In Defence of Linguistics as an Empirical Science in Light of Mario Bunge's Defence of the Scientific Treatment of Biology

Dorota Zielińska[1]

Abstract—Although few linguists currently embrace the empirical paradigm, there are increasing calls for the development of tools for studying language that resemble those in exact sciences. This trend can be observed even in top mainstream linguistic journals, such as the *Journal of Pragmatics*, as exemplified by Xiang (2017). Today, however, linguists who adapt the methodologies from more advanced sciences face isolation from the mainstream linguistic community. This is because the majority of linguists in philological and philosophical departments remain convinced that the object of their studies is fundamentally different from those studied by physicists. Therefore, they argue that linguistic methodology cannot resemble that used in empirical sciences. As a result, linguistics is often seen as requiring interpretation rather than an explanation, and evaluation of linguistic research is based on acceptance within the scholarly community rather than empirical testing.

Résumé — Bien que peu de linguistes adoptent actuellement le paradigme empirique, il y a une demande croissante pour le développement d'outils d'étude du langage similaires à ceux des sciences exactes. Cette tendance peut être observée même dans les principales revues linguistiques, telles que le *Journal of Pragmatics*, comme l'illustre Xiang (2017). Aujourd'hui, cependant, les linguistes qui adaptent les méthodologies des sciences plus avancées sont isolés de la communauté linguistique. En effet, la majorité des linguistes des départements de philologie et de philosophie restent convaincus que l'objet de leurs études est fondamentalement

[1] **Dorota Zielińska** has an M.S. in Physics and a Ph.D. in English Philology from the Jagiellonian University, Poland. She started her career as a physicist at Fermilab and at Northeastern University, USA. Upon returning to the Jagiellonian University, she focused on adapting the methodology of socio-natural sciences to linguistics in the framework of Mario Bunge. In 2013, she received qualification for a professorship in philosophy of language from MIUR, Italy, and now she continues as an independent researcher. She has established two linguistic laws, formulated within Mario Bunge's paradigm. One law, referred to in this article, pertains to the ordering of adjectives in Polish noun phrases (Zielińska 2007b). The other law addresses the position of "counterfactual if clauses" in English and Polish sentences and was presented in more detail in Zielińska (2019).

différent de ceux étudiés par les physiciens. Par conséquent, ils soutiennent que la méthodologie en linguistique ne s'apparente pas à celle utilisée dans les sciences empiriques. Par conséquent, la linguistique est souvent perçue comme une science interprétative plutôt qu'explicative, et l'évaluation de la recherche en linguistique se fonde alors sur le consensus au sein de la communauté scientifique plutôt que sur des tests empiriques.

Keywords—Language laws, Empirical paradigm, Mario Bunge, Expectation field, Operationalization.

> Follow the evidence wherever it
> leads, and question everything.
> NEIL DEGRASSE TYSON

In this article, we will critically analyse several common arguments used to support the misconception that linguistic methodology cannot resemble that used in established empirical sciences and show that they do not hold up to scrutiny. To accomplish this, we will draw inspiration from Mario Bunge's famous defence of biology as an empirical science articulated in the '60s of the previous century, during a time when many biologists vehemently opposed a scientific treatment of their discipline.

1] Part One: The History of Language Laws

Half a century ago, mainstream biologists still strongly opposed introducing the scientific method to their discipline. Groups of biologists put forward a plethora of arguments against treating biology in the same way that physicists approach physical phenomena. The arguments ranged from the objection that live organisms cannot be studied in the same way as inanimate matter, to pointing out the special role of the comparative method in biology. Today, no biologist questions the value of molecular biology, biotechnology, genetics or epigenetics—disciplines firmly placed in the empirical paradigm—despite the continual use of the comparative method when classifying newly discovered species of plants and animals.

A similar dispute had also taken place among psychologists and sociologists before many of them embraced the empirical paradigm. And what about linguistics? Must linguistics and empirical sciences belong to two distinct cultures with incompatible research methods and evaluation criteria, as Snow (2001) framed the question? While scientists and a growing number of maverick linguists adapting the

scientific method to study language answer "no" to this question, mainstream professors of linguistics in high places still advocate for there being a chasm between sciences and linguistics. Why?

According to Grzybek (2006), many contemporary linguists oppose the treatment of linguistics in parallel to empirical sciences due to the vehement criticism received by the Neogrammarians for their allegedly similar approach to studying the history of Indo-European languages. For quite a while, the Neogrammarians attempted, unsuccessfully, to find exceptionless rules of the sound changes taking place in languages over time. As a result, the idea that language can be captured with linguistic laws—as laws were understood by linguists at that time—was widely criticized and rejected by the linguistic community. However, this conclusion was drawn only because the linguists at that time, just as most mainstream linguists today, understood the concepts of a law differently from physicists.

Before explaining what distinguishes the concepts of the law in natural sciences from that entertained by the Neogrammarians and other mainstream linguists, first let me note that physicists search for two types of laws: summarizing and explanatory. Summarizing laws, such as Kepler's laws, are descriptive laws that summarize patterns from observed data, as algebraic formulae, to answer the question of **how** things behave. For instance, Kepler discovered patterns in the movements of planets using data collected by Tycho Brahe and expressed them as formulae for ellipses. Explanatory laws, on the other hand, hypothesize the material causes of such patterns by positing the causative role of some material characteristics of the observed phenomena. Newton's laws, for instance, which explain the movement of material bodies, are such explanatory laws. They can be used to explain Kepler's summarizing laws, elucidate **why** planets orbit the Sun in elliptical paths.

The Neogrammarians searched solely for summarizing laws, which is the first difference in understanding the concept of law in physics and traditional linguistics. However, it should be noted that looking exclusively for summarizing laws is a legitimate goal of proper scientific research, too. The crucial difference in understanding the concept of law in those two disciplines concerned the fact that all linguists at that time believed in the exceptionless nature of all laws and assumed that laws are always deterministic and

thus can always be expressed as algebraic formulae. In this sense, these potential linguistic laws were meant to resemble Kepler's laws. Therefore the Neogrammarians, searching for sound change laws in language viewed as an abstract structure (*langue,* to use Saussure's terminology), hoped to discover laws that fit the data (consisting of sound types, not sound tokens) perfectly and could be expressed with algebraic formulae.

By the end of the 19th century, having failed to find such laws, the Neogrammarians and other mainstream linguists observing them, began to develop an aversion to linking linguistics to the empirical sciences. The linguistic community began to embrace the view that language differed so significantly from physical phenomena that linguistic studies required a methodology completely different from that employed in the natural sciences. These scholars believed that linguistic research required interpretation rather than an explanation, and thus, it was necessary to assess the merit of such research based on acceptance within the discipline-specific scholarly community, rather than through empirical testing.

It was not until the second half of the 20th century that Noam Chomsky, the most cited linguist ever, acknowledged the importance of considering the material causes of language (*langue*) and proposed that language has its origins in psychological processes. However, he also held the view that the task of linguists is to find algebraic-like, exceptionless algorithms that generate various **types**, not **tokens,** of sentences, and he delegated the task of discovering the causes of such linguistic laws to psychologists. In other words, Chomsky and his followers, known as generativists, just like the Neogrammarians, sought to find linguistic laws expressed as algebraic formulae perfectly summarizing the observed data consisting of **types** of linguistic items. This assumption of the existence of an algorithm that captures the generation of every sentence type in a language, made the generativists' effort destined to fail, for reasons that will be explained soon.

At the end of the 20th century, in reaction to generativists' efforts falling short of expectations, linguistics witnessed a cognitive turn. Cognitivist linguists, among them prominently, Ronald Langacker and George Lakoff, independently proclaimed that language mechanisms cannot be captured with laws understood in the same way since the Neogrammarians. Lakoff (1987) illustrated his claim by

arguing that there cannot be a general semantic law concerning the meanings of compound words that, for instance, can derive the meaning of the lexeme *overlook* from the meanings of the lexemes *over* and *look*. The meanings of *over* and *look* can only "motivate" the meaning of the lexeme *overlook*. In other words, these meanings can only indicate that there is SOME relationship between the meaning of *overlook* and the meanings of its components and thus, it makes sense that *overlook* means what it does.

While this observation is true, it is important to note that the reason why these linguists failed to find exceptionless laws was because these linguists were concerned only with language as an abstract structure (Saussure's *langue*), and they understand the concept of law as an exceptionless algebraic formula summarizing data. Bunge explains below why such assumptions prevented these scholars from succeeding:

> Languages [treated as *langue*—D.Z.] do not develop or evolve by themselves and there are no mechanisms of linguistic changes, in particular evolutionary forces. Only concrete things, such as people can develop and evolve. And, of course, as they develop or evolve, they modify, introduce, jettison linguistic expressions. The history of mathematics is parallel: mathematicians do come up with new mathematical ideas, which are adopted or rejected by the mathematical community, but mathematics does not evolve by itself. (Bunge 2003: 62)

In other words, Bunge argues that since abstract systems, such as *langue*, cannot change by themselves, therefore there cannot be empirical laws of *langue* describing such change or its results. However, Bunge's argument implies that, within an empirical paradigm, one may legitimately search both for explanatory and summarizing language laws concerning *situated parol*. *Situated parol* refers to utterances pronounced on a specific occasion by specific interlocutors involved in a specific communication process, which means it is a verbal aspect of the communication[2] process taking place in the system of material bodies of people participating in verbal interactions in specific socio-natural contexts, also known as

[2] It is important to note that since *situated parole* is an aspect of the communication process, describing it fully must involve comprehension. A similar view was already expressed by Dummett (1993: 12), who stated: "a theory of meaning must also be a theory of understanding" (cf. Searl 1983).

situated speech acts. Because *situated parole* is an aspect of a material system, a psycho-socio-natural phenomenon of communication, it can be researched within an empirical paradigm, and described with language laws. At this point an open question remains, as to whether language laws searched for in the empirical paradigm can always be captured in terms of exceptionless algebraic formula. This depends solely on the characteristics of the "material system" that produces *situated parole,* or more precisely, on our knowledge of those characteristics.

So what do we know about that material system generating *situated parol*? Since human cognitive capabilities are the result of self-organising and self-regulating, non-linear processes, it is reasonable to assume that the human ability to form and use language is also shaped by such processes. Consequently, an explanatory theory of a person's idiosyncratic language must be a theory of language acquisition and use by that individual. Such a theory must reflect **the history** of the interlocutor's solving specific communicative challenges in specific situations based on socio-cognitive mechanisms operating against the background of the correlations between language forms and meanings already engraved in their memory.

Assuming language has self-organised and keeps self-regulating, similar to all natural, self-organising, non-linear systems, we cannot expect to be able to predict the occurrence of a particular utterance, a novel sentence pattern, or the meaning of a novel compound word with exceptionless algebraic laws. Just as much, as we cannot predict the exact characteristics of a specific volcano eruption, the shape and timing of a specific avalanche, or of a specific tornado. This is because these outcomes depend on the specific history of the development of the "material system" in question, which can never be known with sufficient precision. Furthermore, being non-linear implies that even slight imprecision in their measurement makes any long-term predictions futile. Therefore, all we can say about such systems is describing trends in their development and results, meaning we can only define stochastic laws for them.

In the same vein, the only type of language laws that can be discovered and tested within the empirical paradigm are **probabilistic** laws that model **trends** in the occurrences of such specific utterances (i.e., trends in *situated parole*). In other words, we cannot

hope to ever find exceptionless algebraic laws concerning *langue* (understood as trends, dominant patters in language use), which is what the Neogrammarians, and the Chomskyians sought to uncover.

Given that the vast majority of linguists still hold the view that *langue* is the object of the science of language and that laws can be expressed as exceptionless algebraic formula, the prevailing belief in mainstream linguistics is that "there can be no language laws". This belief is supported by several accompanying myths, akin to those Mario Bunge dispelled in his defence of biology as an empirical science over half a century ago. In the following sections, I will address some myths in linguistics that discourage many linguists from embarking on the empirical paradigm.

2] Part Two: Myths

2.1] Myth One: Linguistic rules are non-nomothetic, while empirical sciences are concerned with natural phenomena describable with nomothetic laws.

One of the most prominent arguments for the belief that linguistic laws differ fundamentally from physical laws has been the assertion that the latter have exclusively nomothetic character, while the former are non-nomothetic. However, this argument is flawed because not all laws in physics are nomothetic.

The term "nomothetic" was introduced by Windelband in the 19th century, meaning "deterministic, based on deduction." A few years later, Windelband, along with his disciple Ricket, proposed that sciences differ from non-empirical disciplines by being concerned with the phenomena describable with nomothetic laws. In the late 19th century, William Dilthey used this distinction to exclude sociology from the family of disciplines that can be studied within an empirical paradigm. He also declared that the objective of the humanities are singularities and individualities of socio-historical reality.

Let us examine this claim in some detail. First and foremost, we must remember that when we discuss laws in empirical sciences, whether deterministic (nomothetic) or not, we are really talking about **our knowledge of these systems, and not some objective laws of nature**. Furthermore, such knowledge changes with time. When the Neogrammarians presupposed that linguistic sound

laws are deterministic[3] (mechanistic), exceptionless like the laws of Newtonian physics, they were not aware that the situation in physics had undergone a profound change in 1877. That year, Boltzmann introduced a non-deterministic law into the realm of physics by redefining the Second Law of Thermodynamics in terms of probability.

To explain the significance of that shift and to provide the essence of the new interpretation of the Second Law of Thermodynamics, it is necessary to start by introducing the First Law of Thermodynamics. The First Law states that the energy of a closed system, one without external influences, cannot change. However, there may be many states of a given system with the same energy, and the First Law does not indicate which of these states will be realized. The Second Law of Thermodynamics addresses this issue by stating that the entropy of processes occurring in closed systems cannot decrease. Loosely speaking, it means that the system cannot become more orderly without receiving energy from outside.

To illustrate the idea behind the Second Law of Thermodynamics, we can use the analogy of the Law of Messy Rooms, describing the mess in our rooms. This Law can be formulated as follows: "We never make rooms tidier accidentally—without our conscious effort to do so". This law corresponds, to some degree, to the Second Law of Thermodynamics, which states that the entropy of closed systems does not decrease. Why? From a probabilistic perspective, the answer is straightforward. There are a vast number of states (potential arrangements of things in our room), that we would consider messy, but only a few that we would classify as tidy. For example, placing socks on any other square inch of your room except in the proper drawer results in increasing the state of the mess.

Now, let us imagine, we start moving things in a room at random, without conscious effort to place them where they belong. Assuming the frequency definition of probability (as the ratio of the number of states to the number of all possible states), the probability that we will arrive at the exceptional state (a tidy room) is the ratio of the states in which everything is in its proper place to the number of all

[3] The notion of causality has a much more complex meaning in contemporary philosophy of science than in common perception inherited from Descartes, who said that a perfect science is about inferring the consequences from causes. A presentation of the contemporary concept of causality can be found in Bunge (1959).

possible arrangements of things, which is very small. This means that the probability for uncoordinated (random) moves, such as dropping books and socks, to result in a messy room, rather than in a tidy one, is much, much greater, regardless of the initial state of the room. This shows that the Law of Messy Rooms does not describe any fundamental aspect of human nature but rather the lack of human propensity to keep things tidy, coupled with the limitations of the physical space in rooms.

Similarly, the Second Law of Thermodynamics does not describe any fundamental property of nature, any "force" (propensity) determining the behaviour of thermodynamical systems, such as gases, but that it results from the special characteristics of the environment of the system. The behaviour described by the Second Law, implying for instance that the particles of oxygen in the rooms we live in do not gather suddenly in a given cubic inch of the room under the ceiling, causing us to suffocate—is, to a large extent, the result of pure statistics. It reflects the ratio of the number of states in which these particles are all in the same given cubic inch, to the number of all possible positions of oxygen particles in the room. Consequently, after Boltzmann's proof, the concept of "law" stopped being exclusively a term for a deterministic relation of "cause" and "effect" allowing no exceptions, but it also started to include non-nomothetic laws—the descriptions of some complex totality in terms of probability[4, 5].

Non-nomothetic laws are employed in scientific contexts in situations when we lack sufficient information about the system, even when every its elements are governed by strict rules. In many such cases, especially in complex non-linear systems, we may not have complete information about all the elements and interactions in a system, or sufficient computational power required to model the

[4] The part of Myth One down to this point restates arguments presented in Grzybek (2006).

[5] Half a century after Bolzman's work, in 1922, Schrödinger raised the question motivated by quantum mechanical considerations that possibly all natural laws were statistical in nature. John Wheeler, based on his research in general relativity and quantum gravity, again came to a similar conclusion in 1994 stating that "every law of physics pushed to its extreme, will be found to be statistical and approximate, not mathematically perfect and precise." Wheeler (1994:293).

system step by step. At the same time, because of their non-linear character[6], it is not possible to calculate an approximate solution.

A class of phenomena governed by strict rules, yet without deterministic solutions, meaning their overall outcomes can only be learned by carrying out all procedures step by step, are games such as chess, go, or certain card games. Coping with such evolving systems requires powerful tools based on statistics. In his memoir (Ulam 1991) describes how he invented one of such methods of gaining information, the Monte Carlo method, while playing solitaires during his stay at Los Alamos. Since then, this statistical method has become a standard tool in many disciplines.

> I noticed that to assess the probability of laying a solitaire (such one as Canfield, in which the skills of a player are of little importance), it is much more practical to "expound cards", to experiment with that process and put down the percent of wins than to try to calculate all combinatorial possibilities, whose number grows exponentially and is so big that, except for the most basic situations, it is impossible to estimate. This is surprising from the intellectual point of view and although not quite humiliating, it forces one to be modest and shows the limitations of rational thinking.

In scientific contexts, statistical laws are also necessary for estimating the parameters of individual components based on the global characteristics of complex liner systems. For example, to calculate the parameters of a given gas particle at some point in time, we would need to know the initial parameters of every gas particle in the container. However, measuring the initial parameters of each element of such a big system is impossible. Instead, we estimate the speed of an individual particle in a gas based on the global characteristics of that gas, such as volume, temperature, and pressure. This way, however, since the values of these parameters are related

[6] Systems whose behaviour cannot be approximated linearly are characterized by a lack of proportionality between the magnitude of an input and the resulting output. In other words, the relationship between the input and output is not simply a matter of scaling, and doubling the input does not necessarily result in doubling the output. This makes it difficult to predict the behaviour of the system, even if we have a good understanding of its individual elements. In these cases, non-nomothetic laws are often used in scientific enquiry, as they allow for a more flexible and probabilistic understanding of the system's behaviour.

to the average speed of all particles, we can only make a statistical guess as to the speed of the particle being considered.

To sum up, the need for statistical laws in physics is abundant. Therefore, today it is no longer accurate to say that what distinguishes sciences from other fields is the absence of non-nomothetic laws.

2.2] Myth Two[7]: History plays an important part in linguistics, but not in physics.

There is a common misconception that history plays a crucial role in linguistics but not in physics. Some argue that understanding the origin and development of language is essential for understanding language itself, whereas physicists study a world consisting of eternal, unchangeable, identical particles that have no historical context that would be relevant to their present-day characteristics.

However, the belief that the history of physical objects has no relevance to physics is misguided. While individual types of particles, such as an electron, may be eternal, individual electrons are not. Individual electrons may be generated and absorbed in various reactions, which phenomena are the subject matter of elementary particle physics. Similarly, the evolution of atoms, chemical elements, molecules, and materials is studied by chemistry, molecular paleontology, and historical geology, respectively, while the evolution of stars, galaxies, and other astronomical systems is studied by cosmologists. Therefore, the history of the development of objects is also a subject of study in empirical sciences. However, what matters in these studies is not only the description of successive stages of evolution, but also the discovery of relevant laws concerning the evolutionary mechanisms and the conditions under which those laws operate to explain the cause behind the evolution. (This is, by the way, exactly what the Neogrammarians unsuccessfully attempted to do when describing sound changes.)

An extreme way of employing history to learn about physical reality has been offered by the Weak and Strong Anthropic Principles. These Principles propose to explore the consequences of the very fact of the presence of different objects—galaxies, stars, planets with life on at least one of them—to place constrains on how the

[7] The discussion of the myths 2-6 has been inspired by Bunge's reply to biologists arguing against the empirical method in biology.

Universe has been developing. In 1987, using Anthropic Principles, Weinberg demonstrated that the limits on the amount of vacuum energy in the Universe must be at least 118 orders of magnitude smaller—that is, a factor of 10^{118}—than the value obtained from quantum field theory calculations. When dark energy was empirically discovered in 1998, its measurement turned out to be 120 orders of magnitude (a factor of 10^{120}) smaller than that calculated from quantum field theory, and remarkably close to the naïve prediction following from the Weak Anthropic Principle; the difference being only two orders of magnitude[8].

Moreover, the mechanisms driving the evolution of physical objects and language systems have much in common. As Bunge (2003) explains, the evolutionary mechanisms in physics have been self-assembly, spontaneous mutation and the selection by the environment. It may come as a surprise to some, but these three classes of phenomena also manifest in language. Self-assembly in language is evidenced by grammar and by power laws that describe many statistical characteristics of language. Spontaneous mutations in language include *ad hoc* "ungrammatical" constructions, novel lexemes created "inadvertently", so-called slips of the tongue, or even novel items created purposefully (such as *iv3rm3ctin* used to mean *ivermectin* on social media). These mutations are unpredictable, but if they are useful enough to be repeated by a sufficient number of members of a given linguistic community, they will become engraved in the memories of the interlocutors and thus indirectly in the system. If not useful, such novel forms will disappear from language due to not being repeated frequently enough, thus forgotten. In other words, new words and patterns will become retained in language if selected by the environment.

In summary, both physical phenomena and language are subject to historical processes, which are driven by similar evolutionary mechanisms such as self-assembly, spontaneous mutation, and selection by environment. Therefore, history is just as important when searching for the essence of physical phenomena, as it is when learning about language.

[8] This observation was made by Ethan Siegel (2022).

2.3] Myth Three: Linguistics can explain at most the facts which have occurred, while physics both accounts for past observations and makes predictions of future events.

Many humanists believe that while physics can make predictions about future events based on past observations, linguistics can only explain facts that have already occurred. Etymologists look back in search of the origins of words, and no branch of linguistics can predict the specific forms and meanings of future words, which was illustrated in the introduction with a brief discussion of the meaning of the lexeme *overlook*, as composed of the lexemes *over* and *look*.

2.3.1] Predictability in Different Disciplines

However, as Bunge (1973:56) notes, predictability is not inherent in things, but in our knowledge of them: "It depends both on the sophistication of existing theories and on the available precision of the data's description". The sophistication of existing theories refers both to the quality of theories *per se* and to that of models to which theories are applied. The precision of data description reflects the degree to which something can be characterized objectively (independently of the person undertaking the description), for instance, how objectively and precisely someone's height can be measured, or the meaning of some linguistic item described. If a discipline has theories which are too general, or data that cannot be described precisely enough for a specific theory to be applicable, then no specific predictions, or retrodictions can be made.

For instance, Darwin's theory—like general quantum theory, by the way, is very general, and thus, it can predict only general trends, rather than specific events. However, if we included a more specific description of the data in line with a more specific model of the species in question and of their environment, the resultant predictions would be much more precise. As Bunge (1973: 57) notices : "the predictive poverty of the theory of evolution is a mark of its generality, rather than the evidence for the lawlessness of organisms".

Nonetheless, in some circumstances, Darwin's theory is still capable of providing specific answers, too. For example, it can identify missing links in an evolutionary sequence by determining **which** of the exemplars found meets the criteria for being a missing link, even when those criteria are only specified in general terms. In this

sense, Darwin's theory can be predictive and a valuable tool in paleontology research.

Based on the information presented in the introduction of this paper, there is no reason why linguists cannot adopt an approach similar to that taken in socio-natural sciences. Specifically, it should be possible for linguists to propose an empirical theory of language, viewed as an aspect of the psychosocial-natural phenomenon of verbal communication, and postulate and test hypotheses implied by the theory on some linguistic corpora collected in the future or in psychological experiments. However, as with Darwin's theory, due to the complex, non-linear nature of language formation within this approach, linguists should expect to discover theories that enable the postulation and testing only of probabilistic laws that model **trends** in the occurrences of specific utterances.

2.3.2] Examples of Probabilistic Language Laws

An example of research testing probabilistic language laws is Zielińska (2019) study. Zielińska postulated and tested hypotheses concerning linguistic trends implied by the Field Theory of Language (FTL)[9], which was coined within Bunge's (2003) systemism *cum emergentism* framework. The first hypothesis tested was that "counterfactual *before* time clauses" tend to precede main clauses in sentences, and the second was that "counterfactual *before* time clauses"[10] are more likely to be the first clause in a sentence than "non-counterfactual *before* time clauses".

2.3.3] The Field Theory of Language (FTL)

Before explaining, why Zielińska postulated her hypotheses, and how she tested them, it is important to understand the underlying framework of Field Theory of Language (FTL). Coined within the empirical paradigm of socio-natural sciences as explicated by Mario

[9] The field theory of language is an extension of the communicative field theory of language presented in Zielińska (1999, 2003, 2007a,b, 2014) and recently further elaborated on in the chapters "*How Does Language Work*?" in Zielińska (2020a) and "*Testing the Advocated Theory of Language. The Studies of the Order of Polish Adjectives in Noun Clauses and the Order of Unfulfilled Before Clauses"* in Zielińska (2020b).

[10] To clarify the terminology, let us consider at the sentence *She died* ***before she graduated***. In this sentence, the clause *she died*, which can stand on its own, is the main clause, while the clause *before she graduated* is a subordinate time clause, more precisely, a subordinate, counterfactual *before* time clause.

Bunge (2003), FTL aims to capture the material causes of "languaging", by which I mean language use and formation, correlated with the brain activity reflecting socio-cognitive behaviour of the interlocutor.

2.3.4] The Mechanism of Self-Organisation and Self-Regulation in the Field Theory of Language

The first major assumption of the Field Model of Language (FTL) is that the material causes driving language formation and self-regulation are grounded in the characteristics of human bodies functioning in societies. More exactly, the process of "languaging" (language use and its self-regulation) is constrained by the assumption that a language system arising in a society develops through its members' reacting to the properties and requirements of their environment via some sort of adaptation mechanisms, as explicated by Koehler and Altmann (2005). For example, the way human memory and cognitive apparatus function suggests that certain phonetic/graphical representations of words or language constructions and their meanings that co-occur on a given occasion are more likely to become permanently correlated in the brain if certain conditions are met. These conditions influencing the formation and retention of language items and their meanings include

- high frequency of occurrence motivated by frequent need;
- relating to basic level items (e.g. "dog" as opposed to "dachshund" and "animal", which are functionally more distinct);
- being shorter or less complex than close functional alternatives;
- not being too short, thus, putting too much burden on the addressee when decoding (to avoid misunderstanding);
- communicating content with adequate precision;
- fitting the dominant language grammar and semantic structures appropriately, which makes it easier to understand and recall forms used;
- enhancing communication;
- and such.

Based on the above, the pairs of {words and their meanings in use}_events that become engrained in memory best (by forming new neuronal connections or readjusting the strengths of the synapses

that already existed, modifying neuronal activation paths) form the basis for individual "languages" in the brain, which consists of items that are most easily remembered and most useful for communication. The items that are retained in memory allow for efficient communication, while balancing the needs of both listeners and speakers. Therefore, for language to self-regulate, interlocutors do not need to consciously strive to choose language solutions that are optimal for the language system, as postulated by Zipf's Principle of Minimal Effort nearly a century ago. Instead, self-regulation of language is mainly the result of unconscious[11] processes, such as remembering more frequently repeated items best.

2.3.5] The Categorization Mechanism in the Field Theory of Language

The second founding assumption of the Field Theory of Language (FTL) concerns the mechanism of semantic categorization, which generates specific language events (specific form-meaning pairs in use, aka instances of *situated parole*). It is postulated that listeners generate *situated parole* by using words either encodingly or selectively.

People arrive at the interpretation of words and sentences (assign meaning to forms or the other way round) **selectively**, similar to how two points define and identify a line, **assuming you know** we are talking about lines, not circles. In the same way, the encoded content of words serves to identify one of the expectations generated in the minds of the interlocutors. The fact that people generate expectations about what the world around them will look like in a moment, including what can **likely, and with what likelihood,** be said and done next during a verbal interaction, has been well established. These expectations are formed primarily due to the interlocutors' awareness of some aspects of the socio-natural environment of the verbal interaction the interlocutors participate in (the situated speech act), what has been said so far, and the relevant experience available to them at that moment, all of which are passed through their attention and intention filter. In FTL, these

[11] Zipf's (1949) alternative assumption that interlocutors consciously optimize language, known as the Principle of the Least Effort, was criticized for its cognitive feasibility, and rightly so. Unfortunately, this criticism led to the dismissal of Zipf's ground-breaking idea of socio-natural source of language formation for almost half a century until it was rediscovered by the Neo-Zipfians, aiming at grounding linguistics firmly in empirical sciences.

expectations, each with assigned probability of occurrence, are referred to as one's "expectation field". Importantly, then the expectation field is only a tiny fraction of what interlocutors know. It is not merely substantially limited by their attention focus and intentions, but also depends heavily on the associations such limited information accessible to them recently generate.

A visual analogy that helps to emphasize the constraining and guiding function of the expectation field postulated in FTL, which is crucial for interpreting language, has been offered by the picture below. When prompted with the same utterance, *I need a mouse*, interlocutors with different backgrounds and attention foci, generate different expectation fields leading them to select different, idiosyncratic interpretations of the same verbal clue.

Selected meaning arrived at in the specific language events establishes, for the first time or by adjusting, **the current encoded meanings** of the lower-lever units that comprise the just-interpreted construction. The adjusted form-meaning pairings that emerge from this process are subsequently stored in long-term memory, with factors such as repetition frequency, brevity, similarity to common items in language, and such, influencing the likelihood of their retention. A similar retention process guides

establishing word patterns within sentences, sentence patterns within texts, and correlations between specific words and word class patterns in which they tend to occur. Each subsequent unit of organisation is organised in the way that reflects the most efficient patterns both for speakers and listeners that the interlocutors have encountered in the past and remembered. This process of remembering certain co-occurring patterns or form-meaning correlations ultimately contributes to the passive self-regulation of language.

2.3.6] Selected Meaning

When the listeners use encoded content of the words they have just heard to selectively identify some percept in their **expectation field**, top-down[12], these words need not fully encode the content of the item identified in the expectation field. The selected meaning is typically much broader than, and may even differ significantly from, the sum of the encoded meanings of the constituent words[13] used during the selection (interpreting) process.

To illustrate the categorization mechanism just postulated, let's consider the interpretation of the phrase *a red rose*. The selected meaning of this phrase is richer than the sum of the encoded meanings of its constituent parts—the meanings of the words *red* and

[12] Note that the mechanism of categorization introduced in FTL deals with the meanings of novel linguistic forms, somewhat similar to the way paleontologists use Darwin's theory to classify missing links. While paleontologists cannot predict exactly what a missing link will look like beforehand, they recognize it once they have discovered it. Similarly, linguists could not have foreseen what the lexeme *a computer game* would mean when computers were first invented, but the term became a perfect name for computer games once they were invented. People who knew of the existence of computer games were able to understand the term, even when hearing it for the first time, in the appropriate context that generated in the listeners' minds an expectation field of options likely to be discussed in the given situation (situated speech act). Such human capability of choosing the best fitting option has been acknowledged by psychologists in relation to language acquisition by toddlers, who are constantly faced with the need to identify referents of novel vocabulary items in pragmatic contexts. In psychology, this phenomenon is called fast mapping.

[13] The most recent models of vision work in a similar way. They propose that what we perceive at a given moment is what we calculated within the past half-a-second, while making the predictions as to what we would see (interpret to see) in half-a-second, based on the data available to us half a second ago. In other words, our perception at a given moment is the result of continuous process of prediction and selection/interpretation. Our brain is constantly making predictions about the future, based on the data available to us in the past, and our perception is a continuous process of updating and refining these predictions in real time.

rose. A *red rose* is not entirely red, instead, it is mostly green, with only a small portion of it being red—its petals. In addition to the meanings of the lexemes *red* and *rose, which* influence phrase's meaning bottom-up, the shade and distribution of redness involved in the interpretation of the phrase "red rose" come from our experience with flowers, and roses in particular—thus, top-down from our expectation field. When interpreting this phrase, the interlocutor first generates a specific field of expectations about what flowers, in particular roses, look like, based on their prior experience. Second, they use the encoded meanings of the lexemes *red* and *rose* to find a rose that is redder than some other roses (white, yellow, pink) and "rosier" than other items.

Note that a typical dictionary (encoded) meanings of *red* and *rose* should be considered rather as proto-meanings. These proto-meanings assume their actual, selected (pragmatic) meanings, only when used in specific phrases uttered on particular occasions (situated speech acts) similar to how Bunge (2003) discusses proto-entities in self-organising systems. (On second thought, it becomes apparent that we almost never use words solely to convey their encoded content, as exemplified by phrases such as *dust furniture* vs. *dust a cake with sugar, a hot day in Stockholm in winter* vs. *a hot day in Miami in summer, a big child* vs. *a big whale, a red bike* vs. *a red pen, a horse is running* vs. *a baby is running*). The encoded content of words primarily serves to indicate **which** item in the field of expectations we are referring to, and only indirectly, **what** that item is like.

It may also be helpful to note that selective categorization, as postulated by FTL, is similar to how pronouns (*he, she,* …) are commonly believed to operate. Pronouns typically point out most of their content from a set of options that are viable on a given occasion, instead of fully encoding their contextualized referents. For example, the meaning of *you* in the phrase *you are right* when spoken by John to Mary, primarily derives from interlocutors' knowledge of the addressee's identity and the knowledge that *you* singles out the addressee. According to the FTL view, this mechanism is not limited to pronouns, but can apply to all lexemes, linguistic constructions, even texts, to indicate "which one it is", akin to how pronouns function, and only indirectly convey specific characteristics of the referents.

2.3.7] Selecting Illocutionary Force and Strong Pragmatic Meaning

The mechanism of categorization by selection in the expectation field is often also used to identify the purpose behind the sentence uttered, known as the illocutionary force, as well as its strong pragmatic meaning. Due to the mechanism of categorization postulated, the purpose of an utterance and its strong pragmatic meaning don't even need to be semantically related to the meanings of the words used to convey them. For example, in response to the suggestion "Let's go for a walk" the sentence *it's raining* will likely be interpreted as rejecting the proposal. This is because the interlocutor's expectation field contains two options for the purpose of that response: "accepting the invitation" and "rejecting the invitation" and the sentence *it is raining* serves to distinguish between the two options. Since people typically do not enjoy walking in the rain, the sentence *it is raining* selects the option of "rejecting the invitation", essentially conveying the strong pragmatic meaning of "Let's not go for a walk, because it is raining"[14].

2.3.8] The Characteristics of Language Organisation Levels in the Field Model of Language

The Field Theory of Language posits that language is a system of successive levels of meaningful language units (except for the lowest-level building blocks—letters). Letters group into morphemes, morphemes group into words, words into phrases and sentences, and sentences may group into larger functional unites, such as reports, letters of recommendation, or poems. Each successive level of organisation is characterized by qualitatively new properties and these levels interact with each other both bottom-up and top-down.

At the lowest level, letters have form. At the next level morphemes and words acquire the novel quality of having a representation and thus being able to be used to refer to something[15]. Several

[14] Similarly, the FTL categorization mechanism is also of great value in establishing the information structure of utterances, stating which part of the information is already known, has been talked about and which is new. In the FTL view that division need not adhere to the divisions imposed by the grammatical structure of the sentence, which allows one to describe the information conveyed by sentences much more precisely than possible in traditional approaches.

[15] Actually, the form of a word has emergent quality, too. The 1D graphical form of letters becomes additionally 2D when they are put together into a word, and its

emergent qualities arise at the level of forming phrases and sentences. The first novelty at that level is the emergent representational meaning reflecting the fact that *a black car* does not mean simply that it is all black. (Linguists refer to this as enriched meaning or weak pragmatic meaning.) Secondly, the enriched meaning of phrases and sentences can convey information by attributing quality A to B, which is another novelty at that level of organisation[16]. The sentence *The Porsche can go fast* may be used to inform about the Porsch's ability to go fast or convey the message that a thing that can go fast **is a Porsche**. (Dividing the sentence explicitly into the part conveying what is being assessed, the 'Given', and what is the 'New', i.e., stated about the 'Given', is called explicating the information structure of that sentence.) Thirdly, such a message can be evaluated as true or false, which is yet another emergent quality[17] of language at the sentence level.

Note, that the same sentence may express both a true and a false proposition, depending on the assigned information structure. For instance, the sentence 'English is spoken in Burma' is true when it is a reply to the question 'which language is spoken in Burma?' with English being the "New' information. However, if we consider the sentence 'English is spoken in Burma' with the words in Burma as the 'New' information, it does not provide the expected, true answer to the question: 'where (in which countries) is English spoken?' The correct answer to the latter question could be: 'English is spoken primarily in England, Canada, USA, Ireland, Australia, RPA, but also in such countries as Burma.' The difference in truth values between the utterances discussed arises from the changed interpretation of words in the sentence discussed when different elements are assigned the status of 'Given' and 'New' respectively. Such a difference in information structure of a sentence can generate different expectation fields during its interpretation process, resulting in

phonetical form is not a simple phonetic realization of the string of the constituent phonemes pronounced in isolation.

[16] In fact, after adopting Shannon's definition of information, we might even ask how informative a given message is.

[17] Note that depending again on which options the message serves to eliminate when informing (cf. Shannon's definition of information), the given sentence may be true when answering one question and false answering another one. For instance, the sentence *ships unload at night* is true when answering the question *when do ships unload?,* but false in response to *what do ships do at night?* Both these examples come from Barbara Partee.

different outcomes. Different interpretations open the possibility that one messages is true, while the other may not be true.

The difference in the interpretation of the same sentence due to different information structures assigned, however, need not merely parallel the difference between conditional probabilities P(A|B) and P(B|A), as was the case with the sentence discussed above. For instance, when seeking information about [how many people read few books?], the sentence 'Many people read few books', with many being assigned the 'New' status, will be interpreted as meaning that every person reads a different set of books. However, when answering the question 'Are there many books that many people read?' (or 'Are there many books read by many people?'), the same sentence with the word 'few' having the 'New' status refers to one set of commonly read books, such as The Bible, The Torah and The Quran.

By analogy, consider the possible messages conveyed by the sentence 'The university did not accept many candidates' when the status of 'New' is assigned to the words 'did not' and 'many' respectively. For example, it could be used as a response to the question 'Are there many students in the incoming class?', or to the question 'How many applicants were rejected by the university?', respectively. As illustrated again, a structurally and lexically unambiguous sentence does not necessarily have a single representation prior to being interpreted in a communicative situation. Interpreting a sentence may require knowledge of its information structure. In other words, the resulting explication is not related to the well-known issue of lexico-grammatical disambiguation, as is the case of the sentence 'Fruit flies like bananas'. Instead, it arises from the ways expectation fields evolve during the interpretation process of a sentence with different information structures imposed by their respective communicative contexts.

Fourthly, sentences treated as aspects of verbal interaction in communicative contexts acquire the emergent quality of having illocutionary force and strong pragmatic meaning. On one hand, they serve to perform various social actions, achieving various goals, such as that of scaring, instructing, asking, baptizing, and such, which are called the illocutionary forces of a sentence. On the other hand, sentences convey strong pragmatic meaning, which includes information about its illocutionary force, explicating what has been

said and why. Strong pragmatic meaning when selected may differ from the message the sentence in question conveys "literally" (i.e., as the weak pragmatic message), as was illustrated by the phrase *it is raining* used to say "no, let's not go for a walk because it is raining" that was discussed earlier. Strong pragmatic meaning can be revealed by reporting on what someone said using reported speech as in "John: 'Let's go for a walk'." can be reported as "He rejected the suggestion to go for a walk because of the rain".

Finally, at the highest level of language organisation, certain groups of sentences together, such as paragraphs, whole texts, or speeches may acquire a joined illocutionary force, and joined pragmatic meaning, which is not a simple sum of the illocutionary forces and of strong pragmatic meanings of the sentences, respectively. Identifying the joined illocutionary force is necessary to understand the purpose of a given text and what it has accomplished, which may include misleading, manipulating, or constituting a letter of recommendation, among others. For example, when reading a description of a person, such as "He has a beautiful handwriting", to understand what that text conveys, one must know whether it constitutes a letter of recommendation for a graduate school of engineering, or a school essay. Explicating what the speaker meant to convey with his text, what that text accomplishes, such as that they wrote a very strong letter of reference truly recommending the candidate for the job, is called the strong pragmatic meaning of that text.

2.3.9] Selecting the Information Structure

As mentioned earlier when discussing the emergent qualities of language, the Field Theory of Language provides a more comprehensive and effective approach to identifying the information structure of sentences. When traditional grammarians aim to distinguish between the NEW information (what has been said) and GIVEN information (about what the NEW was said) by a sentence, they typically focus only on identifying which structural subpart of the sentence identifies the GIVEN information, serving as the topic of the message, and which indicates the NEW information, serving as the comment on the topic identified. For example, in the sentence *The Porsche is fast*, grammarians might identify two separate pieces of information (messages) conveyed by the sentence; the

message [**about** the Porsche] {that it is fast}, and the message [**about** being fast] {that it is a feature of the Porsche}.

According to FTL, the messages expressed with sentences can be more subtle. Firstly, FTL argues that when words are allowed to select meaning from expectation fields, the same word, or phrase, may serve to select both the NEW and the GIVEN information. For example, consider, the sentence *The chess master teaches chess to beginners,* which appeared in a book catalogue. In this context, the phrase *chess master* selects the author of the book as GIVEN and simultaneously assesses the GIVEN, conveying the information that the author is a chess master. In other words, the sentence conveys both the information that the author is a chess master and that he authored the book in which he teaches chess to beginners.

Secondly, individual words themselves, can also be assigned an information structure, which is referred to by the pair of concepts profile (corresponding to NEW) and base (corresponding to GIVEN). In traditional grammars, which do not allow meanings to be selected from the expectation field, each word is associated with one profile and one base. For example, the word *Porsche* conveys the information [a car make] {produced by Porsche} where a [car make] can be considered the GIVEN (or base), and {produced by Porsche} the NEW, or profile.

However, once selective use of words (as well as phrases and sentences) in the expectation field is allowed, words can select not only what is Given and what is New, based on the expectation field, but also allow for multiple divisions of information within a word into the Given and the New. Noting these possibilities allows for a more nuanced understanding of the information structure conveyed by words (as well as phrases and sentences) and of the information conveyed by language.

For example, in the sentence *Jane did not sprain her ankle, she broke it*, the verb *broke* is used to assess the type of injury Jane suffered rather than to inform what event Jane was involved in, what Jane did. To ever need to use this sentence, the speaker must assume that the listener already knows that an accident had happened to Jane and this sentence is providing further details about the type of injury. In contrast, during a phone call, where a mother is enquiring about her children on a summer camp and asks *How are you doing, guys?*, the verb *broke* in the sentence *Jane broke her*

ankle serves to identify a specific event among the events that happened during camp time: the accident involving a broken ankle.

Similarly, the information structure of the meaning of the word *boys* is different, depending on the expectation field generated by different contexts of its use. For example, in the sentence *This competition is for men, not for boys* the word *boy* stands for {young}[male], while in *this school is for boys, not for girls,* it stands for {male}[child].

Finally, by assigning expectation field dependent information structure to demonstrative pronouns, we can resolve the following paradox pointed out by Chierchia (1990): the truth of the sentences *This is big* and *This is a whale* does not necessarily imply the truth of the sentence *This is a big whale*. The sentence *This is big* can also refer to a baby whale. However, if we consider the information structure expressed by this pronoun in different contexts and assign a field-dependent structure to them, we can clarify the selected meanings of *this* used in each sentence in the following way. The meaning of the demonstrative pronoun *this* used in *this is big* can be explicated as {this [object]}, the whole sentence effectively stating that "this object is big". In contrast, *this* in *This whale is big* can be represented as {This [whale]}, the sentence *This whale is big* effectively saying that "this whale is big". From this perspective, it is clear that the demonstrative *this* in *This is big* and in *This is a whale*, respectively, has not been used to assess the same referent, thus we cannot conclude the truth of the sentence *This is a big whale* from the truth of *This is big* and *This is a whale*.

2.3.10] Motivating Language Laws Concerning Counterfactual Time Clauses

After outlining the general characteristics of FTL, we can now motivate the hypothesis that counterfactual time clauses tend to be positioned at the end of a sentence. We can illustrate this trend with the sentence *Mary died before she graduated*, considering the optimal position of the counterfactual clause *before she graduated*. To understand this sentence, we must interpret the main clause *Mary died* literally and infer from the other clause that Mary died **before completing the process leading to graduation**. If we were to place a counterfactual clause *before she graduated* at the beginning of the sentence and thus, first interpret it literally, which would need to include the information that "Mary has graduated", we would need to reinterpret its initial literal meaning after

establishing that Mary died without ever graduating. This is obviously less efficient than starting the interpretation of the sentence with interpreting the factual clause and in that case interpreting the counterfactual time clause only once.

Furthermore it is reasonable to argue that there is no similar reason that would account for a tendency to position factual time clauses last. In other words, counterfactual clauses should be positioned last in sentences relatively more often than factual time clauses. These two hypotheses were confirmed quantitatively using both the British National Corpus and the Polish National Corpus (Zielińska, 2019) and similar investigations can be easily repeated in relation to some future data, by examining corpora yet to be collected.

In summary, both physics and linguistics are equipped to make predictions, and retrodictions, with their respective theories able to offer precise or trend based predictions. The latter is often employed when exploring uncharted domains, when quantitative theories have not yet been formulated, are of stochastic nature, or when collecting data with sufficient precision is not feasible.

Nonetheless, while it is true that quantitative theories are often considered the hallmark of advanced sciences, it is important to acknowledge the value of qualitative research in advancing our understanding of the world. Firstly, qualitative research, often guided by intuition, lays the foundations for any further quantitative investigations by helping to identify the right qualitative assumptions, which are prerequisite for the success of any study. Secondly, significant knowledge about the world can be gained from observations without resorting to the language of mathematics. For example, in the 3rd century Aristotle determined that the Earth is round by observing the shape of its shadow during a lunar eclipse.

It is worth noting, however, that Aristotle's hypothesis about the shape of the Earth was preceded by purely intuitive, qualitative ideas put forward by Pythagoras a century earlier that the Earth is spherical. This purely abstract hypothesis informed further observations and measurements. The next step in our understanding of the shape of the Earth after Aristotle was taken by Eratosthenes, who calculated the circumference of the Earth using measurements of shadows cast by the Sun at distant locations. Thus, the purely conceptual hypothesis of Pythagoras guided others in what could be

observed, which eventually resulted in devising the measurements that could lead to characterizing numerically the qualitative solution found.

2.4] Myth Four: Physics studies classes of identical objects, while humanities are concerned with idiosyncratic ones (such as the speaker's meaning, specific pieces of literary works). Since mathematics can be of value only when describing classes of identical objects—but not of idiosyncratic objects, it can be used only in physics.

Are all objects studied by physics identical and eternal? While physical theories do not distinguish among different electrons, except for their velocity and position, physicists are also concerned with more complex objects such as pieces of rock, hurricanes, and planets, which are so different one from the other that they often get individual names. Furthermore, these objects cannot always be treated as instances of the same category. For example, the models of Mercury or Mars cannot be derived from one general model of a planet as its exemplifications, as they are not simply different members of the same category. Although the models of both are applications of a single theory, they are not contained within it. The description of each of these models involves additionally some peculiar hypothesis concerning shape, density, distribution, orbital motion, and so on.

Moreover, the assumption that all electrons and other elementary particles are identical except for their movement in space is just an assumption. It is an assumption based on our inability to detect any differences, or intentional disregard of them to address the problem at hand with our current tools. This is similar to how linguists postulate the existence of lexemes with their meaning of each of them being specified in dictionaries. Thus, in both physics and linguistics, categories are formed by disregarding individual features of category members in order to explain anything. Without such approximations, if we only focused on individual idiosyncratic instances, we would be unable to make any general statements or apply mathematical description.

Furthermore, the progress of physics began with the fundamental assumption that only some characteristics of a given idiosyncratic object influence its selected feature or a particular aspect of its behaviour. Newton for instance proposed to model the movement of a given object by neglecting all its other characteristics except for

its mass. Entities as diverse as a man, a piece of rock, a star, the Moon, a bee, and a virus, all are subject to Newton's laws because all of them possess "mass" as one of the parameters in their description.

To describe the movements of bodies more precisely, new laws must be introduced that depend on some other characteristics of an object considered, such as its shape. To account for the impact of air resistance, a law of air resistance must be additionally taken into account. If we wanted to consider other differences between the Moon and an apple, apart from their movement, we would need a new law from a different category, which, we will assume to be independent from the law of gravity. To account for more of the individual characteristics of each object, we would be introducing more and more laws, resulting in a progressively more accurate description of their behaviour and characteristics.

In summary, when building models, first we simplify the reality by disregarding many individual characteristics, and start with a very basic representation. Being able to conjecture the essential similarities and disregard incidental differences within a class of objects is a hallmark of scientific enquiry, rather than art. After empirically confirming the validity of the initial assumption, we refine the model, by incorporating more detailed and nuanced aspects of the phenomenon under investigation, dependent on the purpose of the investigation. This way we will acquire successively more accurate and comprehensive understanding of the objects or processes investigated. Therefore, there is no fundamental reason why one cannot eventually construct a model of an individual exemplar within an empirical paradigm.

The reason why mathematics is more commonly applied in physics than in linguistics is simply a matter of practicality. Physics has a long tradition of approximating aspects of physical phenomena using measurable concepts and quantitative theories, which have proven useful in guiding new applications. In contrast, linguists are still in the process of identifying which parameters can be operationalized, developing methods for doing so, proposing and testing relevant quantitative hypotheses.

2.5] Myth Five: Linguists rely on discrete parameters of description, binary classification, while physicists need continuous ones, inherent in advanced mathematics.

In physics, many parameters of description are not binary, but rather continuous. Bunge (1973: 59) pointed out that the progress of the 17th century physics was driven by the realization that differences between individual systems and changes in them cannot be sufficiently described by merely classifying them into binary categories. Instead, continuous variables are needed to capture the nuances of physical phenomena. For instance, in the case of Newton's theory, all parameters except the one identifying the object considered, are continuous. Thus mathematics became essential for handling the resultant variety and complexity. This *novum* allowed for a revolutionary change in the very goal of research, shifting from striving to provide an exact description of perceptible details to discovering universal patterns and creating models that can account for the characteristics and behaviour of the systems modelled.

The empirical sciences took the next revolutionary step in the 19th century, when statistics came into play, building on the use of continuous variables for modelling. (This was already adumbrated when discussing Myth One.) Physics has since continued to advance its theories and models using successively ever more sophisticated tools of mathematical apparatus, which let physicists develop new concepts and eventually lend them to other disciplines. Quantum mechanical formalism, for instance, first developed for physics, has increasingly been applied within a wide range of fields, including economics, artificial intelligence, complex systems science, organisational decision-making, models of the brain and cognition. Even linguistics has been influenced by these developments as researchers such as Peter Bruza, Kirsty Kitto, Douglas Nelson, and Cathy McEvoy (2009), following an early claim by Nelson and McEvoy (2007) suggesting that word associations can display spooky action at a distance behaviour, have shown that quantum mechanical mechanism can model word entanglement in human mental lexicon. The reference to the concept of quantum entanglement has enabled these researchers to reconcile two earlier somewhat contradictory models of word association, the Spreading Activation hypothesis and the spooky-activation-at-a-distance hypothesis, which were capable of modelling only different subsets of data each, arriving at a more complete model. Interestingly, Bruza et al. (2009)

concluded that QM formalism may reflect the entangled nature of the phenomena modelled, rather than merely the characteristics of physical objects of a quantum scale.

In addition to QM, some researchers of the science of language have even adopted partial differential equations to study language. Peter Grzybek (2006) used this formalism to model certain aspects of texts. While the use of QM or partial differential equations to describe linguistic phenomena is rare, the need for another mathematical formalism, statistical analysis of linguistic data, has been widely accepted in psycholinguistic research. In language acquisition studies, statistical analysis is used to predict, for instance, tendencies in the population, such as the decrease in irregular usage of the form "goed" in children with age. (cf. Skousen, 1989).

Finally, it should be noted that the first statistical investigation of linguistic phenomena was carried out by George Zipf in his works from 1832, 1935 and 1949. Zipf's laws are well known, particularly the one that states that the frequency of any word in a text (of a sufficient length, or in a collection of texts) is, roughly, inversely proportional to its rank in the frequency table for that text. For example, in the Brown Corpus, the most frequently occurring word is *the*, which accounts for nearly 7% of all the word tokens there. (69,971 out of slightly over 1 million). The second-place word in the Brown Corpus, *of*, accounts for slightly over 3.5% of words (36,411 occurrences), followed by *and* (28,852). It turns out that only 135 vocabulary items are needed to account for half the Brown Corpus. Since Zipf, many other statistical regularities of the similar type have been discovered (cf. *Journal of Quantitative Linguistics*, Koehler 2012). It is interesting to notice that such power law dependence, as illustrated here by the relationship between the frequencies and the ranks of words in corpora, characterize self-organising systems at large, which we have postulated language to be.

To conclude, it is not accurate to distinguish between sciences and non-sciences based on the use of complex mathematics versus classification. The choice of tools appropriate for a given discipline depends not on its subject matter *per se*, but on the quality and depth of our knowledge of it.

2.6] Myth Six: While the physicist uses objectively measured empirical data to create his theories, the linguist must rely on his intuition to interpret a text.

Another way to express the misconception that linguists rely on intuition, and scientists on objective data is by stating that while sciences deal with quantities, thus with mathematics, humanities focus on qualitative aspects of the phenomena they study. However, such an argument stems from a lack of understanding of the role of mathematics in sciences, which serves as a tool in constructing a theory. Bunge (1973) reminds us that facts are neither mathematical, no anti-mathematical: only ideas can be open to mathematization if they have sufficient clarity and precision. Alternatively, as Altmann (1985) puts it, neither quality nor quantity are inherent characteristics of objects and phenomena, rather, they are parts of concepts that we use to interpret nature.

In other words, when discussing the quantitative aspects of language, the meaningfulness or meaninglessness of quantitative data is not absolute for a given discipline considered, rather it depends on the discipline's models. Therefore, if language is viewed merely as a set of language patterns, as proposed by structuralists, or as algorithms for generating such patterns, which are part of the organism's genetic endowment, as seen by generativists, than quantitative descriptions are of no use. However if language is considered as a self-organising process of language creation that responds to current communicative needs and changing environments, while taking into account previously noted correlations, then the frequency of occurrence of specific patterns realized in the past becomes crucial for deriving "grammar rules".

Thus, the core issue at hand is determining the degree of precision with which we can articulate our intuition about the concepts involved, that is, the extent to which we can reach a consensus when classifying or measuring entities. Traditional sciences are dominated by concepts that are highly measurable, with many derived from intuitive concepts. In linguistics, such precise, measurable concepts are gaining grounds. More and more often, theories are proposed that operationalize intuitive concepts by establishing corresponding measurable equivalents. Two examples of such concepts are introduced below. Further on, they will be used to formulate a linguistic law that can be objectively tested in quantitative terms.

The first concept to be defined is the **sensitivity of the adjective to the noun it modifies**, which reflects our intuition about the range of variability of the meanings of a given adjective depending on the nouns it accompanies. For example, intuitively we agree that a "big" virus differs in size significantly more from a "big" planet than the shade of "blue" of a forget-me-not differs from the shade of "blue" of a blue sky. In other words, the noun sensitivity of the adjective *big* is intuitively higher than that of the adjective *blue*.

A measurable operationalization of the concept of the **sensitivity of the adjective to the noun it modifies** to be introduced stems from the observation that adjectives whose meanings vary significantly when modifying different nouns are more frequently used in comparative and superlative forms than the remaining adjectives. For instance, in linguistic corpora, "*this … is bigger than …,"* is a more frequent comparison than "*this … is redder than…"*. Using this observation, we can operationalize noun sensitivity of an adjective by considering its gradability, which is the ratio of the number of occurrences of a given adjective in its superlative (e.g., *biggest*) or comparative (e.g., *bigger*) forms to its total occurrences (e.g., either *big*, or *bigger*, or *biggest*) in a given linguistic corpus:

$$\text{gradability (big)} = \frac{\#\,\text{bigger} \;+\; \#\,\text{biggest}}{\#\,\text{big} \;+\; \#\,\text{bigger} \;+\; \#\,\text{biggest}}$$

The other linguistic concept, whose operationalization I shall refer to further on when formulating another linguistic law, is **the degree of the adjective's tendency to form situated subcategories**, or for short: adjectives' subcategory forming tendency. A situated subcategory refers to the intuition that certain adjectives used in Adj+Noun phrases affect the referents of the head noun in more ways than simply by stipulating the value of the parameter of the referent of the head noun expressed directly by the given adjective. A good example of a highly subcategory forming adjective is *wooden*. This can be illustrated by the differences between the situated subcategory of *wooden bridges* vs. *steel bridges*, and between *wooden tables* vs. *steel tables*. A wooden bridge and a wooden table differ from steel bridges and steal tables, respectively, not only in the material used to make them (wood vs. steel), but also in their

construction types, likely sizes, and additional materials needed. For example, steel tables often have glass or ceramic tops, while wooden tables are usually all made entirely of wood, except for steel nails. Steel bridges, in turn, tend to be much longer than wooden bridges.

In Polish, we can operationalize the intuition of an adjective's "degree of situated subcategory forming tendency" by examining the semantic impact of the position of adjectives in noun phrases. When placed after nouns, Polish adjectives often indicate a situated subcategory forming property of that adjective. For example, *barszcz czerwony* (*red borscht*) refers to a specific type of soup made primarily of beetroots, which has a somewhat reddish colour, while *barszcz biały* (white borscht) not only has an off-white colour, but most importantly, is made of a different set of ingredients—fermented wheat. So *barszcz czerwony* and *barszcz biały* refer to functionally distinct, situated subcategories of soups, not merely soups of different colours. On the other hand, *czerwony balon* (*a red balloon*) refers to a balloon that differs from a *blue balloon* in colour only, indicating that the adjectives *red* and *blue*, respectively, while prepositioning the noun *balon*, do not single out functionally different subcategories. Therefore, we can quantify the degree of an adjective's tendency to form functionally distinct subcategories (situated subcategories) in Polish by calculating the ratio of the number of its occurrences after nouns in (N+Adj.) phrases, to the number of its total occurrences in noun phrases (N+Adj or Adj+N) in language corpora:

$$\text{subcategory forming tendency (red)} = \frac{\#\,(\text{Noun} + \text{red})}{\#\,(\text{Noun} + \text{red}) + \#\,(\text{red} + \text{Noun})}$$

Based on the two operationalized concepts defined above, we can formulate a quantitative hypothesis about the ordering of adjectives in (Adj_1+Adj_2+Noun) phrases within the Field Model of Language (FTL). As we remember, according to FTL, language self-regulates by interlocutors passively retaining language solutions that optimize cognitive effort involved in communication, because they are easier to remember, recall, more frequently repeated and such.

Therefore, we postulate that the ordering of adjectives in A_1A_2Noun phrases is optimized for cognitive efficiency. Assuming that adjectives in a noun phrase are interpreted starting with the adjective closest to the noun, cognitive efficiency will be increased if we position highly subcategory-forming adjective closest to the noun. The same will be true if we place the most noun sensitive ones the farthest from the noun. This is because, before assessing the parameters of the referent, such as size, colour, value, or opinion, it is good to know the specific characteristics of the situated subcategory the given noun represents. For instance, we can better interpret the size of a *huge building*, if we already know whether this is a *family building* or a *commercial building*.

Therefore, in A_1A_2Noun phrases, where one of the adjectives is subcategory forming and the other is noun sensitive, we should expect to see the trend for noun sensitive adjectives to precede the subcategory-forming ones. This way the listener avoids reinterpreting these noun sensitive adjectives again after interpreting the noun modified by the other subcategory forming adjective. Hence, we typically end up with phrases like *a long wooden bridge* rather than *a wooden long bridge, a huge commercial building* rather than *a commercial huge building, a cute chubby puppy* rather than *a chubby cute puppy, a strong little boy,* and not *a little strong boy, a beautiful French garden* and not *a French beautiful garden.*

Zielińska (2007) used an early version of FTL to demonstrate quantitatively that the postulated tendencies described above hold true for Polish, despite this hypothesis being counter-intuitive for a language with a rich flection and relatively free word order, like Polish, as opposed to English. Unlike English, where the hypothesis of a dominant order of adjectives in A_1A_2N phrases is well known to grammarians and the trend is almost a rule, the Polish version of the hypothesis had not been noticed by Polish grammar books, because this trend is much weaker. Therefore, a quantitative statistical analysis was required to show it[18,19]. Clearly, measurable data

[18] A purely numerical hypothesis of this kind, one considering even more measurable parameters, was confirmed numerically even earlier by Wulff (2003) based on the English National Corpus. Stephanie Wulff, however, was interested only in numerical analysis of her data and did not look for any explanatory theory that could imply the data patterns she found.

[19] Zielińska (2007b) analysed her data statistically by comparing the distribution of semantic categories corresponding to categories of various noun sensitivity and

concerning *situated parole* is needed to note some of the characteristics of *langue*, which are not always binary, but rather of statistical character. And

> Once you begin to look at language from a quantitative point of view, you will detect features and interrelations that can be expressed only by numbers or ranking whatever detail you peer at. There are dependencies of length (or complexity) of syntactic constructions on their frequency, and on their ambiguity, of homonymy of grammatical morphemes on their dispersion on their age, the dynamics of the flow of information on its size, the probability of change of sound on its articulatory difficulty … in short, in every field and on every level of linguistic analysis—lexicon, phonology, morphology, syntax, text structure, semantics, pragmatics, dialectology, language change, psycho- and sociolinguistics, in prose and lyric poetry—phenomena of this kind are predominant. … Moreover it can be shown that these properties of linguistic elements and their interrelations abide by universal laws, which can be formulated in a strict mathematical way in analogy to the laws of the well-known natural sciences. (Altmann & Köhler 2007)

And coming back to the law discussed in this section, the observation about the order of the two classes of adjectives discussed can also be stated in more general terms as two separate laws. The first law: the more sensitive the adjective is to the noun it modifies, the more likely it is to come first in the A_1A_2N phrase. The second law: the more subcategory forming tendency the adjective manifests, the more likely it is to come second in such noun phrases.

2.7] Myth Seven: Unlike in physics, linguistic data is never "pure", and no collection of linguistic data can ever be complete. Therefore, empirical data cannot serve to build a model of language.

One of Chomsky's arguments against using authentic language data, such as language corpora, for language modelling (in McEnery 2003), was that observed language data is never pure. For instance, when uttering a sentence that was later collected in a corpus, the subject may have been under the influence of alcohol, suffered from some sort of memory loss, had a slip of the tongue, or spoke ungrammatically. Moreover, some information in corpus data, such as the

various category forming capacity, but doing it directly with measurable parameters as proposed here would be preferable.

fact that BNC contains more sentences *I live in NY* than *I live in Danton, Ohio*, does not necessarily reflect a linguistic fact. Therefore, our grammatical commentary based on a corpus data rather than intuition may turn out to be a commentary concerning the health condition of a particular speaker, their level of knowledge of the language, or of the reality surrounding them, and not of a language system. As a result, corpus data cannot be relied on when constructing language models.

Yet, in physics there is no "pure" empirical data, either; all data are theory-laden and require interpretation and thus intuition. For instance, to apply Newton's laws to describe the movement of the Moon around the Sun, one must first approximate the Moon as a point in space having mass, next "attach" vectors expressing forces involved and write down relevant mathematical equations. This requires physical intuition that is so distinct from general reasoning that it is quite possible even for mathematicians to lack it. Moreover, when collecting any kind of data, carrying out measurements, we cannot avoid making some errors due to limited precision of instruments used. However, statistical methods can be used to assess the degree of certainty of the answers obtained.

Whether collected data leads to new insights, or simply confirms known knowledge, depends on the nature of the model being tested. For example, if we observe that galaxies are either spiral (clockwise and anticlockwise) or elliptical, additional observations will not enhance our understanding of galaxy types (assuming no new types are discovered). However, if we are studying the model of the universe's creation, which predicts the distributions of clockwise and anticlockwise galaxies, further observations of galaxies can deepen our knowledge.

The situation is analogous in linguistics. For generative grammar models the frequency of a given structure in a corpus is irrelevant. Yet, for models examining the distribution of preferred grammatical structures based on their impact onto optimizing cognitive effort, analysing their statistical distribution, or even discovering that "more of A leads to more (less) of B" can be most significant.

Chomsky's second argument against constructing empirical models of linguistic phenomena using linguistic corpora, as presented in Tony McEnery *et al.* (2006), was based on the impossibility of including all possible sentences in a corpus. In particular,

corpora do not include sentences of infinite length, which are theoretically possible according to generative grammar. Furthermore corpora, tend to lack many grammatically correct but false sentences, and contain few sentences stating obvious truths. As a result, Chomsky concluded that the corpora, being an incomplete source of language data, cannot serve as a basis for constructing a comprehensive model of language[20].

It is true that a corpus cannot determine whether a given sentence is grammatical or not. Yet, empirical data used in physics is never complete in the sense of providing outcomes for all possible situations implied by a given model, either. Even when confirming Newton's Laws of motion, physicists have not tested them for every possible value of every parameter of every specific model. For instance, when modelling a free fall in the gravitational field, they did not test the laws for every conceivable mass and every possible height of the tree the apple could be dropped from. Therefore, there are many potentially true and "grammatical sentences" that have not been observed in physical experiments. Nevertheless, that has not prevented physicists from forming hypotheses that have been confirmed with a high degree of certainty.

Finally, as it is with a linguistic corpus, the collection of physical data also contains a fraction of "ungrammatical" as well "grammatical, but untrue" sentences. After all, everybody makes mistakes and occasionally arrives at incorrect solutions or incorrect interpretations of collected data. Sometimes experimental results are reported, which after repetition turn out to have been wrong. Yet, these untrue statements found in journals of physics, do not discredit model creation based on the experimental data. Just as linguists reject some data as inadequate, so do physicists. Just as linguists extrapolate from actual data collected, so do physicists—the latter with the help of statistics, because as Durka (2003: 13) puts it, "statistics is the art of drawing conclusions from incomplete data. [translation DZ]."

2.8] Myth Eight: It is commonly believed that physical theories can be tested broadly and with great precision, i.e., received physical theories and

[20] Tony McEnery and Andrew Wilson (2003) and Geoffrey Sampson (2001) offered somewhat similar arguments. These books, additionally, provide very interesting arguments against the use of introspection in language model creation.

models give predictions in perfect agreement with experimental results, while—to use Sapir's words—"all grammars leak" (1921: 38).

It is a common misconception that models of physical phenomena perfectly mirror reality, and experimental results align with the predictions of these models simply because these models use mathematical language. While mathematical advances have allowed for the exploration of new ideas in physics, the core of modelling in natural sciences is rooted in appropriate simplification rather than replication of reality.

Physicists do not aim at creating exact copies of objects, systems and processes they study through models, but rather at creating their simplifications. Which characteristics of the phenomena will be included depend on the purpose of the given model. A hunter shooting ducks will need a different model of a duck then a biologist studying its migration patterns.

Scientists may need to simplify their models even further to enable them to solve the equations that constitute them. Another limitation on the precision of viable theories and models arises from the fact that when creating models, it makes sense to include only parameters that can be measured. Further restriction comes from the constrains imposed by the uncertainty introduced by measurements. Finally, we always need to approximate reality in order to study classes of objects and processes so as to be able to draw conclusions of any generality. All these limitations require accepting a more modest goal for models, which is to partially account for observed data rather than provide a perfect match to reality.

To illustrate how much the approximations made due to restrictions on what we can solve can diverge from reality, Bunge (1973) cites studies published in the *Journal of One-Dimensional Physics,* which model 3D (three-dimensional) solid-state objects as if they were 1D (one-dimensional). This is done to propose models based on equations that physicists can solve. One example of such simplified 1D model is Volkenshtein's explanation of the elasticity of macromolecules and the uncoiling of proteins, based on a one-dimensional (Ising's) model of a chain of atoms. However, the problem is that the reality observed is not 1D, which means that a discrepancy between experimental results and theory is unavoidable. Nonetheless, such simplified models can often provide useful insights into the nature and behaviour of actual 3D objects. For

instance, Volkenshtein's model provides a qualitative explanation of the type and direction of changes taking place.

In some situations, as demonstrated by a recent astronomical discovery reported by Kroupa, the potential value of the results of testing even as simple hypothesis as "there are more As than Bs" in a given system, may result in extremely significant insights. In the paper published in *Monthly Notices of the Royal Astronomical Society,* Kroupa (2022) described the recent observation of several star clusters that appear to violate both the law of gravity proposed by Newton and that by Einstein by dissipating in an asymmetric manner. According to each of these theories star clusters are supposed to dissipate symmetrically into two tails with an equal number of stars each. But recent observations of this cluster show that this is not the case. This simple "more As than Bs" observation has now prompted a search for a new theory of gravity to explain the data, as well as indicated a need to revisit alternative propositions. So summing up, models in physics never constitute copies of reality and therefore the results of testing these hypotheses only approximate some characteristics of the phenomena studied, some more exactly, others more crudely. Nonetheless, even the results of the tests of those crude hypotheses can be of immense importance.

Moving on to Sapir's observation that all grammars leak, it is certainly true. All grammars leak and there is a systemic reason for that failure. On the view that language arises in a society and develops through its members' reacting to the properties and requirements of their environment *via* some sort of adaptation mechanisms, a grammar rule understood as the description of a grammatical language structure, can be viewed only as a probabilistic trend in *situated parol*. Since probabilistic laws concerning trends cannot capture individual cases by definition, all grammars, reflecting merely such trends, not only leak, but they must leak. Counterexamples to such laws (leaks) at the level of a single case (the occurrence of some string of words), not only fail to refute such statistical laws, but are expected and can be quantitatively determined.

In summary, it is essential to note that all grammars, regarded as concise descriptions of grammatical language structures, not only leak, but they must leak. Furthermore, seemingly crude laws resulting from counting tokens and analysing their interrelations, such as "more As than Bs", "the more of As, the more/less of Bs",

can yield significant insights when exploring uncharted territories in all empirical sciences, whether that be in empirical linguistics or physics.

3] Conclusions

The opposition of influential academic linguists to researchers adapting the empirical approach in language research may be rooted in the history of mainstream linguists' unsuccessful efforts to identify deterministic, summative laws governing language grammar and meaning in language. Traditional linguists longed for the discovery of language laws akin to Kepler's summary of Tycho Brahe's data on planetary movement around the Sun, not being aware that the grammar of language cannot be condensed into such deterministic rules perfectly because of the nature of its source.

When attempts to find deterministic summation rules in language data fell short of expectations, mainstream linguists wrongly concluded that language cannot be studied within the framework of the empirical paradigm. This belief led to a number of myths that were meant to corroborate this misguided conviction, some of which have been dispelled in this article. However, the existence of a group of linguists, often physicists-turned linguists, who have already been researching language within the empirical paradigm provides perhaps the strongest argument against this misguided conviction. Koehler (2012) presents an overview of over a hundred language laws developed within this paradigm. In this paper, two groups of additional language laws coined within the empirical paradigm were discussed: one concerning the ordering of adjectives in noun phrases (Adj+Adj+Noun), the other concerning the ordering of counterfactual time clauses.

In addition to refuting common misconceptions underlying the belief that language cannot be studied within an empirical paradigm, this paper also outlines the framework enabling such research. To this end, first of all, language must be seen as an aspect of a material system. With our current knowledge of the brain, such as expressed by Jeff Hawkins' model presented in his paper "Computing Like the Brain: The Path to Machine Intelligence" (2013), it is reasonable to assume that language emerges and evolves in a society through the adaptation mechanisms of its members' reacting to the properties and requirements of their environment, as

explicated by Altmann[21] and Koehler (2007). In particular, efficient language solutions (such as frequently needed, shorter, resembling other already well-entrenched items), are retained in memory, resulting in self-regulation of the system, without speakers actively searching for optimal solutions[22]. Given that language is clearly a self-organising self-regulating system, the mechanisms forming language can be guided best by the empirical framework systemism *cum emergentism* explicated in Bunge (2003).

Regarding the studies of meaning within this framework, what needs to be postulated is the mechanism that allows the interlocutor to calculate situated meaning perceived in a specific socio-natural situation (in a situated speech act) at a given stage of interpretation process that may potentially serve as the input for further inferring processes. With systemism *cum emergentism* in mind, constructing the Field Theory of Language, Zielińska (2007, 2019) proposed that situated meaning is the result of interlocutors selecting in the field of their expectations the item(s) matching the closest the encoded content of the words being interpreted. The expectation field reflects the ideas and words that, a moment ago, came to one's mind as likely to be expressed next during the interpretation process. The expectation field is established by taking into account such factors as the information about the social situation involved (situated speech act), including its purpose and environmental constrains, information comprehended verbally so far in the given verbal encounter, the encoded contents of the items being interpreted, and associations formed on the way. All this information is filtered by interlocutor's knowledge, experience, biases, interests, current attention focus and similar relevant factors[23]. Each option in the field is assigned a likelihood of being intended. "Efficient situated meanings", as defined above, are stored in memory, building and regulating idiosyncratic languages. Statistical trends in such

[21] This idea was expressed already in Altmann (1978), albeit in a more general manner.

[22] This assumption crucially distinguishes current Neozipfian approaches to describing the mechanism of language self-regulation, from Zipf's Principle of Least effort, which posits that speakers consciously search for optimal language solutions when speaking.

[23] This parallels recent models of visual perception. Perceiving visually is the result of processing the stimuli received half a second earlier to calculate the present moment of perception of the surroundings using our models coined based on our prior experience.

individual languages correspond to *langue*—language of the community seen as an abstract structure. It is also worth noting that the meaning of a word, or of some other stable unit in language, is stored in the brain not only along with its form but also with the contextual information it has been correlated with, both verbal and non-verbal.

The assumptions outlined above provide a foundation for studying both the quantitative and qualitative aspects of language within the empirical paradigm, leading to valuable insights of both kinds. When examining quantitative aspects of language, viewed as a dynamic, non-linear system comprised of situated utterances subject to evolutionary processes, many of its characteristics can only be captured quantitatively as probabilistic trends in the interrelations between tokens under examination.

The quantitative laws of language, first explored by Zipf, encompass dependencies such as the relationship between word length and rank, the complexity of linguistic constructions and their frequency, or the dependence of the dynamics of the flow of information on its size. It has been observed that many of these dependencies follow power-law distributions, a characteristic common to other self-organising systems. Moreover, we can study quantitatively correlations among the sets of words correlated with words, specifically examine correlations between words' verbal contexts, thus gaining relational insight into the meanings of these words. By the way, it is also worth noting that it was over a century ago when Firth made the observation that "you shall know the word by the company it keeps," heralding the relational approach to the study of meaning.

Another approach to studying language in the empirical paradigm that is practiced today involves testing quantitative hypothesis implied by qualitative theories of language understanding and processing. This can be done by analysing language corpora or considering the characteristics of physical responses accompanying verbal interactions, such as data resulting from measuring reaction times, recording eye tracking, or monitoring brain activities. Although such hypotheses are often crude, for instance stating "the more of As, the more of Bs", their test results also help gain valuable insight into understanding language, for instance, validate qualitative assumptions made.

All in all, it is clear that in language studies, as in other empirical disciplines, quantitative research is possible and complements qualitative studies. Quantitative results can be used to fine-tune language characteristics hypothesized qualitatively, to draw new hypothesis suggested by the observations, or to rigorously test qualitative assumptions made, among others. However, making effective qualitative assumptions, such as shifting from viewing language as a self-standing structure to seeing language as a self-organising and self-regulating system, selecting the appropriate operationalization of concepts, or utilizing the quantitative information that can be measured, is critical to the quality and significance of all insights, including those gained quantitatively.

For example, postulating that language is an aspect of a material system that has a self-organizing and self-regulating mechanism provides the source of Zipf laws, elevating them from mere trivia to constituting the central argument for language being a self-organizing and self-regulating system. In turn, proposing additionally the qualitative mechanism of interpreting meaning in the expectation field, which uncouples encoded meaning of words from their selected (situated) meanings, allowed us to gain, among others, the following novel insights into language.

Firstly, it allowed for an explanation of the emergence of novel meaning in language. This is crucial for elucidating the meaning of words used in specific situations, accounting for the compositionality of meaning, the self-regulation of meaning in idiosyncratic languages, and ultimately in a community language. Secondly, it offers a more comprehensive account of the messages that can be conveyed with the same sentence in different situations, going beyond what the traditional division into sentence comment and topic (the NEW and the GIVEN) can do. According to the view advocated, any part of a sentence may contribute to identifying in the expectation field of the interlocutor the non-encoded topic and/or the comment (or both), resulting in the possibility of the sentence selecting a much larger number of messages than what the traditional division of a sentence into the GIVEN and the NEW allows one to account for. It also provides the explanation for the observations that the same structurally and lexically unambiguous sentence used with a different purpose, (with different information structure) may have different representations and therefore, even different truth values. The reason is that since the expectation field postulated by FTL

categorization mechanism evolves during the interpretation process, therefore when the parts of the given sentence are interpreted in different order (which is the case when different elements are treated as the GIVEN), the final interpretations of that sentence may differ from each other. Last but not least, it was demonstrated that FTL can serve as a source of semantically motivated quantitative language laws.

Acknowledgments

I would like to express my deep gratitude to a number of people whose expertise and integrity have played a significant role in shaping my research. To late prof. Jacob May, the founder of linguistic pragmatics, the founder and Editor-in-Chief of the *Journal of Pragmatics*, for inspiration, as well as editing Zielińska (2007a), the article introducing an early version of FTL, and even adding his own comments in footnotes. Professor May connected me, in turn, with prof. Alessandro Capone of Messina University, whose assistance has helped me in more ways I could enumerate here. In particular, I am grateful for inviting me to the Editorial Board of the book series *Pragmatics Philosophy and Psychology*.

Next, I would like to thank Dr. Martin Benes from Charles University and Czech Academy of Science, who pleasantly surprised me by translating my book into Czech (Zielińska 2007b) and also incorporated my theory of language into in his own research, which gave me special encouragement to continue when I needed it most. I am also immensely indebted to Professor Helene Włodarczyk and Professor Andre Włodarczyk, respectively, emeritus professors of Charles de Gaulle University emeritus professors of Sorbonne, for the inspiration I gained from their linguistic research clearly impacted by their immensely broad interests ranging from semiology of prehistoric rock art, through Japanese poetry, to cybernetics, as well as for noting my own theory of language and inviting me to present it at the University of Sorbonne.

Last but not least, I extend my special gratitude to Francois Maurice, the founder of the Society for the Progress of Metasciences and this journal, for his helpful comments and editorial patience in whose research Mario Bunge holds a special place and to Professor Teresa Grabinska, a theoretical physicist and philosopher of science, the recipient of the Copernican Award, for her continual encouragement and

inspiration. It was when reading her books that I learned of Mario Bunge's research for the first time.

References

Altmann, G. (1985), « Sprachtheorie und mathematische Modelle », *SAIS Arbeitsberichte aus dem Seminar für allgemeine und indogermanische Sprachwissenschaft*, Vol. 8, p. 1-13.

Altmann G. & Koch W.A. (1998), *Systems: New Paradigms for the Human Sciences*, De Gruyter

Altmann G. & Köhler R. (2007), « Chapter 2: Quantitative Linguistics: An Overview », *in* S. Kepser & M. Reis (eds.), *Linguistic Evidence - Empirical, Theoretical, and Computational Perspectives*, De Gruyter.

Bruza P., Kitto K., Nelson D., & McEvoy C. (2009), « Extracting Spooky-Activation-at-a-Distance from Considerations of Entanglement », *in* P. Bruza, S. Sofge, W. Lawless, K. van Rijsbergen & M. Klush, *Quantum Interaction: Third International Symposium, QI 2009, Saarbrücken, Germany, March 25-27, 2009. Proceedings 3*, Springer, p. 71-83.

Bunge M. (1973), *Method, Model and Matter*, Reidel.

Bunge M. (1959), *Causality: The Place of the Causal Principles in Modern Science*, Harvard University Press.

Bunge M. (2003), *Emergence and Convergence: Qualitative Novelty and the Unity of Knowledge*, University of Toronto Press.

Chierchia, G. & McConnel, S. (1990), Meaning and Grammar - 2nd Edition: An Introduction to Semantics.

Dummett M. (1993), *The Seas of Language*. Clarendon Press.

Durka P. J. (2003), *Wstęp do współczesnej statystyki*, Adamantan.

Grzybek P. (2006), *Contributions to the Science of Text and Language, Word Length Studies and Related Issues*, Springer.

Jary M. (2008), « The Relevance of Complement Choice: A Corpus Study of 'believe' », *Lingua*, 118, p. 1–18.

Köhler R. & Altmann G. (1996), « "Language Forces" and Synergetic Modelling of Language Phenomena », *in* P. Schmidt (ed.), *Glottometrika 15: Issues in General Linguistic Theory and the Theory of Word Length*, WVT.

Köhler R. & Altmann G. (2005), « Aims and Methods of Quanitative Linguistics », *in* G. Altmann, V. Levickij & V. Perebyjnis (eds.), *Problemy kvantytatyvnoïlingvistyky/Problems of Quantitative Linguistics*, Ruta, p. 12–41.

Köhler R. (2012), *Quantitative Syntax Analysis*, De Gruyter.

Köhler R. & Altmann G., « Introduction to Quantitative Linguistics » (private communication).

Kroupa P., Jerabkova T., Thies I., Pflamm-Altenburg J., Famaey B., Boffin H.M. J., Dabringhausen J., Beccari G, Prusti T., Boily C., Haghi H., Wu X., Haas J., Zonoozi A.H., Thomas G., Šubr L. & Aarseth S.J. (2022), « Asymmetrical Tidal Tails of Open Star Clusters: Stars Crossing Their Cluster's Práh Challenge Newtonian Gravitation », *Monthly Notices of the Royal Astronomical Society*, 517(3), p. 3613–3639.

Lakoff G. (1987), *Woman Fire and Dangerous Things*, Chicago University Press.

McEnery T. & Wilson A. (2003), *Corpus Linguistics. An Introduction*, Edinburgh University Press.

McEnery T., Xiao R. & Tono Y. (2006), *Corpus-Based Language Studies An Advanced Resource Book*, Routledge.

Nelson D. & McEvoy C. (2007), « Is There Something Quantum-Like About the Human Mental Lexicon? », *Journal of Mathematical Psychology*, 53, p. 362-377.

Sampson G. (2001), *Empirical Linguistics*, Continuum.

Sapir E. (1921), *Language: An Introduction to the Study of Speech*, Harcourt Brace**.**

Searl J.R. (1983), *Intentionality: An Essay in the Philosophy of Mind*, Cambridge University Press.

Siegel E. (2022), « We exist. What can that fact teach us about the Universe? », *Starts With a Bang!* https://medium.com/starts-with-a-bang/we-exist-what-can-that-fact-teach-us-about-the-universe-fdae9463a996

Skousen R. (1989), *Analogical Modeling of Language*, Kluwer.

Snow C. P. (2001) [1959], *The Two Cultures*, Cambridge University Press.

Ulam S.M. (1991), *Adventures of a Mathematician*, University of California Press.

Wheeler J.A. (1994), *At Home in the Universe*, Woodbury, American Institute of Physics.

Wulff S. (2003), « A Multifactorial Corpus Analysis of Adjective Order in English », *International Journal of Corpus Linguistics* 8(2), p. 245-282.

Xiang M. (2017), « Toward a Neo-Economy Principle in Pragmatics », *Journal of Pragmatics*, 107, p. 31-45.

Zipf G.K. (1932), *Selected Studies of the Principle of Relative Frequency in Language*, Addison-Wesley.

Zipf G.K. (1935), *The Psycho-Biology of Language*, Houghton-Mifflin.

Zipf G.K. (1949), *Human Behavior and the Principle of Least Effort*, Addison-Wesley.

Zielińska D. (1999), « The Selective Mode of Language Use – The Way Natural Language Adapted Itself to Describing the World Around Us », *Zeszyty Naukowe UJ Prace Językoznawcze*, 119, p. 173-176.

Zielińska D. (2003), « On the Selective Mode of Language Use », *Biuletyn Polskiego Towarzystwa Językoznawczego*, LIX, p. 27-35.

Zielińska D. (2007a), « The Selective Mode of Language Use and the Quantized Communicative Field », *Journal of Pragmatics*, 39, p. 813-830.

Zielińska D. (2007b), *Proceduralny model języka: Proceduralny model języka: Językoznawstwo z pozycji teorii modeli nauk empirycznych*, Wydawnictwo Uniwersytetu Jagiellońskiego[24].

Zielińska D. (2014), *Procedurální model jazyka: Lingvistika z pohledu teorie modelu empirickych ved*, Edice Qfwfq.

[24] Since Zielińska (2007b) is out of print, therefore I provide here its Czech translation available for free as PDF. Zielińska, D. (2014), Procedurální model jazyka Lingvistika z pohledu teorie modelu empirickych ved:

https://oltk.upol.cz/fileadmin/userdata/FF/katedry/kol/publikace/publ_qfwfq/Zielińska-Proceduralni_model_jazyka.pdf

https://oltk.upol.cz/fileadmin/userdata/FF/katedry/kol/publikace/publ_qfwfq/Zielińska-Proceduralni_model_jazyka.pdf

Zielińska D. (2019), « The Field Model of Language and Free Enrichment", *in* A. Capone, M. Carapezza & F. Lo Piparo (eds), *Further Advances in Pragmatics and Philosophy: Part 2 Theories and Applications. Perspectives in Pragmatics, Philosophy & Psychology*, vol 20, Springer, p. 239-249.

Zielińska D. (2020a), « How does language work? », (manuscript), https://www.researchgate.net/publication/344239308_3_How_does_language_work

Zielińska D. (2020b), Testing the Advocated Theory of Language. The Study of the Order of Adjectives in Noun Clauses and of the Order of Counterfactual Before, https://www.academia.edu/84422633/4_Testing_the_advocated_theory_of_language

COPYRIGHT AND DISTRIBUTION POLICY