	7292.04	7294.03	49.53	45.49	893.65	5.61	0.00
Icelandic	24340.78	831.13	833.13	0.00	1.94	1841.66	305.52	307.52
Indonesian	27476.04	1800.38	1802.38	0.00	1.08	721.10	41.17	41.86
Italian	37854.79	1944.06	1946.06	1.10	0.00	2920.54	1417.74	1419.67
Japanese	50290.71	8524.56	8526.56	0.58	0.00	1890.84	1302.96	1303.51
Korean	18840.12	5032.98	5034.98	141.29	139.91	314.13	4.31	0.00
Polish	22403.20	1251.25	1253.25	0.00	0.73	867.34	104.69	106.27
Portuguese	34681.44	1852.22	1854.22	0.00	0.34	2649.16	1196.46	1198.33
Russian	23723.74	1129.28	1131.28	0.00	1.37	1540.80	451.36	453.32
Spanish	35817.93	1976.75	1978.74	1.07	0.00	2557.41	1264.87	1266.59
Swedish	22503.61	671.54	673.54	0.00	0.88	2487.59	497.78	499.75
Thai	42561.20	4033.29	4035.29	197.05	198.82	722.50	0.00	1.23
Turkish	15618.79	4193.97	4195.96	117.05	112.74	452.16	5.51	0.00

Table B11. Best parameters estimated by maximum likelihood on mixed sentence lengths in PSUD collection.

			1	2		3			4				5		6			7			
language	$max(n)$	$max(d)$	q	q	d_{max}	q_1	q_2	d^*	q_1	q_2	d^*	d_{max}	d_{max}	γ	γ	q	d^*	γ	q	d^*	d_{max}
Arabic	50	30	0.434	0.434	30	0.668	0.269	3	0.668	0.269	3	30	30	1.973	1.840	0.240	7	1.842	0.238	7	30
Czech	44	29	0.418	0.418	29	0.499	0.270	5	0.500	0.269	5	29	29	1.837	1.385	0.347	3	1.385	0.347	3	29
German	50	42	0.322	0.322	42	0.485	0.222	4	0.485	0.222	4	42	42	1.675	1.351	0.234	5	1.351	0.234	5	42
English	56	31	0.395	0.395	31	0.453	0.256	6	0.454	0.255	6	31	31	1.759	0.930	0.380	2	0.930	0.380	2	31
Finnish	39	21	0.446	0.446	21	0.639	0.388	2	0.562	0.360	3	21	21	1.855	1.440	0.374	3	1.440	0.374	3	21
French	54	36	0.396	0.396	36	0.491	0.197	6	0.492	0.196	6	36	36	1.826	1.462	0.296	4	1.462	0.296	4	36
Hindi	58	42	0.303	0.303	42	0.671	0.175	3	0.672	0.174	3	42	42	1.735	1.798	0.170	4	1.800	0.169	4	42
Indonesian	47	27	0.442	0.442	27	0.629	0.305	3	0.630	0.304	3	27	27	1.935	1.713	0.295	5	1.714	0.294	5	27
Icelandic	52	34	0.432	0.432	34	0.531	0.309	4	0.531	0.309	4	34	34	1.888	1.412	0.359	3	1.412	0.359	3	34
Italian	60	35	0.403	0.403	35	0.490	0.207	6	0.491	0.206	6	35	35	1.830	1.446	0.307	4	1.446	0.307	4	35
Japanese	70	65	0.337	0.337	65	0.521	0.119	6	0.522	0.118	6	65	65	1.849	1.754	0.130	13	1.755	0.130	13	65
Korean	43	37	0.364	0.364	37	0.700	0.197	3	0.701	0.197	3	37	37	1.886	1.886	0.180	5	1.888	0.179	5	37
Polish	39	27	0.449	0.449	27	0.569	0.288	4	0.569	0.287	4	27	27	1.936	1.653	0.324	4	1.653	0.324	4	27
Portuguese	58	34	0.396	0.396	34	0.504	0.234	5	0.504	0.233	5	34	34	1.812	1.436	0.301	4	1.436	0.301	4	34
Russian	47	32	0.440	0.440	32	0.529	0.261	5	0.529	0.260	5	32	32	1.916	1.564	0.332	4	1.564	0.332	4	32
Spanish	58	32	0.399	0.399	32	0.510	0.231	5	0.510	0.230	5	32	32	1.816	1.460	0.300	4	1.460	0.300	4	32
Swedish	49	31	0.404	0.404	31	0.462	0.257	6	0.462	0.257	6	31	31	1.791	1.214	0.358	3	1.214	0.358	3	31
Thai	63	38	0.409	0.409	38	0.653	0.258	3	0.653	0.258	3	38	38	1.933	1.770	0.230	7	1.770	0.230	7	38
Turkish	37	34	0.343	0.343	34	0.670	0.201	3	0.671	0.200	3	34	34	1.797	1.812	0.195	4	1.815	0.194	4	34
Chinese	49	39	0.323	0.323	39	0.569	0.233	3	0.569	0.232	3	39	39	1.694	1.439	0.219	6	1.440	0.219	6	39

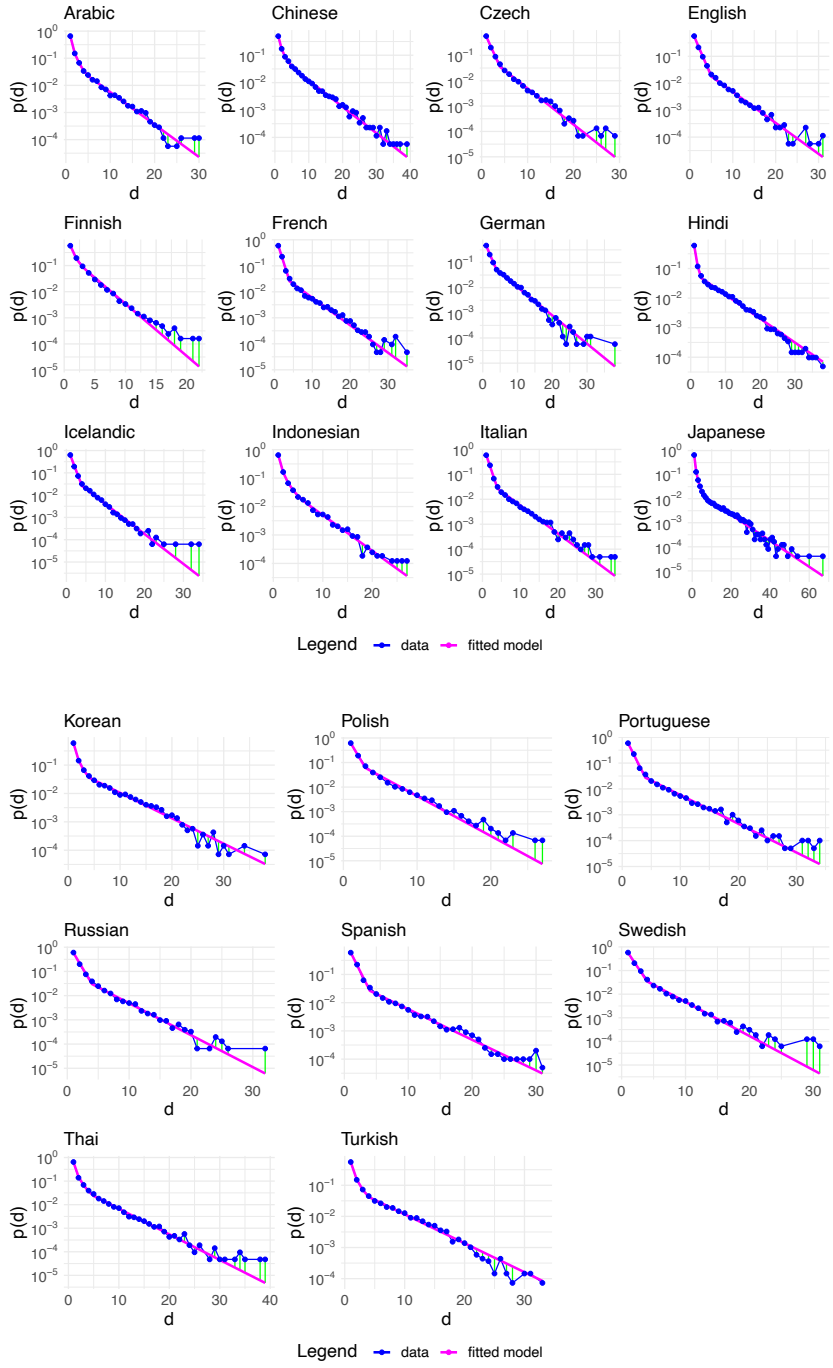
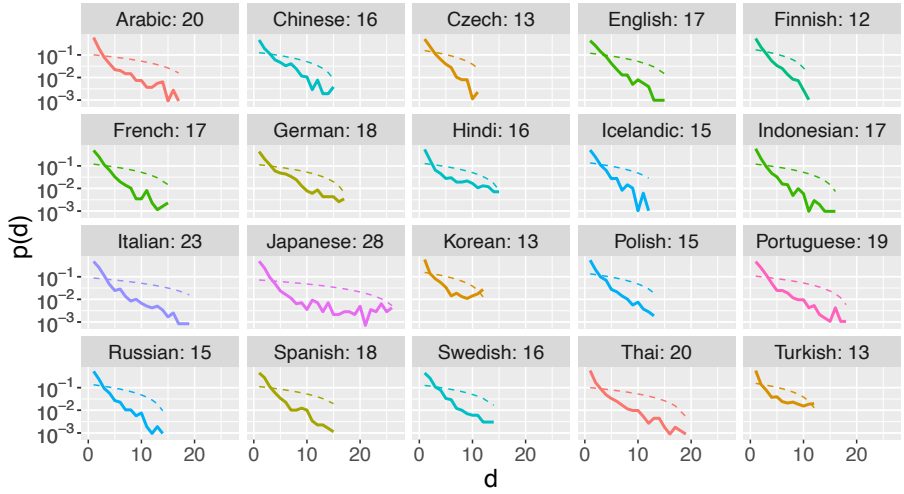


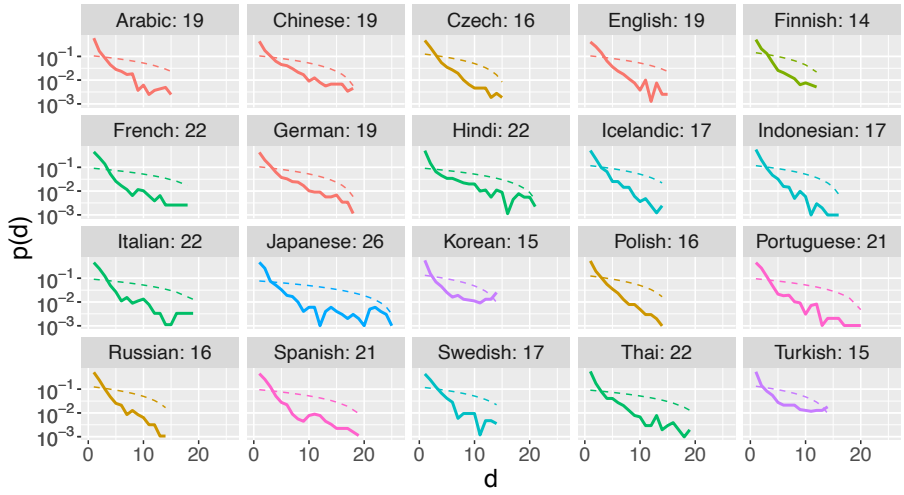
Figure B3.$\boldsymbol{p(d)}$, the probability that a dependency link is formed between words at distance $\boldsymbol{d}$ according to the best model for every language in PSUD.

Appendix C. Gallery

We here display the probability distribution of dependency distances in typical sentences for each language. Figure C1 (a-b) shows the distributions for modal and mean sentence lengths respectively in PUD, while figure C2 (a-b) shows the ones for PSUD.

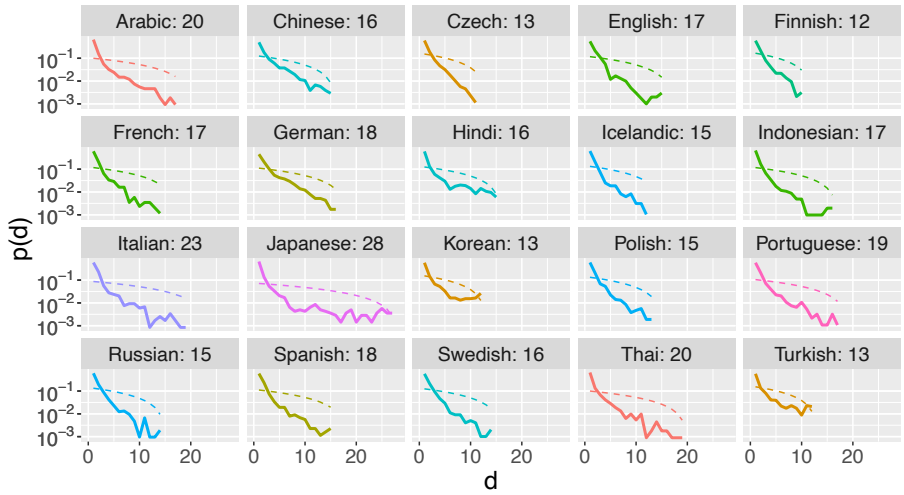


(a) Modal sentence length.

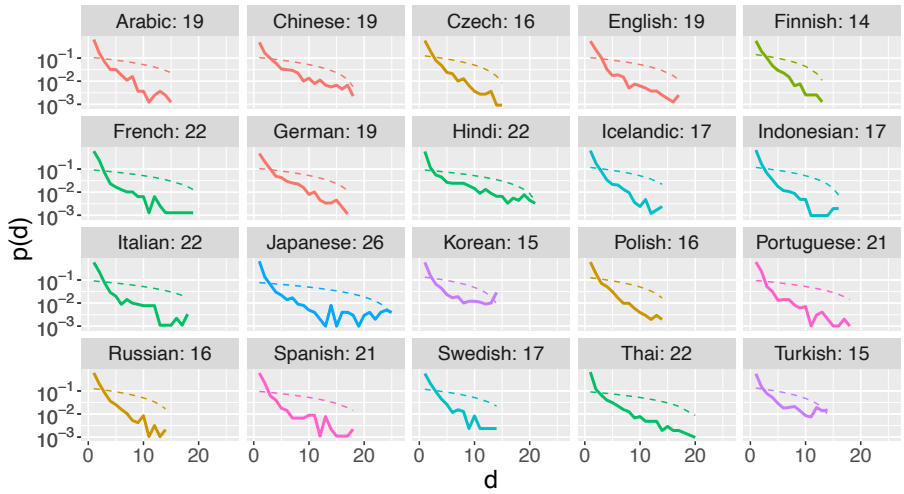


(b) Mean sentence length.

Figure C1. $\boldsymbol{p(d)}$, the probability that linked words are at distance $\boldsymbol{d}$ in sentences of modal (a) and mean (b) length for each language of PUD. Modal and mean sentence values are shown next to the respective language label. For each language, the dashed line shows the probability in shuffled sentences (equation (1)). Points where $\boldsymbol{p(d) = 0}$ are not shown.



(a) Modal sentence length.



(b) Mean sentence length.

Figure C2. $\boldsymbol{p(d)}$, the probability that linked words are at distance $\boldsymbol{d}$ in sentences of modal (a) and mean (b) length for each language of PSUD.Modal and mean sentence values are shown next to the respective language label. For each language, the dashed line shows probability in shuffled sentences (equation (1)). Points where $\boldsymbol{p(d) = 0}$ are not shown.