EKATERINE MARUASHVILI

Associated Professor LEPL Georgian State Teaching University of Physical Education and Sport (Georgia)

STATISTICAL ANALYSES IN LANGUAGE USAGE

DOI: https://doi.org/10.52340/isj.2024.28.06

Language has a fundamental social function, it is a widely used mean of communication, dynamic, robust and still so simple; a specific human capacity, capable of carrying our thoughts and maybe the only feature that make us humans fundamentally different from other species, and still so vaguely understood. Approximately from 3000 to 7000 languages are spoken nowadays, all of them hold remarkable distinctions one from another, but still have much in common. Recent research on cognitive sciences have concluded that patterns of use strongly affect how language is perceived, acquired, used and changes over time. It is argued that languages are self-organizing systems, and that language usage creates and shapes what languages are. The linguistic competence of a speaker is attributed to self-organization phenomena, but not to a nativist hypothesis. The purpose of this study is to develop statistical analyses of language usage based on a detailed investigation of the Zipf's law and other laws of quantitative linguistics. We will develop a systematic empirical investigation of phenomena via statistical, mathematical and computational techniques. We carry out, first, a horizontal analysis across different languages using the UCLA Phonological Segment Inventory Database. This analysis is followed by a vertical investigation of English patterns in different linguistic structural levels. In addition to the results obtained with Zipf's law, information theoretical analyses are done in order to understand the trade-off between the efficiency of language information transmission and language complexity. We observe that the features of linguistic elements and their interrelations abide by universal laws (in the stochastic sense). These analyses are important for a quantitative comprehension of linguistic concepts that are already well known qualitatively, providing a means to understand the processes underlying language usage and evolution. Understanding how languages work and evolves might be the only hope to create technological artifacts that truly exhibit human-level communication capabilities, being able to understand and produce human-like sentences/ utterances.

Language is a biological, psychological and social process. The study of language as a communication process involves insight on these subjects and a scientific analysis of data produced as a mean of information transfer. Performing a statistical analysis of language is a way of acknowledging its unpredictable nature, as the uncertainty intrinsic to it is the way in which it is possible to carry information. Although language has a random nature, it holds an order, coordination and structuration that imposes an amount of redundancy to the transmitted message. It is important to characterize the process and understand what variables are into play in the communication process. Language is not a process controlled by a single agent, rather it is driven by interactions of multiple agents, it is wholly decentralized or distributed over all the components of the system. All languages attain such characteristics and therefore it is important to analyze languages from this common ground and try to understand, based on the common patterns observed in languages, how languages work. We need then to change our paradigm of 'linguistic universals'. As we might observe, language speech inventories are quite diverse and there are vanishingly few linguistic universals in direct sense left. On the

other hand, as we regard language as a adaptive complex system, we shall observe that there are patterns in language that are also usual in natural phenomena. The ubiquity of power laws is the most notorious one and for that reason we will deeply investigate the well know Zipf's law.

In this study the focus will be on a statistical analyses of language usage, performing a detailed investigation of quantitative linguistics laws. The approach chosen consist on first analyze different languages and then perform a deep examination on English patterns. The communication process is observed under the information theoretical point of view to understand the relation existing between the efficiency in information transfer and the complexity of the system. These analyses are important to comprehend how the language communication phenomenon works and correlate the findings with the well-known linguistic concepts. The analysis of language as a complex system is radically different from the traditional analysis based on a static system of grammatical principles, as a result of the generativist approach. This new approach to language is important for it may allow a unified understanding of seemingly unrelated linguistic phenomena, such as: "variation at all levels of linguistic organization; the probabilistic nature of linguistic behavior; continuous change within agents and across speech communities; the emergence of grammatical regularities from the interaction of agents in language use; and stage like transitions due to underlying nonlinear processes" (Beckner et al., 2009). The language patterns are important for language usage, acquisition and efficiency. A well-known example is the word frequency effect on lexical access (Whaley, 1978; Grainger, 1990; Andrews, 1989). Low frequency words require greater effort than high frequency words on recognition task, leading to a poorer performance on speed and accuracy tests. Words might be ranked in order of their frequencies of occurrence and that leads to the observation of a power law relation between word rank and frequency. Length of words is also not a mere hazard but a rational deliberation aiming a thrifty and efficient use of resources in a communication process. The way a language sound system is organized seeks a maximal dissimilarity between stimuli. This is an important choice in order to convey maximal information transfer between speaker and listener in a noisy environment. In this huge universe of multiple possible combination of structures, we believe the formation of a language is guided by choices, which organize and structure the random process of communication. Languages are complex systems whose emergence is an event of central importance to human evolution. Several remarkable features suggest the presence of a fundamental principle of organization that seems to be common among all languages.

In this dichotomy of 'language as chance' - 'language as choice', applying quantitative methods are fundamental to let us draw insights on nature of this communication phenomenon. This dichotomy, rightly understood, might appear as the bridge between the two dichotomies proposed by Saussure: 'langue-parole' and 'significant-signifies. "In fact, the relation is quite close: language as change refers to the langue-parole dichotomy in its interpretation as that between statistical universe and sample, whereas language as

"If a statistical test cannot distinguish rational from random behavior, clearly it cannot be used to prove that the behavior is rational. But, conversely, neither can it be used to prove that the behavior is random. The argument marches neither forward nor backward" (Miller, 1965). Contrary to Miller's belief, we argue that a statistical characterization of language as a communication process is of central importance to trace the line that distinguishes a mere random event from another, also random in nature, but that stands in the watershed between chaos and order, establishing a balance between information transfer and communication cost. The idea of statistical treatment of language data is not new, and we might even say that linguistics is not possible without some degree of statistical classification. Linguists have

always used patient recording, annotations and classifications in order to imagine what would be a possible grammar for that language. Moreover, a regularity in the historic observation of language, like the Grimm's law consonantal shift, could only be realized after an investigation on a long and patient collection of data. Comparative philology also uses the comparison of a great mass of linguistic data to establish the relationships among languages and families.

"The effectiveness of language as a means of communication depends, naturally, on its being highly patterned, and hence on its users' behavior being predictable, not necessarily as to the meanings they will convey in each individual situation, but as to the phonological, morphological, and syntactical paths they will follow in so doing. Yet no set of speech-habits is entirely rigid or ultra systematic... There are always loose ends within the system of speech behavior. It is this inherent looseness of linguistic patterning, together with built-in redundancy, that makes change not only normal but inevitable, and thus a basic part of language. The great mistake of the idealists (determinists) is their overemphasis on vocabulary choice as the only source of linguistic change, and their consequent neglect of the habitual aspects of language. Our linguistic behavior is very largely a matter of habit, and, in Twaddell's words, 'below and above the control of the individual' – below because it is so largely unreflecting habit in brain, nerve, and muscle; above, because it is so largely influenced, from its very inception in each of us, by the behavior of other members of the community.

Each individual builds up his own set of speech-habits, his idiolect, in himself, and of course the idiolect is the only ultimate linguistic reality. Entities such as 'dialect' or 'languages' are always abstractions formed on the basis of a comparison of two or more idiolects... Yet this does not mean that each individual 'creates' his language ex novo; virtually all our speech-habits are built up through imitation of those of other individuals, and what little is 'original' with each speaker derives from combination of already existing patterns. An idiolect is effective as a means of communication only because it closely resembles the idiolects of other speakers. There is never an absolute identity between any two idiolects, but there can be a very close similarity which justifies our abstracting (naively or analytically) what is common to them and treating it as an entity. Each language, each dialect has its phonemic structure, and only what is within that structure is possible for the speaker and listener of the language or dialect. And within the limits of structure imposed by the community, the individual speaker makes his choices... He sees his choices as free and... comes to ignore the limitations and move about them comfortably, so that the real choices become the only choices he sees" (Hall, 1964). In order to capture language as an emergent identity on the vast universe of idiolects and spoken realizations, it is important to observe the recurring patterns on a large dataset and extract linguistic meaning from it. The quantitative analysis of languages is important to produce a systematic empirical investigation of the language phenomenon via statistical, mathematical or computational techniques. It is grounded on a large data of empirical observations, which are used to develop and employ mathematical models, theories and hypothesis pertaining the phenomenon.

The quantitative approach to language analysis data back to the ancient Greek who have used combinatorics to investigate the formation of linguistic structures. Later, the philologist and lexicographer Al-Khalil ibn Ahmad (718-791) used permutations and combinations to list all possible Arabic words with and without vowels. William Bathe (1564-1614) published the world's first language teaching texts, called 'Janua Linguarum', where he had compiled a list with 5.300 essential words, according to their usage. From the end of the 19th century many scientific works on language started using the quantitative approach. Augustus De Morgan

(1851), for example, on the statistical analysis of literary style, suggested that one could identify an author by the average length of his words. Many scientific counts of units of language or text were published in the 19th century as a means of linguistic description: in Germany, F¨orstemann (1846, 1852) and Drobisch (1866); in Russia, Bunjakovskij (1847); in France, Bourdon (1892); in Italy, Mariotti (1880); and in the USA, Sherman (1888). From the 20th century on, many scientific works have been produced on quantitative linguistics.

The linguistic analysis of a language is the observation of certain recurring patterns, their transformation over time and interactions. Patterns that occur systematically across natural languages are called linguistic universals. An important goal of linguistics is to explain the reason why these patterns, emerge so often, which is also a concern of cognitive studies. Some approaches might be used to carry out systematic research and to analyze the role of these regularities on languages. We are here concerned with a statistical analysis based on real world data, through the usage of linguistic corpora, and with computer simulations of models mimicking language interactions.

We know that speech sounds used in spoken communication vary from one language to the other. We propose to perform a statical analysis of the speech inventories used in different languages. For this purpose, we will use the UCLA Phonological Segment Inventory Database (UPSID) which has 451 languages in its database. We will observe the different speech inventories used and their characteristics. Among these various languages, we will observe that some speech sounds are very common while others are quite rare. All these analyses presupposes that a speech utterance might be segmented into distinctive speech segments, phones. The UPSID has a detailed description of the phones used in each language and much information might be extracted by means of this database. It is still unclear what is the nature of the language constituent elements, how they are used and organized, and how they change over time. The phoneme, taken as a mental representation, the basic element of spoken language, has been questioned over its status on the study of language. Port (2007) argues that “words are not stored in memory in a way that resembles the abstract, phonological code used by alphabetical orthographies or by linguistic analysis”. According to him, the linguistic memory works as an exemplar memory, where the information stored is an amalgam of auditory codes which include nonlinguistic information. The acceptance and usage of the phonetic model is a reflex of our literacy education (Port, 2007; Coleman, 2002). The assumption of a segmental description of speech is also desired since it guarantees a discrete description at the lower level, what implies discreteness at all other levels. All formal linguistics is based on one a priori alphabet of discrete tokens. There are many interactions between speech and writing. Lev S. Vygotsky was a psychologist who took an active interest in the cognitive consequences of writing, studying how speech affected writing and vice versa. “Writing requires deliberate analytic action on the part of the speaker1. In speaking, he is hardly conscious of the sounds he produces and quite unconscious of the mental operations he performs. In writing, he must take cognizance of the sound structure of each word, dissect it, and reproduce it in alphabetic symbols, which he must have studied and memorized before” (Vygotsky, 1934). The relationship between writing systems and spoken language is also a theme covered by Coulmas (2003). According to him, “the introduction of writing implies a cognitive reorientation and a restructuring of symbolic behavior. Names of objects are conceptually dissociated from their denotata, as signs of physical objects are reinterpreted as signs of linguistic objects, names. In a second step, signs of names are recognized as potentially meaningless signs of bits of sound, which are then broken down into smaller components” (Coulmas, 2003).

Considering words as unities of mental

processing, it is important to investigate the aspects involving this hypothesis. Miller (1956) suggested that the short-term memory storage capacity is constant in terms of the number of chunks. If we could consider words as chunks, then the short-term memory capacity should be the same regarding the size or duration of words. Baddeley et al. (1975) explores the relations between the memory span and length of words. They observed that memory span is inversely proportional to word's length. Word's duration was recognized as an important aspect, since it was recognized that words of short temporal duration were better recalled than words of long duration, even when the number of syllables and phonemes are held constant. The results achieved by Baddeley et al. (1975) have some implications on Miller (1956)'s suggestions, "that memory span is limited in terms of number of chunks of information, rather than their duration. It suggests a limit to the generality of the phenomenon which Miller discusses, but does not, of course, completely negate it. The question remains as to how much of data subsumed under Miller's original generalization can be accounted for in terms of temporal rather than structural limitations" (Baddeley et al., 1975).

In this study we have focused on applying a statistical analysis on language use data, in special investigating some already know quantitative linguistic laws. Initially we have performed a horizontal analysis across different languages, using the UPSID. Afterwards we have attained our attention to the patterns of use of only one language: English, performing the in multiple linguistic levels. We have also the classical Information Theory to carry out a systematic inquiry to examine language under this perspective, in order to understand the trade-off between efficiency in information transmission and complexity of the system. These analyses have shown important to achieve a quantitative comprehension of linguistic concepts which are already well known and described in the literature. Such study is important to understand the processes underlying language use and evolution, what is necessary to create a better model that might be applied to create technological artifacts that truly exhibit human-level communication capabilities.

In order to better understand the role played by these aspects into the way a language is structured, organized and used, we propose here a statistical analysis using a corpus. It would be time-spending and would require a great amount of work to collect a speech corpus and make use of it. Instead, we propose the usage of a text corpus, pronunciation dictionary and speech samples provided by online dictionaries. The analysis here will concern only the statistical aspects of written and spoken words length, what is important as length is regarded as an aspect of mental representation, among other features (Port,2007). Mendenhall realized that the study of word length, specifically, the analysis of the distribution of words of different lengths was important to establish comparisons of styles. Mendenhall (1887) investigated the differences in the literary styles of Dickens and Thackeray insofar as word-length distribution was concerned. The same approach was afterwards used (Mendenhall, 1901) to analyze the authorship of Shakespeare's plays. In count of words of length three. Comparing with Bacon, the count of words of length three was greater than four, and Bacon also present a distinctly higher proportion of longer words than Shakespeare.

.

Bibliography:

[1]. Manin, D. Y. (2008). Zipf's law and avoidance of excessive synonymy. Cognitive Science: A Multidisciplinary Journal, 32(7):1075–1098.

[2].Manin, D. Y. (2009). Mandelbrot's model for Zipf's law: Can Mandelbrot's model explain Zipf's

law for language? Journal of Quantitative Linguistics, 16(3):274–285.
[3].Odden, D. (2005). Introducing Phonology. Cambridge University Press
[4]. Port, R. (2006). Second Language Speech Learning: The Role of Language Experience in Speech Perception and Production, chapter the graphical basis of phones and phonemes. John Benjamins, Amsterdam.

ЕКАТЕРИНА МАРУАШВИЛИ

Доцент LEPL Грузинский государственный педагогический университет физического воспитания и спорта (Грузия)

СТАТИСТИЧЕСКИЙ АНАЛИЗ ИСПОЛЬЗОВАНИЯ ЯЗЫКА

Резюме

В этом исследовании мы сосредоточились на применении статистического анализа данных об использовании языка, в частности, на исследовании некоторых уже известных количественных лингвистических законов. Сначала мы провели горизонтальный анализ по разным языкам, используя UPSID. После этого мы обратили внимание на закономерности использования только одного языка: английского, выполняя на нескольких языковых уровнях. У нас также есть классическая теория информации для проведения систематического исследования языка с этой точки зрения, чтобы понять компромисс между эффективностью передачи информации и сложностью системы. Эти анализы показали важность достижения количественного понимания лингвистических концепций, которые уже хорошо известны и описаны в литературе. Такое исследование важно для понимания процессов, лежащих в основе использования и эволюции языка, что необходимо для создания лучшей модели, которая может быть применена для создания технологических артефактов, которые действительно демонстрируют возможности общения на уровне человека.

Чтобы лучше понять роль, которую играют эти аспекты в том, как язык структурирован, организован и используется, мы предлагаем здесь статистический анализ с использованием корпуса. Это заняло бы много времени и потребовало бы большого объема работы, чтобы собрать речевой корпус и использовать его. Вместо этого мы предлагаем использовать текстовый корпус, словарь произношения и образцы речи, предоставленные онлайн-словарями. Анализ здесь будет касаться только статистических аспектов длины написанных и произнесенных слов, что важно, поскольку длина рассматривается как аспект ментального представления, среди прочих особенностей (Port,2007). Менденхолл понял, что изучение длины слова, в частности, анализ распределения слов разной длины, важно для установления сравнений стилей. Менденхолл (1887) исследовал различия в литературных стилях Диккенса и Теккерея в той мере, в какой это касалось распределения длины слова. Тот же подход впоследствии использовался (Mendenhall, 1901) для анализа авторства пьес Шекспира. По подсчету слов длиной три. По сравнению с Бэконом, количество слов длиной в три слова было больше, чем у четырех, и у Бэкона также было заметно большее количество длинных слов, чем у Шекспира.