Dependency distance minimization predicts compression

Ramon Ferrer-i-Cancho
Complexity and Quantitative Linguistics Lab,
LARCA Research Group
Departament de Ciències de la Computació
Universitat Politècnica de Catalunya
Campus Nord, Edifici Omega
Jordi Girona Salgado 1-3
08034 Barcelona, Catalonia, Spain
rferrericancho@cs.upc.edu

Carlos Gómez-Rodríguez
Universidade da Coruña, CITIC
FASTPARSE Lab, LyS Research Group
Departamento de Ciencias
de la Computación y Tecnologías
de la Información
Facultade de Informática, Elviña,
15071, A Coruña, Spain
cgomezr@udc.es

Abstract

Dependency distance minimization (DDm) is a well-established principle of word order. It has been predicted theoretically that DDm implies compression, namely the minimization of word lengths. This is a second order prediction because it links a principle with another principle, rather than a principle and a manifestation as in a first order prediction. Here we test that second order prediction with a parallel collection of treebanks controlling for annotation style with Universal Dependencies and Surface-Syntactic Universal Dependencies. To test it, we use a recently introduced score that has many mathematical and statistical advantages with respect to the widely used sum of dependency distances. We find that the prediction is confirmed by the new score when word lengths are measured in phonemes, independently of the annotation style, but not when word lengths are measured in syllables. In contrast, one of the most widely used scores, i.e. the sum of dependency distances, fails to confirm that prediction, showing the weakness of raw dependency distances for research on word order. Finally, our findings expand the theory of natural communication by linking two distinct levels of organization, namely syntax (word order) and word internal structure.

1 Introduction

According to the dependency distance minimization (DDm) principle, the distance between heads and their dependent words in a sentence has to be reduced (Ferrer-i-Cancho, 2004; Gildea and Temperley, 2007). The principle has been supported widely by studies whose coverage of languages and families is increasing over time (e.g., Liu (2008), Futrell et al. (2015), Futrell et al. (2020), Ferrer-i-Cancho et al. (2021)). For simplicity, such distance is usually measured in words (see Ferrer-i-Cancho (2015b) for an exception) but it could be measured with more precision in syllables or phonemes (Ferrer-i-Cancho, 2015b). In that way, the distance of a dependency would be a function of the length of the words defining the dependency and that of the words in-between. For this reason, it was predicted that word lengths should be minimized to minimize dependency distances (Ferrer-i-Cancho, 2017a; Ferrer-i-Cancho, 2017b), namely the DDm principle predicts compression, i.e., another principle whereby L, the mean word length, has to be minimized. However, to our knowledge, such a prediction has never been tested in spite of its great theoretical importance. First, it is crucial for the construction of a theory of language and other natural communication systems (Semple et al., 2021). A critical component of a theory are the predictions that it can make. For instance, DDm predicts the scarcity of crossing dependencies (Gómez-Rodríguez and Ferrer-i-Cancho, 2017) and compression of word lengths predicts Zipf's law of abbreviation (Ferrer-i-Cancho et al., 2019). These are examples of first order predictions, namely manifestations that are predicted by a certain principle. The focus of this article are second order predictions, namely, principles that are predicted by other principles, such as (a) the prediction of DDm from Dm, a general principle of distance minimization (Ferrer-i-Cancho et al., 2021; Ferrer-i-Cancho, 2003) in the spirit of Behaghel's pioneering views (Behaghel, 1932), or (b) a principle of surprisal or entropy minimization (Hm) from compression (Ferrer-i-Cancho, 2018). Table 1 summarizes these and other first order and second order predictions from previous research stemming from a general principle of energy minimization (Em) in the spirit of Zipf's least effort principle (Zipf, 1949) but extended to other biological systems (Semple et al., 2021).

In the quantitative linguistics tradition, syllables are considered to be one of the best units (if not the best one) for measuring word length, mainly to warrant cross-linguistic validity (Popescu et al., 2013; Grzybek, 2013). In addition, phonemes (and graphemes) are considered to not be appropriate because they are not immediate constituents of words (words are made of syllables that are made in turn of phonemes). Here we will revise this view arising from research on word lengths and check if it still applies for research on the interaction between dependency distance (in words) and word length.

This article is aimed at testing the hypothesis that DDm implies compression, which we refer to as repertoire unit weight minimization (RUWm). We follow the convention that a lower case m at the end of the abbreviation of a principle indicates "minimization" (Ferrer-i-Cancho and Gómez-Rodríguez, 2021). RUWm is the minimization of the weight (or cost) of units in a repertoire. In the context of word types and their length as their weight, RUWm corresponds to the minimization of word lengths. RUWm, predicts that more likely units in a repertoire (e.g., more frequent words in a vocabulary) should be lighter (e.g., be shorter). In particular, RUWm has been argued to lead to the emergence of the law of abbreviation across linguistic levels (Ferrer-i-Cancho et al., 2019): e.g., the lexical level where frequent word types or meanings tend to have shorter forms (Zipf, 1949; Kanwal et al., 2017; Brochhagen, 2021), and the sublexical level where more frequent cases or case marker types tend to have shorter forms (Liu, 2021). RUWm is a specialization of a more general principle of unit weight minimization (UWm) arising from research on unifying compression with the origins of both the law of abbreviation and Menzerath's law (Gustison et al., 2016). The law of abbreviation is the tendency of more frequent words to be shorter (Zipf, 1949) while Menzerath's law is the tendency of linguistic constructs with more parts to be made of smaller parts (Altmann, 1980). A specialization of UWm on sequences, sequence unit weight minimization (SUWm) sheds light on the possible origin of Menzerath's law (Gustison et al. (2016); Table 1). In the context of words as sequences of syllables and the length of as syllable as its weight (or cost), SUWm corresponds to the minimization of the lengths of syllables in the words they appear. SUWm predicts that units in longer sequences (e.g., syllables in words with more syllables) should have smaller weight (e.g., be shorter).

Formally, we aim to test

$$DDm \longrightarrow RUWm. \tag{1}$$

As an alternative to that prediction one could consider a compensation hypothesis, where less optimized languages at the level of dependency distances would have more pressure for shorter words (or more optimized languages at the level of dependency distances would tolerate longer words). The idea of compensation has been applied in word order research in different ways (Ferrer-i-Cancho, 2018; Ferrer-i-Cancho, 2015b). For instance, the suboptimal placement of the verb in SOV orders with respect to DDm (DDm predicts SVO or OVS) has been argued to be compensated by the short length of clitics (Ferrer-i-Cancho, 2015b).

2 Methods

To measure dependency distance in a sentence, we considered two scores, both computed using dependency distances measured in words. The first score is Ω, a recently introduced normalized score that takes the value 1 when the dependency distances are fully optimized and is expected to take a value of 0 if there is no bias on dependency distances, for or against DDm (Ferrer-i-Cancho et al., 2021). Ω can be seen as score of the intensity of DDm, which is maximum when $\Omega = 1$, and missing when $\Omega \approx 0$. $\Omega < 0$ indicates that DDm is surpassed by other word order principles. The score has the virtue of satisfying a series of mathematical and statistical properties: dual normalization (i.e. normalization with respect to both the minimum and the random baseline), constancy under minimum linear arrangement, stability under random linear arrangement, invariance under linear transformation and boundedness under maximum linear arrangement (Ferrer-i-Cancho et al., 2021). These properties are particularly useful when calculating an average score over the sentences of a treebank. For comparison, we consider also D, the

Em	⟶	***Dm***	⟶	***SDm***	⟶	pairs of primarily alternating word orders (Ferrer-i-Cancho, 2016)
			⟶	***DDm***	⟶	acceptability (Morrill, 2000)
					⟶	word order preferences (Morrill, 2000)
					⟶	scarcity of crossing dependencies (Gómez-Rodríguez and Ferrer-i-Cancho, 2017)
					⟶	tendency to uncover the root (Ferrer-i-Cancho, 2008)
					⟶	projectivity & planarity with high probability
				DDm (+ projectivity)	⟶	medial placement of the root (Gildea and Temperley, 2007; Alemany-Puig et al., 2022)
				DDm (+ planarity)	⟶	medial placement of the central vertex (Iordanskii, 1987; Hochberg and Stallmann, 2003; Alemany-Puig et al., 2022)
				DDm + extreme V placement in SVO triples	⟶	placement of adjectives with respect to nominal heads (Ferrer-i-Cancho, 2008; Ferrer-i-Cancho, 2015a)
					⟶	placement of auxiliary V with respect to main V (Ferrer-i-Cancho, 2008)
					⟶	consistent branching for dependents of nominal heads (Ferrer-i-Cancho, 2015b)
					⟶	"unnecessity" of headedness parameter (Ferrer-i-Cancho, 2015b)
					⟶	***RUWm*** (to be tested in this article)
	⟶	***UWm***	⟶	***RUWm***	⟶	Zipf's law of abbreviation (Shannon, 1948; Ferrer-i-Cancho et al., 2019)
						reduction (Ferrer-i-Cancho, 2017a)
				RUWm + unique segmentation	⟶	***Hm*** (Ferrer-i-Cancho, 2018)
				SUWm	⟶	Menzerath's law (Gustison et al., 2016; Ferrer-i-Cancho et al., 2019)

Table 1: Optimization principles and their predictions. Arrows link principles with their predictions. Predictions can take the form of principles or manifestations. Principles are marked in boldface. The two principles whose relationship is the target of this article are marked in blue. *Em*: energy minimization. *Dm*: distance minimization. *UWm*: unit weight minimization, popularly known as *compression*. *RUWm*: repertoire unit weight minimization. *SUWm*: sequence unit weight minimization. *SDm*: swap distance minimization. *DDm*: dependency distance minimization. *Hm*: entropy (or surprisal) minimization. We use parentheses for assumptions that are likely to be predictions of DDm and thus likely to be unnecessary assumptions to a large extent (see Gómez-Rodríguez and Ferrer-i-Cancho (2017) for further details about the argument). Awareness of such unnecessary assumptions is vital for the construction of a parsimonious theory.

sum of dependencies of a sentence. While Ω is a measure of closeness, D is a measure of distance. D is the most widely used score (Gildea and Temperley, 2007; Gildea and Temperley, 2010; Futrell et al., 2015; Futrell et al., 2020) but does not satisfy any of the remarkable mathematical properties enumerated above. The intensity of DDm is difficult to assess just from the value of D.

For each treebank considered in this study, we calculated an average Ω and average D over all the sentences of the treebank. To measure word length, we considered two different units: phonemes and syllables. For each language, we calculated L_s, the mean word length in syllables (mean syllables per word token) and L_p, the mean word length in phonemes (mean phonemes per word token).

To control for the content or the source text of the treebanks, we used a parallel collection of treebanks, in particular, the Parallel Universal Dependencies (PUD) collection version 2.6 (Zeman et al., 2017). PUD contains 20 languages from 9 distinct families. PUD follows the UD annotation style (Zeman et al., 2020). To control for annotation style, we also use SUD, i.e. Surface-Syntactic Universal Dependencies (Gerdes et al., 2018). We use PSUD to refer to the PUD collection following the SUD annotation (Ferrer-i-Cancho et al., 2021). The PUD and PSUD treebank collections are borrowed from a recent study (Ferrer-i-Cancho et al., 2021). The preprocessing of these treebanks involves the removal of punctuation marks and reparalellization to warrant there is no loss of parallelism after punctuation mark removal (Ferrer-i-Cancho et al., 2021). The data are available from `https://github.com/lluisalemanypuig/optimality-syntactic-dependency-distances` in two levels of preprocessing: (a) the preprocessed treebanks as head vectors and (b) the raw text tables that were extracted from them and used to feed the statistical analyses. The transformation of the raw head vectors into the raw text tables can be replicated easily with the Linear Arrangement Library (Alemany-Puig and Ferrer-i-Cancho, 2022).

The PUD data does not include syllable or phoneme annotations that would allow us to measure word lengths in these units, and obtaining them would be highly costly. Thus, we instead borrowed mean word lengths from the dataset of another recent study (Fenk-Oczlon and Pilz, 2021). In that study, mean word lengths were estimated from 22 simple declarative sentences encoding one proposition and using basic vocabulary. Three languages from PUD/PSUD are missing in that dataset: Arabic (Afro-Asiatic), Indonesian (Austronesian) and Swedish (Indo-European). As a result, the final collection has 17 languages from 7 distinct families that are displayed in Table 2.

We consider two approaches to investigate the relationship between dependency distance and word length. First, a Kendall τ correlation test between the mean dependency distance score (D or Ω) and mean word length (L_s or L_p). With respect to plain Pearson correlation, τ is more robust to extreme observations and to non-linearity (Newson, 2002). A significant negative correlation between mean Ω and L_s or L_p would confirm the prediction in Eq. 1. Note that Ω and D have opposite interpretations (larger values of Ω imply shorter dependencies whereas larger values of D imply longer dependencies) hence a positive correlation with respect to D would be analogous to a negative correlation with respect to Ω. Therefore, a positive correlation between D and L_s or L_p could also be interpreted as confirming the prediction in Eq. 1 but D lacks the mathematical and statistical properties that are required for robust assessment (Ferrer-i-Cancho et al., 2021).

Second, the fact that the Indo-European family is over-represented and the only one that is represented by more than one language (Table 2) motivates the need to control for the effect of family size. Accordingly, we also consider a couple of kinds of generalized linear models. First, a null model with mean word length as response (L_s or L_p) and language family as random factor. Second, a mixed effects model with mean word length as response (L_s or L_p), language family as random effect and a dependency distance score (mean D or mean Ω) as fixed effect. To test the prediction with these models, we use information theoretic model selection (Burnham and Anderson, 2002; Winter, 2019). AIC of the mixed effects model being lower than that of the null model would confirm the prediction provided that the weight of the association between the predictor and the response has a sign that matches that of the prediction. Again, some caution is needed when D is involved because of its technical limitations (Ferrer-i-Cancho et al., 2021).

The linear models were fitted with the `lme4` R package. Confidence intervals for the weight of the

Family	Languages
Turkic (1)	Turkish
Indo-European (11)	Czech, English, French, German, Hindi, Icelandic, Italian, Polish, Portuguese, Russian, Spanish
Japonic (1)	Japanese
Koreanic (1)	Korean
Sino-Tibetan (1)	Chinese
Tai-Kadai (1)	Thai
Uralic (1)	Finnish

Table 2: The 17 languages from 7 families used in the present study. The counts attached to each family name indicate the number of different languages included in the present study.

Collection	distance	length	n	τ	p
PUD	Ω	L_s	17	-0.052	0.773
		L_p	17	-0.37	0.039
PSUD	Ω	L_s	17	-0.111	0.536
		L_p	17	-0.37	0.039
PUD	D	L_s	17	-0.258	0.149
		L_p	17	-0.459	0.011
PSUD	D	L_s	17	-0.185	0.303
		L_p	17	-0.385	0.032

Table 3: The correlation between the mean dependency distance score and mean word length. We show the annotation style, the distance score, the word length score, the value of the Kendall τ correlation statistic and p, the p-value of the corresponding two-sided test. We assume a significance level of 0.05.

fixed effect were computed using the parametric bootstrapping method of function `confint` of the `lme4` package.

3 Results

Fig. 1 shows the relationship between word length and Ω. Table 3 indicates that the correlation between Ω and L_p is negative and statistically significant whereas the correlation between Ω and L_s is also negative but not significant. That is, the shorter the syntactic dependencies of a language upon dual normalization (Ω), the shorter the words when their length is measured in phonemes. Table 4 indicates that the result is confirmed by a linear mixed effects model that predicts L_s or L_p based on Ω with family as random effect. The null model (only family as random factor) always yields an AIC value that is larger

Collection	distance	length	AIC mixed effects	AIC null
PUD	Ω	L_s	23.15	23.9
		L_p	42.16	47.64
PSUD	Ω	L_s	23.54	23.9
		L_p	42.96	47.64
PUD	D	L_s	32.28	23.9
		L_p	45.23	47.64
PSUD	D	L_s	32.22	23.9
		L_p	48.03	47.64

Table 4: Information theoretic selection of models to predict the mean word length. We show the annotation style, the distance score, the word length score, the Akaike Information Criterion (AIC) of the mixed effects model and the AIC of the null model.

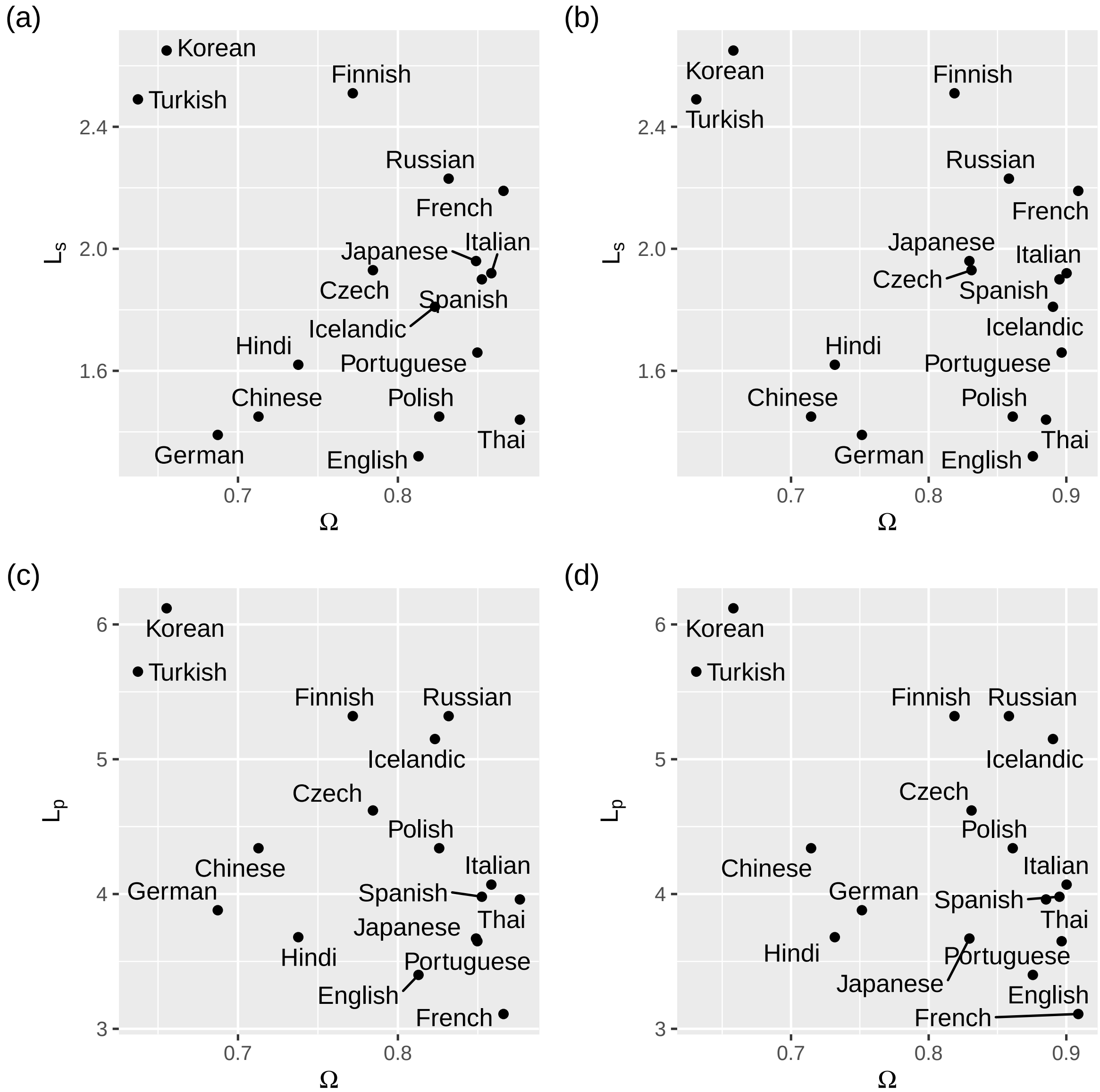

Figure 1: Mean word length (L) as a function of the mean degree of optimality of syntactic dependency distances (mean Ω). (a) L_s as a function of mean Ω in PUD. (b) L_s as a function of mean Ω in PSUD. (c) L_p as a function of mean Ω in PUD. (d) L_p as a function of mean Ω in PSUD.

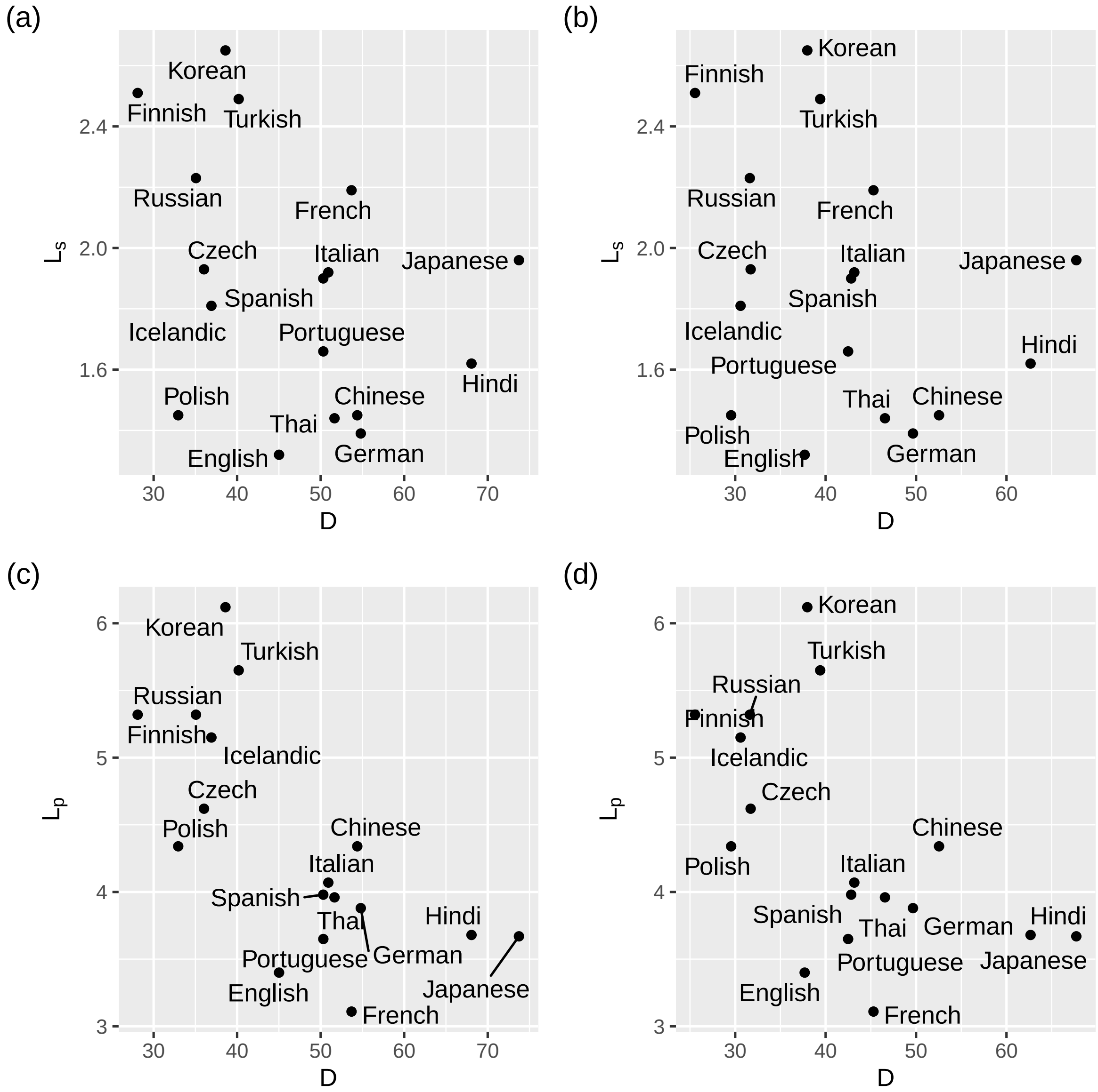

Figure 2: Mean word length (L) as a function of the mean sum of dependency distances (mean D). (a) L_s as a function of D in PUD. (b) L_s as a function of D in PSUD. (c) L_p as a function of mean D in PUD. (d) L_p as a function of mean D in PSUD.

Collection	Fixed effect	Response	Lower	Upper
PUD	Ω	L_p	-10.85	-1.19
PSUD	Ω	L_p	-9.21	-0.69
PUD	D	L_s	-0.03	0
		L_p	-0.07	-0.02
PSUD	D	L_s	-0.03	0.01
		L_p	-0.07	-0.02

Table 5: Lower and upper bounds of the 95% confidence interval for the weight of the fixed effect (Ω or D) in the mixed effects models. The confidence intervals for Ω with L_p as response could not be computed due to numerical problems.

than that of the mixed effects model (Ω as fixed effect and family as random effect). However, the AIC of the null model is only sufficiently large (i.e., differing by various units) when the predictor is L_s, in agreement with the plain correlation analysis (Table 3). The analysis of the confidence intervals supports the conclusions: the confidence interval for the weight of Ω with L_p as response comprises exclusively negative values, consistently with a negative correlation between Ω and L_p (Table 5).

If Ω is replaced by D in the preceding analyses, opposite conclusions are drawn concerning the direction of the correlation. First, Figure 2 suggests that mean word length tends to decrease as D increases when word length is measured in phonemes in both PUD and PSUD. Accordingly, Table 3 confirms that the correlation between D and L_p is negative and statistically significant while the correlation between D and L_s turns out to not be significant. That is, the longer the syntactic dependencies of a language, the shorter the words when their length is measured in phonemes, the opposite conclusion that is reached with Ω. Table 4 indicates that the results are confirmed by linear mixed effects linear models that predict L_s or L_p based on D with family as random effect. When measuring word lengths in syllables, the AIC of the null model is much smaller than that of the mixed effects model with D as predictor, supporting that D and L_s are uncorrelated. In contrast, the AIC of the null model is larger (by more than unit) than that of the mixed effects model with D as predictor, supporting that D and L_p are correlated. The analysis of the confidence intervals (Table 5) confirms a negative correlation between D and L_p because the confidence intervals for the weight of D with L_p as response comprise exclusively negative values; that does not happen when L_s is the response.

4 Discussion

We have confirmed the prediction that DDm leads to compression (Eq. 1) by showing that Ω and L_p are negatively correlated: the closer the syntactically related words, the shorter the words. Although syllables are the preferred unit of measurement of word length over phonemes in quantitative linguistics (Popescu et al., 2013; Grzybek, 2013), we have failed to confirm the prediction with that unit. We cannot exclude the possibility that a negative correlation between Ω and L_s exists but has not surfaced because it is weaker and the sample of languages is not large enough. However, the failure with syllables may be a confirmation of the higher capacity of phonemes over syllables to capture the so-called 'phonological complexity' (Pimentel et al., 2020). Furthermore, we suspect that word lengths in phonemes lead eventually to a more accurate estimation of the true distance between words, currently measured in words, than syllables. To see the relationship between word length and dependency distance and to build an explanation, suppose that δ_p is the phonemic distance between two words, δ_s is their syllabic distance and d is their distance in words ($d = 1$ if two words are consecutive). Then assuming that a word and its dependent word are connected by the middle of their phonemic or syllabic sequence, a mean field approximation yields $\delta_p \approx dL_p$ (the head and their dependent contribute with $L_p/2$ phonemes; the words in-between the head and the governor contribute with $(d-1)L_p$ phonemes). Analogously, $\delta_s \approx dL_s$. Then δ_p may be more strongly correlated with the time that is needed to keep unresolved dependencies in memory (Morrill, 2000) than δ_s because syllables vary in phonemic length. That time was considered to be the key to understand what determines the acceptability of sentences and other word order phenomena

(Morrill, 2000). Then δ_p looks like a better proxy for the combination of memory decay and interference that is believed to cause DDm (Liu et al., 2017; Temperley and Gildea, 2018). To recap, the argument that syllables are more appropriate units than phonemes (or graphemes) because syllables are immediate word constituents (Popescu et al., 2013; Grzybek, 2013) may not be appropriate for research on dependency distances because distances (or time) could be measured more accurately with constituents located farther down in the hierarchy of constituents.

We have also shown that D, a widely used dependency score (Gildea and Temperley, 2007; Gildea and Temperley, 2010; Futrell et al., 2015; Futrell et al., 2020), fails to confirm the prediction that DDm implies compression. It is not the first time that Ω shows a superior performance when testing theoretical predictions. When testing that DDm should be surpassed by other word order principles in short sentences (Ferrer-i-Cancho, 2014; Ferrer-i-Cancho, 2017a), Ω was able to find many more languages with anti-DDm effects than D (Ferrer-i-Cancho et al., 2021). Our findings reinforce the view that raw dependency distances are a poor reflect of DDm and that advanced scores such as Ω are crucial for progress in research on that optimization principle and related memory constraints (Ferrer-i-Cancho et al., 2021), and eventually, for the construction of a general theory of natural systems with human language as a particular case and energy minimization at the center (Semple et al., 2021). An early sketch of that theory is shown in Table 1.

For each language, we have measured dependency distances (Ω and D) on the collection of sentences included in PUD/PSUD whereas mean word lengths (L_p and L_s) come from an independent collection of sentences (Fenk-Oczlon and Pilz, 2021). It could be argued that mean word lengths should have been estimated on PUD/PSUD, too. This would make the whole study fully parallel, and word length data potentially more accurate, as it would come from a larger sample and, more specifically, from the same sentences where dependency lengths have been measured. It is conceivable, for example, that a given kind of syntactic construction could produce shorter dependency distances in language A than in language B, while being prevalent in PUD/PSUD but not in the word length dataset (e.g. due to genre, style or topic differences). This could cause us to observe increased dependency distance minimization in language A with respect to B without being able to observe the associated compression, as our word length data would not include sentences with that specific phenomenon. In turn, this could cause us to underestimate the correlation between Ω and L_s or L_p. Unfortunately, the technical complexity of replicating the present study with actual phonemic or syllabic lengths on PUD/PSUD goes beyond the scope of the present article. These considerations notwithstanding, notice that the stress of this article is on testing a prediction rather than theoretically agnostic data description. In this context, it is rather astonishing that the theoretical prediction (Eq. 1) is confirmed even though the collections of sentences for dependency distances and the collections for word lengths are independent. That confirmation offers two major interpretations: (a) the results are due to biases in the sentences, either in PUD or in Fenk-Oczlon's dataset (Fenk-Oczlon and Pilz, 2021) or, crucially, (b) there is actually a deep reason for the prediction to hold. Further research on a fully parallel scenario or alternative sources is required.

By having shown that DDm implies compression of word lengths, we do not mean that compression is produced exclusively by DDm. Our work does not rule out other sources for compression. Following previous research testing successfully the prediction that DDm weakens in small sequences (Ferrer-i-Cancho and Gómez-Rodríguez, 2021; Ferrer-i-Cancho et al., 2021), we formulate a new prediction, namely that compression *per se* (independently from DDm) may surface in short sequences. Future research should clarify the weight of the contribution to compression from DDm, compression itself and other principles.

Once one integrates our findings into the piece of a mathematical theory of communication in Table 1, it turns out that we have actually uncovered the following chain,

$$Dm \longrightarrow DDm \longrightarrow RUWm \longrightarrow \text{ Zipf's law of abbreviation}$$

Namely, a general principle of minimization of the distance between elements leads to the prediction of the law of abbreviation.

Our findings have implications for competing views and frameworks. First, the compensation hypothesis outlined in the introduction, i.e. less optimized languages at the level of dependency distances would have more pressure for shorter words (or more optimized languages at the level of dependency distances would tolerate longer words), would predict that the correlation between Ω and word length should not be negative. Instead, a zero or positive correlation would be expected depending on the strength or the nature of the compensation effect. Our findings of a negative correlation rule out the compensation hypothesis as a primary explanation for the global trend. However, we cannot exclude that compensation has some secondary role in general, an important role in specific languages or an important role with certain domains of a language, as suggested for the suboptimal placement of clitics with respect to DDm in Romance languages (Ferrer-i-Cancho, 2015b). Second, DDm predicts compression and compression in turn predicts reduction (see Section 3.4 of Ferrer-i-Cancho (2017a)). By reduction, here we mean the shortening or omission of predictable utterances, a phenomenon that has been used to justify the uniform information density and related hypotheses (see Lemke et al. (2021) and references therein). Therefore, the need for uniform information density and related hypotheses as standalone hypotheses needs to be revised. Third, our finding of an association between word length and dependency distance also challenges the recent conclusion that *(compositional) morphology and graphotactics can sufficiently account for most of the complexity of natural codes – as measured by code length* (Pimentel et al., 2021) as if no strong independent pressure for compression (beyond morphology and graphotactics) really existed. Our findings suggest that the two components of the problem of compression as defined in standard information theory, i.e. the minimization of word length and the conditions of the coding scheme (Ferrer-i-Cancho et al., 2019), may not be as easy to dissociate from (compositional) morphology and graphotactics as expected from the traditional reductionistic division into linguistic levels. Indeed, our article demonstrates how constraints on the "syntactic level" (DDm) may be shaping the "(sub)lexical level" (compression on word lengths).

Our work has implications beyond the current state of development of the theory of the communication efficiency reviewed in Table 1. We have confirmed that DDm implies compression but the outcome of our correlation analysis does not exclude that it could also the other way around, namely that compression is actually leading to DDm. A possible track could be that pressure for shorter words may lead to a loss of information that would require more words to convey the same message, which in turn would imply longer sentences and then higher pressure for DDm. Another track could be that a general principle of compression operates on top both at the level of words and at higher levels (phrases, clauses, sentences) and then close packaging (DDm) at all these levels is simply a consequence of maximizing compression, in line with the now-or-never bottleneck (Christiansen and Chater, 2016). We hope that our work stimulates further research on general principles and their predictions.

Acknowledgements

We are grateful to S. Semple and M. Gustison for extremely valuable and inspiring discussions on theory construction (Semple et al., 2021). We are also grateful to two anonymous reviewers, T. Pimentel and G. Fenk-Oczlon for helpful comments. RFC is supported by the grant TIN2017-89244-R from MINECO (Ministerio de Economía, Industria y Competitividad) and the recognition 2017SGR-856 (MACDA) from AGAUR (Generalitat de Catalunya). CGR is supported by the European Research Council (ERC), under the European Union's Horizon 2020 research and innovation programme (FASTPARSE, grant agreement No 714150), the ANSWER-ASAP (TIN2017-85160-C2-1-R) and SCANNER-UDC (PID2020-113230RB-C21) projects from ERDF/MICINN/AEI, Xunta de Galicia (ED431C 2020/11 and an Oportunius program grant to complement ERC grants); the CITIC research center is funded by ERDF and Xunta de Galicia (ERDF - Galicia 2014-2020 program, grant ED431G 2019/01).

References

Lluís Alemany-Puig and Ramon Ferrer-i-Cancho. 2022. The Linear Arrangement Library. A new tool for research on syntactic dependency structures. In *Proceedings of the SyntaxFest 2022*, Sofia, Bulgaria. Association for Computational Linguistics.

Lluís Alemany-Puig, Juan Luis Esteban, and Ramon Ferrer-i-Cancho. 2022. Minimum projective linearizations of trees in linear time. *Information Processing Letters*, 174:106204.

Gabriel Altmann. 1980. Prolegomena to Menzerath's law. *Glottometrika*, 2:1–10.

Otto Behaghel, 1932. *Deutsche Syntax: eine geschichtliche Darstellung*, chapter Bd 4: Wortstellung. Periodenbau. Carl Winters Univeritätsbuchhandlung, Heidelberg, Germany.

Thomas Brochhagen. 2021. Brief at the risk of being misunderstood: Consolidating population-and individual-level tendencies. *Computational Brain & Behavior*, Feb.

Kenneth P. Burnham and David R. Anderson. 2002. *Model selection and multimodel inference. A practical information-theoretic approach*. Springer, New York, 2nd edition.

Morten H. Christiansen and Nick Chater. 2016. The now-or-never bottleneck: a fundamental constraint on language. *Behavioral and Brain Sciences*, 39:1 – 72.

Gertraud Fenk-Oczlon and Jürgen Pilz. 2021. Linguistic complexity: Relationships between phoneme inventory size, syllable complexity, word and clause length, and population size. *Frontiers in Communication*, 6:66.

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

Ramon Ferrer-i-Cancho, Christian Bentz, and Caio Seguin. 2019. Optimal coding and the origins of Zipfian laws. *Journal of Quantitative Linguistics*, page in press.

Ramon Ferrer-i-Cancho, Carlos Gómez-Rodríguez, Juan Luis Esteban, and Lluís Alemany-Puig. 2021. The optimality of syntactic dependency distances. *Physical Review E*, page in press.

Ramon Ferrer-i-Cancho. 2003. *Language: universals, principles and origins*. Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona.

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

Ramon Ferrer-i-Cancho. 2008. Some word order biases from limited brain resources. A mathematical approach. *Advances in Complex Systems*, 11(3):393–414.

Ramon Ferrer-i-Cancho. 2014. Why might SOV be initially preferred and then lost or recovered? A theoretical framework. In E. A. Cartmill, S. Roberts, H. Lyn, and H. Cornish, editors, *The Evolution of Language - Proceedings of the 10th International Conference (EVOLANG10)*, pages 66–73, Vienna, Austria. Wiley. Evolution of Language Conference (Evolang 2014), April 14-17.

Ramon Ferrer-i-Cancho. 2015a. The placement of the head that minimizes online memory. A complex systems approach. *Language Dynamics and Change*, 5(1):114–137.

Ramon Ferrer-i-Cancho. 2015b. Reply to the commentary "Be careful when assuming the obvious", by P. Alday. *Language Dynamics and Change*, 5(1):147–155.

Ramon Ferrer-i-Cancho. 2016. Kauffman's adjacent possible in word order evolution. In *The evolution of language: proceedings of the 11th International Conference (EVOLANG11)*.

Ramon Ferrer-i-Cancho. 2017a. The placement of the head that maximizes predictability. An information theoretic approach. *Glottometrics*, 39:38–71.

Ramon Ferrer-i-Cancho. 2017b. Towards a theory of word order. comment on "dependency distance: A new perspective on syntactic patterns in natural language" by Haitao Liu et al. *Physics of Life Reviews*, 21:218 – 220.

Ramon Ferrer-i-Cancho. 2018. Optimization models of natural communication. *Journal of Quantitative Linguistics*, 25(3):207–237.

Richar Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Richard Futrell, Roger P. Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–412.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 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, November. Association for Computational Linguistics.

Daniel Gildea and David Temperley. 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, June. Association for Computational Linguistics.

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.

Carlos Gómez-Rodríguez and Ramon Ferrer-i-Cancho. 2017. Scarcity of crossing dependencies: A direct outcome of a specific constraint? *Physical Review E*, 96:062304.

Peter Grzybek. 2013. Word length. In John R. Taylor, editor, *The Oxford Handbook of the Word*. Oxford University Press.

Morgan L. Gustison, Stuart Semple, Rarmon Ferrer-i-Cancho, and Thore Bergman. 2016. Gelada vocal sequences follow Menzerath's linguistic law. *Proceedings of the National Academy of Sciences USA*, 13:E2750–E2758.

Robert A. Hochberg and Matthias F. Stallmann. 2003. Optimal one-page tree embeddings in linear time. *Information Processing Letters*, 87:59–66.

Mikhail A. Iordanskii. 1987. Minimal numberings of the vertices of trees — Approximate approach. In Lothar Budach, Rais Gatič Bukharajev, and Oleg Borisovič Lupanov, editors, *Fundamentals of Computation Theory*, pages 214–217, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jasmeen Kanwal, Kenny Smith, Jennifer Culbertson, and Simon Kirby. 2017. Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, 165:45–52.

Robin Lemke, Lisa Schäfer, and Ingo Reich. 2021. Modeling the predictive potential of extralinguistic context with script knowledge: The case of fragments. *PLoS One*, 16:e0246255.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9:159–191.

Zoey Liu. 2021. A multifactorial approach to crosslinguistic constituent orderings. *Linguistics Vanguard*, page in press.

Glyn Morrill. 2000. Incremental processing and acceptability. *Computational Linguistics*, 25(3):319–338.

Roger Newson. 2002. Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *Stata Journal*, 2(1):45–64.

Tiago Pimentel, Brian Roark, and Ryan Cotterell. 2020. Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics*, 8:1–18, 01.

Tiago Pimentel, Irene Nikkarinen, Kyle Mahowald, Ryan Cotterell, and Damián Blasi. 2021. How (non-)optimal is the lexicon? In *North American Chapter of the Association for Computational Linguistics*.

Ioan-Iovitz Popescu, Sven Naumann, Emmerich Kelih, Andrij Rovenchak, Haruko Sanada, Anja Overbeck, Reginald Smith, Radek Cech, Panchanan Mohanty, Andrew Wilson, and Gabriel Altmann. 2013. Word length: Aspects and languages. In *Studies in Quantitative Linguistics*, volume 3, pages 224–281. RAM-Verlag, Lüdenscheid.

Stuart Semple, Ramon Ferrer-i-Cancho, and Morgan Gustison. 2021. Linguistic laws in biology. *Trends in Ecology and Evolution*, page in press.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423 623–656.

David Temperley and Daniel Gildea. 2018. Minimizing syntactic dependency lengths: Typological/Cognitive universal? *Annual Review of Linguistics*, 4(1):67–80.

Bodo Winter. 2019. *Statistics for linguists: an introdution using R.* Routledge, New York & London.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, et al. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada, August. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Lene Antonsen, et al. 2020. Universal dependencies 2.6. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

George Kingsley Zipf. 1949. *Human behaviour and the principle of least effort*. Addison-Wesley, Cambridge (MA), USA.