Natalia Levshina* and Steven Moran

# Efficiency in human languages: Corpus evidence for universal principles

https://doi.org/10.1515/lingvan-2020-0081

**Abstract:** Over the last few years, there has been a growing interest in communicative efficiency. It has been argued that language users act efficiently, saving effort for processing and articulation, and that language structure and use reflect this tendency. The emergence of new corpus data has brought to life numerous studies on efficient language use in the lexicon, in morphosyntax, and in discourse and phonology in different languages. In this introductory paper, we discuss communicative efficiency in human languages, focusing on evidence of efficient language use found in multilingual corpora. The evidence suggests that efficiency is a universal feature of human language. We provide an overview of different manifestations of efficiency on different levels of language structure, and we discuss the major questions and findings so far, some of which are addressed for the first time in the contributions in this special collection.

**Keywords:** efficiency, corpora, typology, information theory, language universals

## 1 Aims and background

According to a general functionalist view, languages are efficient in meeting the communicative needs of their users (e.g., Du Bois 1985; MacWhinney et al. 1984; Zipf 1949). The goal of this special collection is to demonstrate how different manifestations of efficiency in the lexicon, grammar and discourse can be tested and explained with novel methods born out of the increasing availability of corpus data from typologically diverse languages. The contributions in this collection focus on different cross-linguistic manifestations of efficiency, written by experts in corpus linguistics, information theory, cognitive science, psycholinguistics, discourse analysis and typology. These papers bring together new insights on efficiency as a universal principle of language.

Corpus evidence has played a major role in the study of linguistic efficiency. Corpora of different languages have been used to detect efficient patterns in numerous areas of grammar, including in the lexicon (e.g., Bentz 2018; Piantadosi et al. 2011; Zipf [1935]1965); in syntax (e.g., Futrell et al. 2015; Hawkins 1994: Ch. 4); and in phonology (e.g., Cohen Priva and Jaeger 2018; Coupé et al. 2019; Fenk-Oczlon and Fenk 2010; Pellegrino et al. 2011). The use of corpora gives researchers two main advantages in comparison with other methods. First, with the emergence of new multilingual collections of comparable and uniformly annotated corpora, we can test if a particular efficient pattern occurs in many diverse languages, which is much more challenging for the experimental approach. Second, many corpora contain language produced in naturalistic settings, in which the communicative pressures leading to the emergence of efficient patterns come into play. As new and diverse corpora with detailed linguistic information are increasingly available, we have more opportunities for studying different types of efficiency cross-linguistically, as demonstrated by the contributions to this collection. With the help of large-scale corpus data, we can disentangle different factors that have been claimed to explain language users' preferences, to test the links between linguistic and extralinguistic variables, and to fine-tune previous theoretical ideas by making them quantifiable and testable.

***Corresponding author: Natalia Levshina**, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, E-mail: natalia.levshina@mpi.nl. https://orcid.org/0000-0003-4224-2959
**Steven Moran,** Language and Space Lab, University of Zurich, Zurich, Switzerland, E-mail: steven.moran@uzh.ch. https://orcid.org/0000-0002-3969-6549

This article has the following structure. First, we provide an overview of attested manifestations of efficiency in language (Section 2). Then we focus on the contribution of corpus linguists to this area (Section 3). Finally, we formulate a number of big theoretical and methodological questions that need to be answered in the field and give an overview of how the contributions in this special collection help answer them (Section 4).

## 2 Efficiency: Main principles, predictions and manifestations

Various linguistic phenomena related to efficiency have been explained in functional linguistics by the principle of economy (Haiman 1983), the Principle of Least Effort (Zipf 1949), Maxims of Quantity (Grice 1975), the form-frequency correspondence universal (Haspelmath in press), and similar principles, but it is only recently that these phenomena have been systematized under one theoretical umbrella (Gibson et al. 2019; Hawkins 2004, 2014; Levshina 2018). And although there remain many theoretical and methodological questions and challenges (see below), we can offer the following working definition of efficiency:

(1) *A speaker/signer uses language efficiently if he or she spends not more effort than necessary in order to convey intended information, while at the same time maximizing processing ease for the recipient(s).*

Using concepts from economics, efficiency can also be defined as the maximization of the benefit-to-cost ratio. Although communication can bring diverse benefits (social status, aesthetic pleasure, etc.), the main focus of most current studies on efficiency is on the successful transfer of information through a noisy communication channel, which is studied from an information theoretic perspective (Shannon 1948; see the recent review in Gibson et al. 2019). The costs are related to articulation and diverse cognitive operations, such as extraction of lexemes from long-term memory and processing of long syntactic dependencies. Assuming that the language users' behavior is efficient, one can formulate at least three falsifiable predictions:

*Prediction 1.* A communication system should exhibit a positive correlation between costs and benefits. For example, we do not expect a language to exist in which highly informative units are systematically easier to produce than less informative units.

*Prediction 2.* We also expect to find a negative correlation (or in other words, a trade-off) between the costs of communication for different agents (e.g., the speaker/signer and the recipient) or within different subsystems of one communication system (e.g., overt morphological markers vs. word order), provided the benefits (i.e., amount of information) are kept constant. These expectations are similar to the state of Pareto efficiency in economics,[1] in which resources cannot be reallocated to make one individual better off without making at least one individual worse off.

*Prediction 3.* Language users will tend to keep the benefit-to-cost ratio approximately constant per time unit. If the costs are distributed evenly in time, the benefits (transferred information) are also constant. This can explain Fenk and Fenk's (1980) hypothesis of constant information flow, Levy and Jaeger's (2007) Uniform Information Density hypothesis, as well as Coupé et al. (2019)'s finding that languages tend to keep the information rate more or less uniform.

Efficiency as management of the amount of the language users' effort has been observed in many linguistic domains. For example, in phonology, there is ample evidence that words and segments that are more predictable generally, and in a given context, undergo phonetic reduction more frequently than less predictable units (Jurafsky et al. 2001; Seyfarth 2014). For example, the schwa before /r/ and /n/ is absent in high-frequency words, such as *every*, and present in low-frequency words like *artillery* (Hooper 1976).

As for the lexicon, according to Zipf's ([1935]1965, 1949) law of abbreviation, more frequent words tend to be shorter than less frequent ones (see also Kanwal et al. 2017). This can be explained by formal reduction due to clipping (e.g., German *Automobil* > *Auto*), as well as by the use of shorter synonyms to express more frequent and accessible concepts, e.g., *car* instead of *automobile* (Zipf [1935]1965: 33).

[1] https://www.investopedia.com/terms/p/pareto-efficiency.asp. Accessed on 25.10.2020.

In grammar, efficiency manifests itself in various markedness phenomena. The asymmetries in formal marking (e.g., shorter or zero singular markers and longer plural forms) arise and persist due to the tendency to provide less formal marking to more predictable categories (e.g., singular), and more marking to less predictable ones (e.g., plural) (Haspelmath and Karjus 2018). Another manifestation of efficiency is the well-known interplay between case marking and fixed word order: in order to convey 'who did what to whom', languages tend to use either explicit case marking (e.g., Lithuanian) or rigid word order (e.g., English) (Bentz and Christiansen 2013; Blake 2001: 15; Sapir 1921: 66; Sinnemäki 2014). Moreover, it has been argued that some word order patterns are preferred if they facilitate processing. For example, when arranging the constituents that belong to the same head, speakers of VO languages tend to put long constituents after short ones, e.g., longer prepositional phrases after shorter ones (Hawkins 2004). This word order minimizes the domains required for recognizing syntactic structure, reducing simultaneous processing and working memory load. These preferences are known as the principles "Minimize Domains" and "Early Immediate Constituents" (Hawkins 2004, 2014; see also Gibson 1998).

Beyond the sentence level, efficient communication is observed when more accessible referents are expressed by shorter forms than less accessible ones, e.g., nouns for new referents vs. pronouns or zero anaphora for the referents that have already been introduced in discourse (see Ariel's 1990 Accessibility Theory). Avoidance of high cognitive costs is also associated with restrictions on the introduction of new referents because the integration of new referents in discourse structure involves high processing costs (e.g., Chafe 1987; Gibson 1998).

To summarize, efficient linguistic structures and language use are extremely diverse and can be described by different mathematical relationships. Further examples can be found in Hawkins (2014), Jaeger and Buz (2017), Levshina (2018) and Gibson et al. (2019) and other reviews. Despite these and other new findings, no language is perfectly efficient. Or as Joseph Greenberg once put it, "[a] speaker is like a lousy auto mechanic: every time [s]he fixes something in the language, [s]he screws up something else" (Croft 2002: 5). As such, we also observe other factors, such as analogy and ease of learning, which shape language structure. Therefore, efficiency is a gradual phenomenon, which is best described in terms of probabilistic or so-called 'soft' constraints (Bresnan et al. 2001). In order to test these constraints, we need large corpora and quantitative techniques. The contribution of corpus linguistics to the study of efficiency is discussed in the next section.

## 3 Corpus-based approaches to efficiency

Although ideas about the efficient use of language have a long history (e.g., Gabelentz 1901; Haiman 1983; Zipf [1935]1965), efficiency has remained below the radar of mainstream linguistic theory for most of the 20th century, with the few notable exceptions mentioned above. The interest towards efficiency as a universal principle has only recently re-emerged due to the development of large, richly annotated corpora, which allow new forms of quantitative cross-linguistic investigation. An important role is also played by cutting-edge statistical methods, which include correlation analysis, regression, and computer simulations, as well as by concepts from information theory, such as entropy and information content (e.g., contextual surprisal).

This trend started with corpus-based psycholinguistic research on phonological and grammatical reduction and enhancement, such as hypo- and hyperarticulation in language production, where research has focused mainly on English (e.g., Aylett and Turk 2004; Levy and Jaeger 2007). A more recent trend is to use large-scale cross-linguistic corpora to investigate efficiency due to increasingly easy access to large and parallel electronic corpora, including the Universal Dependencies Corpora (Zeman et al. 2020), the parallel Bible translations (Mayer and Cysouw 2014), and the OPUS corpus (Tiedemann 2012).[2] Moreover, nowadays there are new projects that allow the investigation of efficiency in spoken data, including less well documented

[2] Additionally, it is increasingly easy to create a corpus of texts from numerous languages from scratch by using web-crawling tools and automatic tokenizers, lemmatizers and parsers (e.g., the UDPipe software) – a trend which will continue to increase the number of openly available (and parallel) corpora for linguistic study.

languages, e.g., Multi-CAST (Multilingual Corpus of Annotated Spoken Texts; Haig and Schnell 2016a)[3] and DoReCo (Language DOcumentation REference COrpus).[4]

The availability of big data from corpora and the increasingly detailed information they contain (e.g., part-of-speech tags, syntactic dependencies) enables linguists to fit more complex and precise models and to ask more challenging questions than was possible during Zipf's time and for much of the 20th century. As such, researchers can and will move from investigating basic formal properties of the lexicon to deeper and more complex issues regarding morphosyntax, semantics and discourse pragmatics. For example, newly available corpora have been used to test well-known efficiency effects on large and typologically diverse samples of languages and to challenge previous claims in the literature. Some of the most important findings include:

- Speakers of different languages minimize syntactic dependency lengths in order to optimize processing (Futrell et al. 2015), in line with Gibson's (1998) Hawkins' (2004, 2014) theory (see above). At the same time, there is evidence that at least some languages do not follow this principle in some constructions, probably due to conflicting processing pressures (Liu 2020).
- Information conveyed by word order is negatively correlated with information conveyed by internal word structure (Koplenig et al. 2017), which is a generalization of the interplay between case marking and word order mentioned above.
- It has been shown that Zipf's correlation between word length and frequency is supported by a sample of texts in almost 1,000 languages (Bentz and Ferrer-i-Cancho 2016). Moreover, it has been shown by Piantadosi et al. (2011) that average predictability of words from context is more strongly correlated with word length than simple frequency.

## 4 New directions of corpus-based research on efficiency: Contributions to this special collection

Since quantitative corpus-based research on efficiency is relatively new, naturally there remain many fundamental theoretical and methodological questions, which are addressed in the contributions in this special collection. These include the four "big" questions below.

*Big Question #1.* How can we disentangle efficiency from other factors, as well as to identify different pressures related to efficient language use? These issues are related to what has been called competing motivations (Du Bois 1985) and their explanations. For example, linguists have discovered numerous universals and statistical tendencies in both the functional and formalist traditions. But it is still unclear which of them are related to efficiency, and which are not? Specifically, which types of efficient behavior explain which linguistic phenomena?

For example, Zipf's well-known law of abbreviation is usually explained by the *Principle of Least Effort* (Zipf 1949). However, a famous argument by Miller (1957) suggests that even a monkey typing randomly would at some point produce strings of characters, such that more frequent strings would be shorter than less frequent ones (similar to what we observe in natural language produced by humans). At the same time, it has been demonstrated that the random typing model does not approximate the frequency distributions found in real corpora (Ferrer-i-Cancho and Elvevåg 2010; see also Piantadosi 2014).

In studies on morphosyntax, there has been a debate on the explanations of efficient patterns in morphological marking, e.g., the tendency of languages in general to have longer marking for plural forms, which are less frequent, with the exception of languages like Welsh in which nouns that occur more frequently in the plural (e.g., birds or strawberries) have longer singulative forms (cf. Haspelmath and Karjus 2018). Some linguists argue that this has nothing to do with communicative pressures – and thus with efficiency. Instead, they explain that longer forms emerge from longer source constructions (e.g., Cristofaro 2019). However, this

[3] https://multicast.aspra.uni-bamberg.de/.

[4] http://doreco.info.

does not explain why languages overwhelmingly display a bias towards the efficient use of form-frequency correspondences, as in the example with singular and plural marking (see more details in Schmidtke-Bode et al. 2019).

In this collection, the question of efficient patterns in morphological marking is addressed by two case studies. First, it has been suggested that the scarcity of crossing dependencies in syntactic trees is due to the tendency to minimize dependency lengths (Ferrer-i-Cancho 2006). Is this tendency sufficient to explain the observed syntactic fact, or are there other independent factors at play? This question is addressed by **Yadav, Husain and Futrell** (this collection). Using dependency treebanks from 52 languages, they compare the actual trees against random baselines with the same dependency lengths as the real data. The authors find that the actual trees have fewer crossing dependencies than the random baselines, which suggests that the minimization of dependency lengths alone cannot explain the rarity of crossing dependencies. At the same time, however, the rarity of crossing dependencies and the constraints on dependency lengths can account for gap degree and well-nestedness – two well-studied formal restrictions on dependency trees.

Similarly, the length of a linguistic unit depends on its frequency or predictability (as noted above). However, length can also be thought of as a function of complexity of the superordinate unit, as posited by Menzerath's Law (Altmann 1980; Menzerath 1954), which states that longer linguistic constructions will contain shorter constituents. When applied to words and morphemes, this means that longer words will be composed of shorter morphemes, and vice versa (shorter and more complex words will contain longer morphemes). Which of these linguistic laws, i.e., Zipf's vs Menzerath's, has a stronger effect on the length of linguistic units? This question is addressed by **Stave, Paschen, Pellegrino and Seifart** in this collection. Their paper is based on the data from nine spoken corpora from the large-scale DoReCo project (see Section 3). Although the Zipfian morpheme frequency is a more powerful predictor of morpheme length, Menzerath's Law, operationalized as the number of morphemes in a word given its length (in terms of characters), has its independent contribution on the length of morphemes, especially those that occur in morphologically complex words. This study, as well as the study by Yadav et al., provide a good illustration of how one can disentangle competing motivations in language with the help of multilingual corpus data and cutting-edge methods.

*Big Question #2.* Which old observations and generalizations that have been claimed to be manifestations of efficiency can be corroborated by tests with new data? For example, it has been argued that languages use intransitive subjects as entry points for new discourse referents, serving as a kind of 'safety valve' that helps to avoid cognitive overload (Durie 2003; cf. Lambrecht's [1994: 185] 'Principle of the Separation of Reference and Role'). This role is particularly salient under conditions of high information pressure, where a relatively high number of referents has to be processed in a given stretch of discourse (Du Bois 1987: 834–836). However, this view has been challenged by Haig and Schnell (2016b) (see also Du Bois 2017). In a paper for this collection, **Schiborr, Schnell and Haig** investigate these claims against a sample of narrative spoken texts from a variety of different languages taken from the Multi-CAST collection (see Section 3). Their results show that the cognitive challenge of introducing referents is only weakly reflected in the morphosyntax. In other words, intransitive subjects do not specialize in this task. New referents are seamlessly integrated in the narrative using those structures that reflect their semantic role the best; no specialization of predicates is observed.

*Big Question #3.* While the previous questions focus on the specific communicative and cognitive tasks and linguistic phenomena, we can also "zoom out" and ask, how do global biological and cultural factors shape the language system as a whole, making it more efficient as a cultural tool? For example, Bentz (2018) has demonstrated that a high proportion of L2 speakers decreases the lexical diversity of a language as measured by the unigram-based entropy of a text. This question is also addressed in the paper by **Koplenig** (this collection), who investigates the efficiency of written language based on a parallel corpus of Bible translations in more than 1,000 languages. Efficiency is measured in information-theoretic terms, as the ratio of information transmission rate to the channel capacity. Put simply, systems that are more efficient have lower predictability of their units from preceding context. Koplenig quantifies this efficiency for every doculect (translated text in a specific language) at the level of characters and words. Notably, languages with a larger number of speakers usually have more efficient written language than languages with a smaller number of speakers. This intriguing result suggests that cultural evolution can shape language structure in many subtle ways, which can be detected only with the help of corpora and statistics.

*Big Question #4*. How do we formalize and quantify linguistic intuitions about efficiency, and how do we make these theories falsifiable? Previous studies of efficiency contain many inspiring ideas and hypotheses, but often they are not formulated precisely or in testable ways. For example, what exactly do words like 'predictability' or 'expectedness' or 'typicality' actually mean? Which frequency-based measures represent these constructs the best? What is the role of type frequency? These topics are addressed in two case studies in this collection. The focus of the contribution by **Levshina** is on the cross-linguistic tendencies in differential case marking of the transitive subject and object. Based on data from corpora of spontaneous conversations in five typologically diverse languages, she argues that the use of markers is communicatively efficient because it depends on the predictability of the syntactic roles given specific features of the referents, such as animacy or givenness, rather than on the predictability of the 'typical' features given the roles. This suggests that the conditional probability of roles given features is more relevant than the conditional probability of features given roles for explaining differential case marking.

The focus of the contribution by **Guzmán Naranjo and Becker** is on form-frequency correspondences in nominal inflection. Whereas coding length studies in the Zipfian tradition have focused on token frequency of units, the authors investigate the less well-explored measures that reflect the distribution of inflection markers as types: relative marker frequency in a paradigm, relative cell frequency, flexibility, and entropy of a marker and a paradigm cell. They find that all of these measures help predict the length of inflectional morphemes. Their data come from a cross-linguistic morphological database called UniMorph, which is supplemented by corpus data from several morphologically rich languages. These two contributions demonstrate how one can bring the theory of efficiency forward by trying out different ways of formalizing the theoretical claims.

Thus, the contributions in this special collection provide a thematically unified perspective on efficiency, which has never before been presented in one collection. The authors address the big fundamental questions on linguistic efficiency, which is now possible due to increasing access to corpus data on typologically diverse languages, and due to the careful application of statistical methods. The authors extend and challenge the existing knowledge about universal performance-based principles of human language by bringing to light new insights regarding efficiency as a universal principle of language. They also exemplify how cutting-edge quantitative approaches, coupled with cross-linguistic and multilingual corpora, are being used to investigate the many open theoretical and methodological questions in human language efficiency.

More broadly speaking, the present collection provides new insights to the fundamental question in functional linguistics and in typology: why are human languages the way they are? We believe that convergent linguistic strategies in different languages reported in the case studies emerge due to common cognitive, biological and social pressures shared by language users around the world. Of course, correlations found in corpora do not automatically imply causation. In the future, we expect that evidence from experimental linguistics will support, refine, or possibly even refute, these claims. This collection also provides illustrations of how different explanatory factors and different operationalizations can be disentangled with the help of advanced statistical methods and rich data, which will hopefully resolve the various controversies in present-day functional linguistics (see, e.g., Schmidtke-Bode et al. 2019).

**Acknowledgments:** We would like to express our gratitude to Kaius Sinnemäki and Susanne Flach (area editors) and to Jeff Good (editor-in-chief) of *Linguistics Vanguard*, for helping make this collection a reality. We also thank the many reviewers for generously sharing their time and expertise. Natalia Levshina gratefully acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no 670985, Martin Haspelmath, PI) and from the Dutch Research Foundation NWO (Gravitation grant Language in Interaction, grant number 024.001.006). Steven Moran was funded by the Swiss National Science Foundation (grant number 176305).

# References

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

Ariel, Mira. 1990. *Accessing noun-phrase antecedents*. London: Routledge.

Aylett, Matthew & Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47(1). 31–56.

Bentz, Christian. 2018. *Adaptive languages: An information-theoretic account of linguistic diversity*. Berlin: Mouton.

Bentz, Christian & Morten H. Christiansen. 2013. Linguistic adaptation: The trade-off between case marking and fixed word orders in Germanic and Romance languages. In Feng Shi & Gang Peng (eds.), *Eastward flows the great river: Festschrift in honor of Prof. William S-Y. Wang on his 80th birthday*, 48–56. Hong Kong: City University of Hong Kong Press.

Bentz, Christian & Ramon Ferrer-i-Cancho. 2016. Zipf's law of abbreviation as a language universal. In Christian Bentz, Gerhard Jäger & Igor Yanovich (eds.), *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics*, Leiden: University of Tubingen, online publication system: https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68558.

Blake, Barry J. 2001. *Case*. Cambridge: Cambridge University Press.

Bresnan, Joan, Shipra Dingare & Christopher D. Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG01 conference*, 13–32. Stanford: CSLI publications.

Chafe, Wallace. 1987. Cognitive constraints on information flow. In Russell Tomlin (ed.), *Coherence and grounding in discourse*, 21–51. Amsterdam: Benjamins.

Cohen Priva, Uriel & T. Florian Jaeger. 2018. The interdependence of frequency, predictability, and informativity in the segmental domain. *Linguistics Vanguard* 4(S2). https://doi.org/10.1515/lingvan-2017-0028.

Coupé, Christophe, Yoon Mi Oh, Dan Dediu & François Pellegrino. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communication niche. *Science Advances* 5. eeaw2594.

Cristofaro, Sonia. 2019. Taking diachronic evidence seriously: Result-oriented vs. source-oriented explanations of typological universals. In Karsten Schmidtke-Bode, Natalia Levshina, Susanne M. Michaelis & Ilja Seržant (eds.), *Explanation in typology: Diachronic sources, functional motivations and the nature of the evidence*, 25–46. Berlin: Language Science Press.

Croft, William A. 2002. On being a student of Joe Greenberg. *Linguistic Typology* 6: 3–8.

Du Bois, John W. 1985. Competing motivations. In John Haiman (ed.), *Iconicity in syntax,* 343–365. Amsterdam: John Benjamins.

Du Bois, John W. 1987. The discourse basis of ergativity. *Language* 63. 805–855.

Du Bois, John W. 2017. Ergativity in discourse and grammar. In Jessica Coon, Diane Massam & Lisa D. Travis (eds.), *Oxford handbook of ergativity*, 23–58. Oxford: Oxford University Press.

Durie, Mark. 2003. New light on information pressure. Information conduits, "escape valves", and role alignment stretching. In John Du Bois, Lorraine Kumpf & William Ashby (eds.), *Preferred argument structure. Grammar as architecture for function*, 159–196. Amsterdam: Benjamins.

Fenk, August & Gertraud Fenk. 1980. Konstanz im Kurzzeitgedächtnis – Konstanz im sprachlichen Informationsfluß. *Zeitschrift für experimentelle und angewandte Psychologie* XXVII (3). 400–414.

Fenk-Oczlon, Gertraud & August Fenk. 2010. Measuring basic tempo across languages and some implications for speech rhythm. In *Proceedings of the 11th annual conference of the international speech communication association, Interspeech 2010*, 1537–1540. Makuhari, Japan: ISCA.

Ferrer-i-Cancho, Ramon. 2006. Why do syntactic links not cross? *Europhysics Letters* 76(6). 1228.

Ferrer-i-Cancho, Ramon & Brita Elvevåg. 2010. Random texts do not exhibit the real Zipf's law-like rank distribution. *PloS One* 5(3). e9411.

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

Gabelentz, Georg von der. 1901. *Die Sprachwissenschaft. Ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Leipzig: Tauchnitz.

Gibson, Edward. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68. 1–76.

Gibson, Edward, Richard Futrell, Steven Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen & Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Science* 23(5). 389–407.

Grice, H. Paul. 1975. Logic and conversation. In Peter Cole & Jerry L. Morgan (eds.), *Syntax and semantics*, vol.3. Speech acts, 41–58. New York: Academic Press.

Haig, Geoffrey & Stefan Schnell (eds.). 2016a. *Multi-CAST (Multilingual Corpus of Annotated Spoken Texts)*. Bamberg: University of Bamberg. https://lac.uni-koeln.de/multicast/.

Haig, Geoffrey & Stefan Schnell. 2016b. The discourse basis of ergativity revisited. *Language* 92(3). 591–618.

Haiman, John. 1983. Iconic and economic motivation. *Language* 59(4). 781–819.

Haspelmath, Martin. In press. *Explaining grammatical coding asymmetries: Form-frequency correspondences and predictability*. https://goo.gl/zcJdYk.

Haspelmath, Martin & Andres Karjus. 2018. Explaining asymmetries in number marking: Singulatives, pluratives and usage frequency. *Linguistics* 55(6). 1213–1235.

Hawkins, John. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Hawkins, John. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Hawkins, John. 2014. *Cross-linguistic variation and efficiency*. Oxford: Oxford University Press.

Hooper, Joan B. 1976. Word frequency in lexical diffusion and the source of morphophonological change. In William Christie (ed.), *Current progress in historical linguistics*, 96–105. Amsterdam: North Holland.

Jaeger, T. Florian & Esteban Buz. 2017. Signal reduction and linguistic encoding. In Eva M. Fernández & Helen Smith Cairns (eds.), *The handbook of psycholinguistics*, 38–81. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781118829516.ch3.

Jurafsky, Daniel, Alan Bell, Michelle L. Gregory & William D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In Joan L. Bybee & Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 229–254. Amsterdam: John Benjamins.

Kanwal, Jasmeen, Kenny Smith, Jennifer Culbertson & 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.

Koplenig, Alexander, Peter Meyer, Sascha Wolfer & Carolin Müller-Spitzer. (2017). The statistical trade-off between word order and word structure – large-scale evidence for the principle of least effort. *PloS One* 12. e0173614.

Lambrecht, Knud. 1994. *Information structure and sentence form: Topic, focus, and the mental representation of discourse referents*. Cambridge: Cambridge University Press.

Levshina, Natalia. 2018. *Towards a theory of communicative efficiency in human languages*. Leipzig: Leipzig University habilitation thesis. https://doi.org/10.5281/zenodo.1542857.

Levy, Roger & T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In Bernhard Schlökopf, John Platt & Thomas Hoffman (eds.), *Advances in neural information processing systems (NIPS)*, vol. 19, 849–856. Cambridge, MA: MIT Press.

Liu, Zoey. 2020. Mixed evidence for crosslinguistic dependency length minimization. STUF – Language Typology and Universals 73(4). 605–663.

MacWhinney, Brian, Elizabeth Bates & Reinhold Kliegl. 1984. Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior* 23. 157–150.

Mayer, Thomas & Michael Cysouw. 2014. Creating a massively parallel bible corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri,ThierryDeclerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis (eds.), *Proceedings of the international conference on language resources and evaluation (LREC)*, Reykjavik 3158-3163: Reykjavik: European Language Resources Association (ELRA).

Menzerath, Paul. 1954. *Phonetische studien. Vol. 3: Die Architektonik des deutschen Wortschatzes*. Bonn, Hannover & Stuttgart: Dümmler.

Miller, George A. 1957. Some effects of intermittent silence. *American Journal of Psychology* 70. 311-314.

Pellegrino, François, Christophe Coupé & Egidio Marsico. 2011. A cross-language perspective on speech information rate. *Language* 87(3). 539–558.

Piantadosi, Steven, Harry Tily & Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108(9). 3526.

Piantadosi, Steven. 2014. Zipf's word frequency law in natural language: a critical review and future directions. *Psychonomic Bulletin & Review* 21(5). 1112–1130.

Sapir, Edward. 1921. *Language, an introduction to the study of speech*. New York: Harcourt, Brace & Co.

Schmidtke-Bode, Karsten, Natalia Levshina, Susanne M. Michaelis & Ilja Seržant (eds.). 2019. *Explanation in typology: Diachronic sources, functional motivations and the nature of the evidence*. Berlin: Language Science Press.

Seyfarth, Scott. 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition* 133(1). 140–155.

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

Sinnemäki, Kaius. 2014. Complexity trade-offs: A case study. In Frederick J. Newmeyer & Laurel B. Preston (eds.), *Measuring grammatical complexity*, 179–201. Oxford: Oxford University Press.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC-2012)*, 2214–2218: Istanbul: European Language Resources Association (ELRA).

Zeman, Daniel, Joakim Nivre, Mitchell Abrams, 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. Prague: Charles University. http://hdl.handle.net/11234/1-3226. See also http://universaldependencies.org.

Zipf, George. [1935]1965. *The psychobiology of language: An introduction to Dynamic Philology*. Cambridge, Mass.: M.I.T. Press.

Zipf, George. 1949. *Human behavior and the principle of least effort*. Cambridge, MA: Addison–Wesley.