Patterns in Language Usage

Leonardo Araujo
supervisors: Hani Yehia and Thaïs Cristófaro-Silva
UFMG

October 13, 2013

Outline

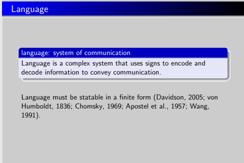
1. Introduction
   - Language
   - Quantitative Linguistics
   - Structuralism
   - Language Structure
2. Interlanguage Statistical Analysis
3. Intralanguage Statistical Analysis

Language

language: system of communication

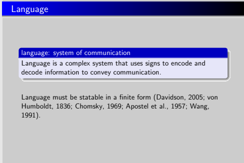
Language is a complex system that uses signs to encode and decode information to convey communication.

Language must be statable in a finite form (Davidson, 2005; von Humboldt, 1836; Chomsky, 1969; Apostel et al., 1957; Wang, 1991).

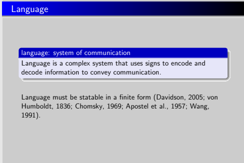


Language is a human capacity for acquiring and using complex systems of communication. A language in this sense is a system of signs for encoding and decoding information. Language is a social process of human iteration, an then it is made possible by our biological capabilities (and so limited by its biological restrictions) and shaped by our psychological aspects.

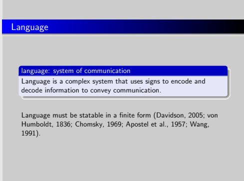
Language has a fundamentally social function. Processes of human interaction along with domain-general cognitive processes shape the structure and knowledge of language.



"Along with its social function, language is importat to humans as a mental instrument. Indeed, the invention of language – that is, the accumulation of symbols to represent emotions, objects, and acts – may be the most important event in human evolution, because so many developments follow from it"(Freedman and Wang, 1996).

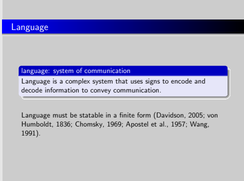


Recent research across a variety of disciplines in the cognitive sciences has demonstrated that patterns of use strongly affect how language is acquired, is structured, is organized in cognition, and changes over time. However, there is mounting evidence that processes of language acquisition, use and change are not independent from one another but are facets of the same system.



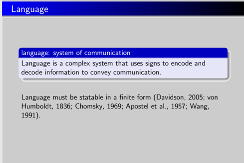
Is there Thought without Language?

"Thoughts are forms conceived in the mind, rather than the forms perceived through the five senses. Thought and thinking are the processes by which these concepts are perceived and manipulated. Thinking allows beings to model the world and to represent it according to their objectives, plans, ends and desires.".



In 1967 Donald Davidson (philosopher) published "Truth and Meaning", in which he argued that any learnable language must be statable in a finite form, even if it is capable of a theoretically infinite number of expressions – as we may assume that natural human languages are, at least in principle. If it could not be stated in a finite way then it could not be learned through a finite, empirical method such as the way humans learn their languages.

We are finite beings whose mastery of the indefinitely many expressions of our language must somehow arise out of our mastery of finite resources. Otherwise, there would be an unbounded number of distinct things to learn in learning a language, which would make language learning impossible for finite beings like ourselves. The linguistic competence of a finite being of our sort must be the result of the interaction of a finite number of basic competencies.



Ever since Humboldt (1836/1999), researchers have hypothesized that language makes "infinite use of finite means." Yet the study of language had to wait nearly a century before the technical devices for adequately expressing the unboundedness of language became available through the development of recursion theory in the foundations of mathematics (cf. Chomsky, 1965). Recursion has subsequently become a fundamental property of grammar, permitting a finite set of rules and principles to process and produce an infinite number of expressions.

"Segmental phonology and hierarchic syntax are of partcular importance" Wang (1991).

Language

language: system of communication
Language is a complex system that uses signs to encode and decode information to convey communication.

Language must be statable in a finite form (Davidson, 2005; von Humboldt, 1836; Chomsky, 1969; Apostel et al., 1957; Wang, 1991).

Apostel's article: "Logique et langage considérés du point de vue de la précorrection des erreurs". Trubetskoy attempted to characterize phonemic systems as maximally 'coherent' the more they consist of 'bilateral,homogeneous, privative, proportional, and neutralizable' oppositions. We might understand as a branching-diagrams of phonemic oppositions which is characterized by the binarity of the nodes, symmetry of the branches, and parsimony in the use of phonetic features. Thus, to require a 'coherent' system to contain 'privative' oppositions, as opposed to 'gradual' and 'equipollent' ones (which is to say that each pair of opposed phonemes must share at least one feature and differ in at least one feature, while no more than two phonemes which share some set of features can differ by any one feature), is simply to require that all nodes, or branching-points, in the branching-diagram of the phonemes by their phonetic features be binary.

language: system of communication

Language is a complex system that uses signs to encode and decode information to convey communication.

Language must be statable in a finite form (Davidson, 2005; von Humboldt, 1836; Chomsky, 1969; Apostel et al., 1957; Wang, 1991).

Similarly, requiring oppositions to be 'proportional' rather than 'isolated', i.e. requiring that there be more than one opposition for which the opposed phonemes differ by some given set of features, is simply to reduce the number of different features used for a given number of oppositions and thus to increase economy, or parsimony.

language: system of communication
Language is a complex system that uses signs to encode and decode information to convey communication.

Language must be statable in a finite form (Davidson, 2005; von Humboldt, 1836; Chomsky, 1969; Apostel et al., 1957; Wang, 1991).

Apostel then goes on to show that such constraints on the kind of oppositions allowed in a ‘coherent’ phonological system are exactly analogous to the constraints which one would impose upon an error-correcting code for optimality of coding and error-correction. Briefly, an error-correcting code is one in which redundancy has been introduced by the addition to certain message units of extra symbols so chosen that a random, noise-produced alteration in one or more message symbols will result in a detectable change in some prearranged arithmetic or logical function and thus permit the detection and correction of the error.

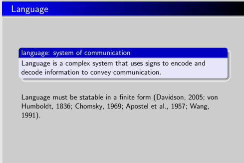
Language

language: system of communication

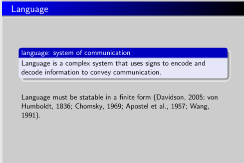
Language is a complex system that uses signs to encode and decode information to convey communication.

Language must be statable in a finite form (Davidson, 2005; von Humboldt, 1836; Chomsky, 1969; Apostel et al., 1957; Wang, 1991).

One scheme for detecting one error per word would require each code word to differ from every other code word by at least two letters, for an alteration of one letter in any word would render that word different from all admissible code words (but the error would not be correctable, for any given nonword received might be an alteration of several code words). Now optimal coding requires that these ‘distances’ between code words be all equal, i.e. consist of the same number of letter differences.



This and certain other simple requirements for optimality may now be interpreted as requirements of symmetry, binarity, and parsimony in a branching-diagram; for just as a code word is an ordered sequence of letter positions which can take on only certain values (say 0 and 1), so a phoneme may be taken as an ordered set of features each of which can take on only certain values (say + or -, in the case of the binary choices admitted here). Thus, as far as symmetry of the branching-diagram is concerned, code words in a simple Hamming error-detecting code are a model of phonemes considered as sets of phonetic features under the requirement of coherence.

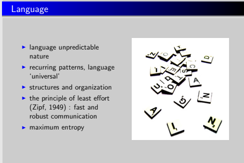


However, in order to generate most efficiently all the sentences of a language, and to do this in such a manner as to segregate the optional phonemic rules from the obligatory phonetic rules, and in such a manner as to represent properly all the morphophonemic and phonemic facts known about sentences, it is necessary to build into the grammar of a language-indeed, into the phonological part-a great deal more specificity of structure than is given by these relatively superficial constraints of binarity, symmetry, and parsimony. In other words, our coding model, though it may be mathematically correct, is compatible with too many different possible grammars, for it does not reach to a deep enough level of linguistic structure.

(Logique, langage et theorie de l'information reviewed by R. B. LEES)

Language

- language unpredictable nature
- recurring patterns, language 'universal'
- structures and organization
- the principle of least effort (Zipf, 1949) : fast and robust communication
- maximum entropy

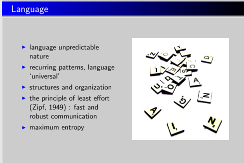


Languages have unpredictable nature and without it there would be no communication at all. If it was a certain deterministic event, it would have no information associated.

The linguistic analysis of a language is the observation of certain recurring patterns. Some patterns that are said to be frequently found in the languages of the world are called 'universal'.

To study and understand how languages works, it is important to find/define its structures and the way they are organized to build communication.

The idea proposed by Zipf (1949) is that language works seeking the principle of least effort. The process of communication is better when it is possible to transmit information in a fast and robust way.



The 'linguistic universals' paradigm gives the impression that languages are all built to a common pattern. In fact, there are vanishingly few universals of language in the direct sense that all languages exhibit them. Instead, diversity can be found at almost every level of linguistic organization. This fundamentally changes the object of enquiry from a cognitive science perspective.

Language

Language is a complex adaptive system.

(1) multiple agents

(2) adaptive

(3) perception and motivation

(4) emergence of patterns

Language

Language is a complex adaptive system.
(1) multiple agents
(2) adaptive
(3) perception and motivation
(4) emergence of patterns

We argue that this system is best construed as a complex adaptive system (CAS). This system is radically different from the static system of grammatical principles characteristic of the widely held generativist approach. Instead, language as a complex adaptive system of dynamic usage and its experience involves the following key features: (1) The system consists of multiple agents (the speakers in the speech community) interacting with one another. (2) The system is adaptive, that is, speakers' behavior is based on their past interactions, and current and past interactions together feed forward into future behavior. (3) A speaker's behavior is the consequence of competing factors ranging from perceptual mechanics to social motivations. (4) The structures of language emerge from interrelated patterns of experience, social interaction, and cognitive processes.

Language

Language is a complex adaptive system.

(1) multiple agents
(2) adaptive
(3) perception and motivation
(4) emergence of patterns

Humans are great at statistical learning, which is the ability to track, sort and categorize sounds and visual patterns.

When it comes to learning either auditory or visual languages, infants are tracking the pattern of how words and parts of words come together. And by tracking the patterns, they can, based on the statistical regularities of which sounds came together, figure out what a word is, or where a word begins and ends. From this, their brains can form sound categories for their native language, which will cue them to pay more attention to those sounds in the future.

Language

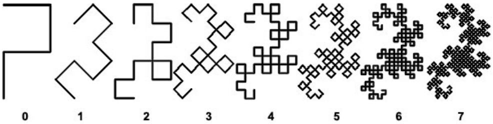


Figure: Dragon fractal.

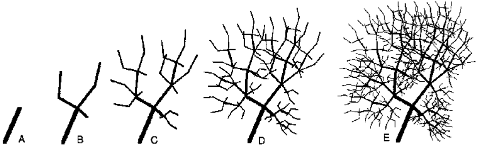


Figure: Purkinje cell, fractal model.

Language

Figure: Dragon fractal.

Figure: Purkinje cell, fractal model.

Our assumption is that the complex patterns and rules observed in language communication are emergent from the multiple interactions of speech agents over time. We believe there are simple fundamental laws that rule a multitude of microscopic interactions which create the observed macroscopic properties apparently complex. We believe language presents a fractal nature.

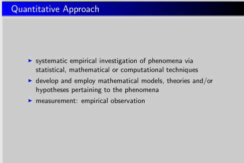
Quantitative Approach

- systematic empirical investigation of phenomena via statistical, mathematical or computational techniques
- develop and employ mathematical models, theories and/or hypotheses pertaining to the phenomena
- measurement: empirical observation

An immense number of properties and processes in language which can be detected and analysed only with quantitative methods on the basis of quantitative concepts: features and interrelations which can be expressed only by numbers or rankings

It can be shown that these properties of linguistic elements and their interreations abide by universal laws, which can be formulated in a strict mathematical way - in analogy to the laws of the well-known natural sciences. Emphasis has to be put on the fact that these laws are stochastic; they do not capture single cases (this would neither be expected nor possible), they rather predict the probabilities of certain events or certain conditions in a whole.

abide: accept or act in accordance with (a rule, decision, or recommendation)



A law can be said to be a statement representing universal patterns in the world (the phenomenological type of law) or universal mechanisms (the representational or mechanistic type).

types of laws: 1) probability distributions, i.e. it makes predictions about the number of units of a given property - Zipf-Mandelbrot Law 2) functional type, because these laws link two (or more) variables, i.e. properties - Menzerath's Law, which relates the size of linguistic constituents to the size of the corresponding construct 3) developmental type - a property is related to time - Piotrowski Law, which represents the development (increase and/or decrease) of the portion of new units or forms over time.

Quantitative Linguistics - History I

- date back in the ancient Greek - applications of combinatorics
- 718-791, philologist and lexicographer Al-Khalil ibn Ahmad - permutations and combinations to list all possible Arabic words with and without vowels
- 1564-1614, William Bathe - *Janua Linguarum*, the world's first language teaching texts, where he had compiled a list with 5.300 essential words

Quantitative Linguistics - History II

- the first scientific counts of units of language or text were published already in the 19th century as a means of linguistic description - in Germany, Förstemann (1846, 1852) and Drobisch (1866), in Russia, Bunjakovskij (1847), in France Bourdon (1892), in Italy, Mariotti (1880), in England, Augustus De Morgan (1851) and in the USA, probably Sherman (1888)
- the Russian mathematician Andrey Andreyevich Markov who created the base of the theory of Markov chains in 1913
- George Kingley Zipf was the first to set up a theoretical model in order to explain the observations and to find a mathematical formula for the corresponding function - the famous “Zipf's Law” (1935, 1949)
- Benoît Mandelbrot (1953, 1959, 1961a, 1961b)

Quantitative Linguistics - History III

- Shannon and Weaver (1949) - Information Theory
- Gustav Herdan (1954, 1956, 1960, 1962, 1964 1966, 1969), Rajmund.G. Piotrowski (1959, 1968, 1979) and Walter Meyer-Eppler (1959)
- today, Quantitative Linguistics is a well-developed scientific discipline with a broad applicational impact

Laws in Quantitative Linguistics

- Zipf-Mandelbrot - frequency and rank are inversely related
- Menzerath - the size of linguistic constituents decreases as the size of the corresponding construct increases
- Heaps (Herdan) - relation between lexical size and text size
- Piotrowski - development of new units or forms over time

Laws in Quantitative Linguistics (Universität Trier) http://lql.uni-trier.de
Glottopedia http://www.glottopedia.org/

Laws in Quantitative Linguistics

- Zipf-Mandelbrot - frequency and rank are inversely related
- Menzerath - the size of linguistic constituents decreases as the size of the corresponding construct increases
- Heaps (Herdan) - relation between lexical size and text size
- Piotrowski - development of new units or forms over time

Laws in Quantitative Linguistics (Universität Trier) http://lql.uni-trier.de
Glottopedia http://www.glottopedia.org/

The Piotrowski's law states over the language change, proposing a mathematical model to it and delimiting which path could changes produce. The basic idea is that the change comes from the realization of one pearson and that may produces a gradual spread. The greater the number of people that takes on this change, the faster will be the spreading process.

This idea seems in accordance with the 'lexical diffusion' proposed by Wang.

Structuralism

Structural linguistics

- Ferdinand de Saussure - *Course in General Linguistics* in 1916.
- theoretical paradigm: elements must be understood in terms of their relationship to a larger, overarching system or structure
- linguistic signs were composed of two parts: *signifier* and *signified*
- linguistic levels: the phonemes, morphemes, lexical categories, noun phrases, verb phrases, and sentence types

"En matière de langue on s'est toujours contenté d'opérer sur des unités mal définies"[1] (de Saussure, 1916).

[1] In language's matter it has always been sufficient to operate on ill-defined units.

Structuralism

Structural linguistics

- Ferdinand de Saussure - *Course in General Linguistics* in 1916.
- theoretical paradigm: elements must be understood in terms of their relationship to a larger, overarching system or structure
- linguistic signs were composed of two parts: *signifier* and *signified*
- linguistic levels: the phonemes, morphemes, lexical categories, noun phrases, verb phrases, and sentence types

"En matière de langue on s'est toujours contenté d'opérer sur des unités mal définies"[1] (de Saussure, 1916).

[1] In language's matter it has always been sufficient to operate on ill-defined units.

The structuralism brings a new paradigm where we understand language as system where elements stablish relations one to another and to a larger structure creating the whole.

It creates a dissociation of the signifier and signified and brings the linguistic analysis to multiple levels.

In order to stablish a logical interpretation and understanding of language, it is necessary to do so by establishing linguistic units that my be used in the process of structuring of the understanding of the linguistic system. As pointed by Saussure, it has always been suffcient to operate on ill-denied units.

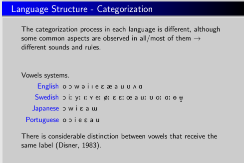
Language Structure - Categorization

The categorization process in each language is different, although some common aspects are observed in all/most of them $\rightarrow$ different sounds and rules.

Vowels systems.

English	o ɔ w ə i ı e ɛ æ a u ʊ ʌ ɑ
Swedish	ɔ iː yː ɪː ʏ eː øː ɛ ɛː œ a uː ʊ oː ɑː ɵ ʉ̞
Japanese	ɔ w i ɛ a ɯ
Portuguese	o ɔ i e ɛ a u

There is considerable distinction between vowels that receive the same label (Disner, 1983).



The sound of languages are also not equal across the multitude of languages in the World. Not only the sounds differ from language to language, but also the way they might be combined.

One big difficulty in learning a foreign language lays on the fact that different languages make different categorizations.

Disner (1983) present the comparison between different vowel systems in many languages. Vowels that are classified, according to the IPA classification as the same, may have considerable distinction in their qualities.

Vowel system: Yoruba x Italian

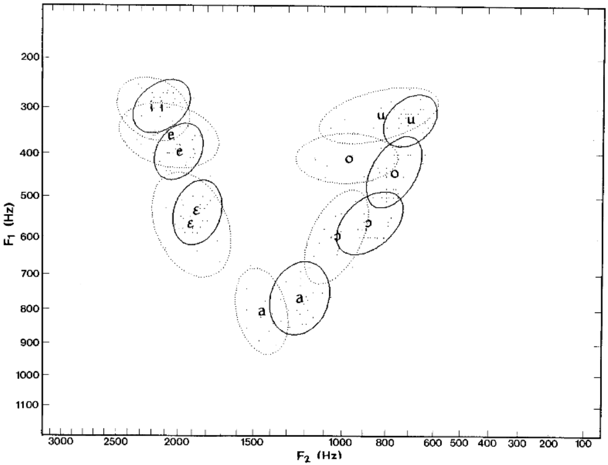


Figure: Shared vowels of Yoruba (dotted) and Italian (solid) (Disner, 1983).

Vowel system: German x Dutch

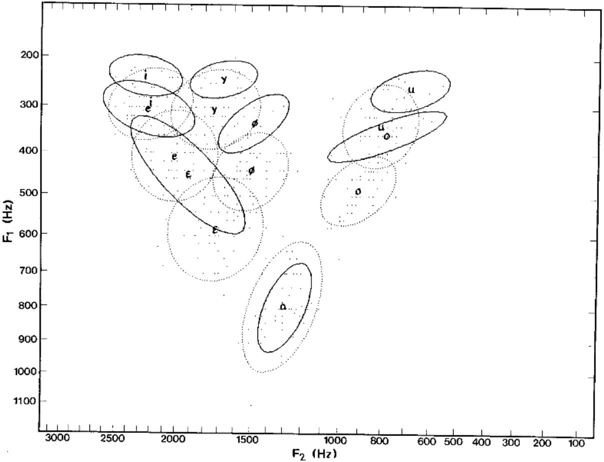


Figure: Shared vowels of German (solid) and Dutch (dotted) (Disner, 1983).

Language Structure - Combinations

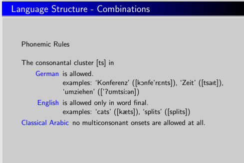
Phonemic Rules

The consonantal cluster [ts] in

German is allowed.
examples: ‘Konferenz’ ([kɔnfe’rɛnts]), ‘Zeit’ ([tsaɪt]), ‘umziehen’ ([’ʔʊmtsiːən])

English is allowed only in word final.
examples: ‘cats’ ([kæts]), ‘splits’ ([splits])

Classical Arabic no multiconsonant onsets are allowed at all.

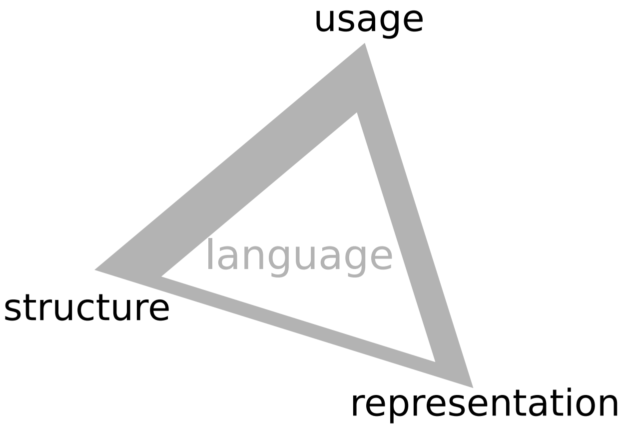


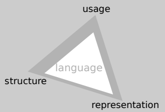
We observe that speech inventory among languages is quite diverse, but there seems to exist some cornerstones and some patterns that are preferred over others. Phonemic rules are also different among different languages, creating different patterns, but some of them are more frequent.

It is important to investigate and understand what the reasons, what are the ubiquitous patterns and from this knowledge formulate/construct an explanation for how language works.

Classical Arabic, also known as Quranic Arabic, is the form of the Arabic language used in literary texts from Umayyad and Abbasid times (7th to 9th centuries). It is based on the Medieval dialects of Arab tribes. Modern Standard Arabic (MSA) is the direct descendant used today throughout the Arab World in writing and in formal speaking.

Language Structure, Usage and Representation





"In particular, the frequency with which individual words or sequences of words are used and the frequency with which certain patterns recur in a language affects the nature of mental representation and in some cases the actual phonetic shape of words" (Bybee, 2003).

"It is certainly possible that the way language is used affects the way it is represented cognitively, and thus the way it is structured" (Bybee, 2003).

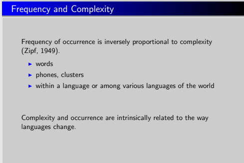
"The proposal that frequency of use affects representation suggests a very different view of lexical storage and its interaction with other aspects of the grammar or phonology than that assumed in most current theories. Structuralist and generative theories assume that the lexicon is a static list, and that neither the rules nor the lexical forms of a language are changed at all by instances of use" (Bybee, 2003).

Frequency and Complexity

Frequency of occurrence is inversely proportional to complexity (Zipf, 1949).

- words
- phones, clusters
- within a language or among various languages of the world

Complexity and occurrence are intrinsically related to the way languages change.



Zipf (1949) proposes that the frequency a phoneme in a language occurs is inversely proportional to its complexity. The frequency of clusters should also be inversely proportional to the complexity of the cluster and now, the complexity of a cluster needs to be defined by the relation created by its parts.

In the history of any language, the phonemic system undergoes constant changes that may affect the complexity and occurrence frequency of phonemes.

Outline

1. Introduction
2. Interlanguage Statistical Analysis
   - UPSID - Statistical Analysis
3. Intralanguage Statistical Analysis

UCLA Phonological Segment Inventory Database

The UCLA Phonological Segment Inventory Database (or UPSID) is a statistical survey of the phoneme inventories in 451 of the world's languages. The database was created by American phonetician Ian Maddieson for the University of California, Los Angeles (UCLA) in 1984 and has been updated several times.

`http://www.linguistics.ucla.edu/faciliti/sales/software.htm`

Book: Patterns of sounds (Maddieson, 1984).

UCLA Phonological Segment Inventory Database

The UCLA Phonological Segment Inventory Database (or UPSID) is a statistical survey of the phoneme inventories in 451 of the world's languages. The database was created by American phonetician Ian Maddieson for the University of California, Los Angeles (UCLA) in 1984 and has been updated several times.

http://www.linguistics.ucla.edu/faciliti/sales/software.htm

Book: Patterns of sounds (Maddieson, 1984).

Patterns of Sounds describes the frequency and distributional patterns of the phonemic sounds in a large and representative sample of the world's languages. Questions of the frequency and co-occurrence of the particular segment types are discussed in detail and possible explanations for the patterns observed are evaluated.

UPSID - Number of Segments (919)

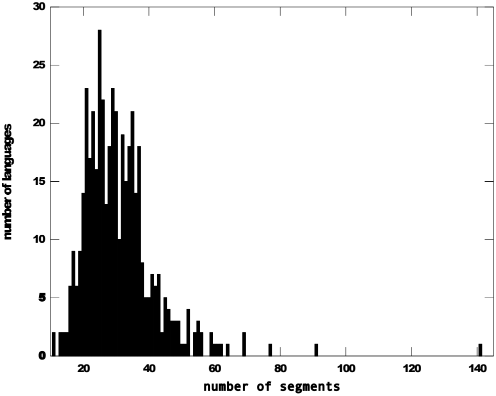


Figure: Number of phones in various languages of the world.

UPSID - Number of Segments (919)

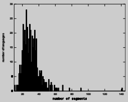


Figure: Number of phones in various languages of the world.

Among the 451 languages in the UPSID databese, the minimum number of segments used by a language is 11, in only two languages (Pirahã, spoken in Brazil by around 300 speakers; and Rotokas, spoken in Papua New Guinea by approximately 4,300 speakers). The language with the highest number of segments found in the UPSID database has 141 segments (the !Xu language, also called !Kung, is spoken by fifteen thousand speakers in Namibia and Angola). On average, the languages are built on 31 segments. Figure 5 bellow shows the histogram of languages regarding the number of segments used in each of them. 66.3% of the languages have a repertoire from 20 to 37 speech sounds. The languages use between 1.2% and 15.3% of all available speech sounds (919).

UPSID - Segments in Languages (451)

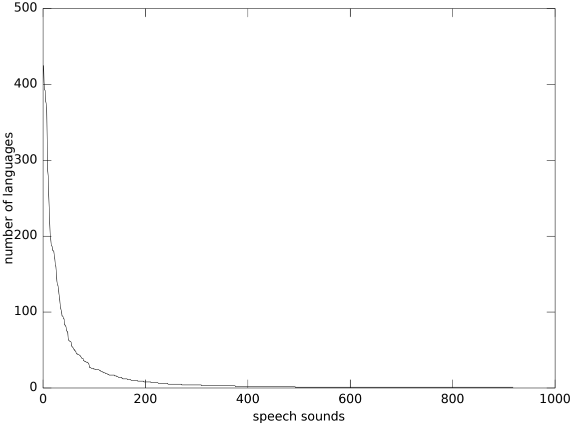


Figure: Number of languages where a give speech sounds occurs.

UPSID - Segments in Languages (451)

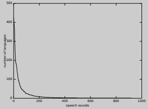


Figure: Number of languages where a give speech sounds occurs.

46.5% of the sounds appear in only one language. 80% of the sounds appear in 10 or fewer languages.

UPSID - Segments in Languages

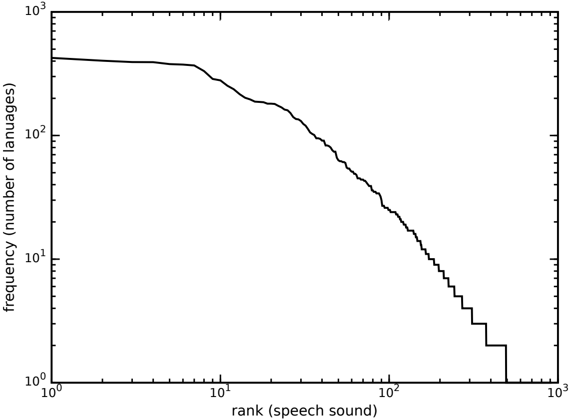


Figure: Number of languages for a given speech sound.

UPSID - Number of Vowels (269)

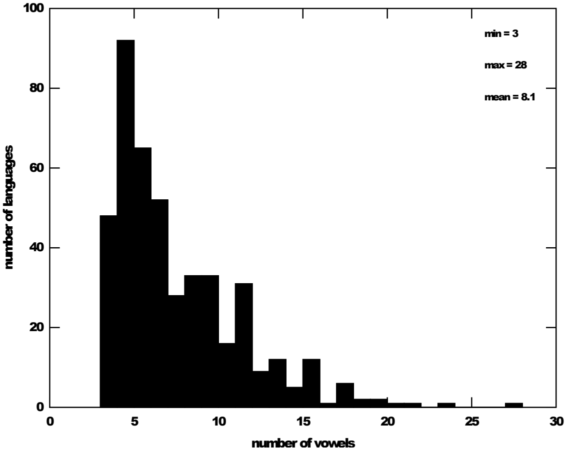


Figure: Number of vowels in various languages of the world.

UPSID - Number of Vowels (269)

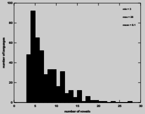


Figure: Number of vowels in various languages of the world.

Considering that the database with 451 languages has in its inventory 652 consonants (71%) and 269 vowels (29%), we see that across the languages, the number of vowels used corresponds from 0.9% to 17.9% (averaging 3.0%) of the possible vowels.

UPSID - Number of Consonants (652)

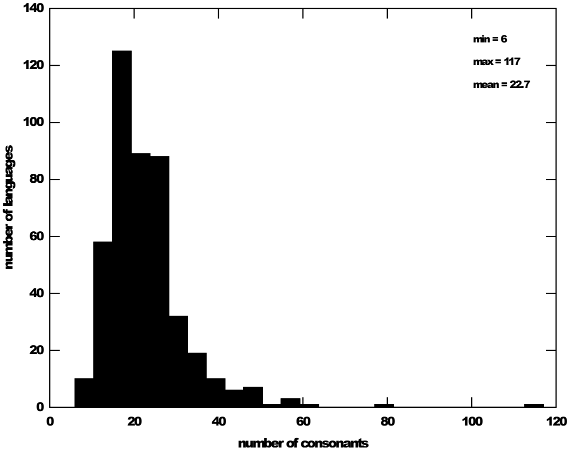


Figure: Number of consonants in various languages of the world.

UPSID - Number of Consonants (652)

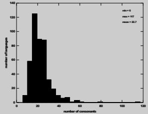


Figure: Number of consonants in various languages of the world.

The number of consonants used goes from 0.9% to 14.6% (averaging 3.5%) of the consonants in the database (which has an inventory of 652 consonants).

UPSID - Consonant-Vowels Ratio

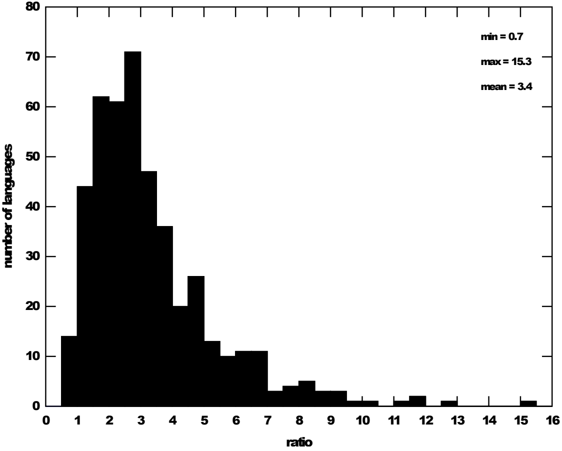


Figure: Consonants to Vowel ratio in various languages of the world.

UPSID - Consonant-Vowels Ratio

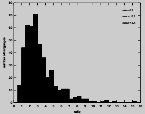


Figure: Consonants to Vowel ratio in various languages of the world.

klao	0.6875	yagua	0.91667	(...)		rutul	11.8
andoke	0.73333	cubeo	0.91667	acoma	9.2		
vanimo	0.75	panare	0.92308	shilha	9.3333	jaqaru	12
apinaye	0.76471	kaingang	0.92857	coola	9.3333		
dan	0.85	kashmiri	0.96429	yanyuwa	9.6667	tsimshian	12.667
bruu	0.90909	japreria	1	dahalo	10.2		
barasano	0.91667	maxakali	1	hadza	11.4	haida	15.333

UPSID - Consonant-Vowels Ratio CDF

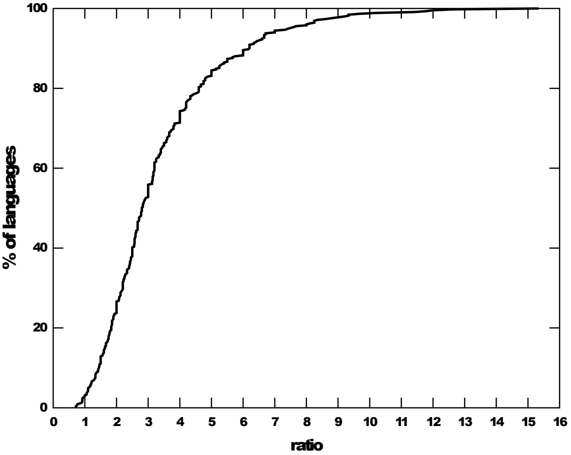


Figure: Cumulative distribution of the Consonant-Vowel ratio.

UPSID - Consonants

Table: List of the 20 most frequent consonants in UPSID.

consonant	m	k	j	p	w	b	h
n. of languages	425	403	378	375	332	287	279
frequency	94.2	89.4	83.8	83.2	73.6	63.6	61.9

consonant	g	ŋ	ʔ	n	s	tʃ	ʃ
n. of languages	253	237	216	202	196	188	187
frequency	56.1	52.6	47.9	44.8	43.5	41.7	41.5

consonant	t	f	l	n̪	t̪	ɲ
n. of languages	181	180	174	160	152	141
frequency	40.1	39.9	38.6	35.5	33.7	31.3

UPSID - Vowels

Table: List of the 10 most frequent vowels in UPSID.

vowel	i	a	u	ɛ	o/ɔ
n. of languages	393	392	369	186	181
frequency	87.1	86.9	81.8	41.2	40.1
vowel	e/ɛ	ɔ	o	e	a
n. of languages	169	162	131	124	83
frequency	37.5	35.9	29.0	27.5	18.4

UPSID - Vowels

Table: List of the 10 most frequent vowels in UPSID.

vowel	i	a	u	ɛ	o/ɔ
n. of languages	393	392	369	186	181
frequency	87.1	86.9	81.8	41.2	40.1
vowel	e/ɛ	ɔ	o	e	a
n. of languages	169	162	131	124	83
frequency	37.5	35.9	29.0	27.5	18.4

The most frequent vowels are not so frequent across languages in comparison to the most frequent consonants.

UPSID - number of phones vs. frequency index

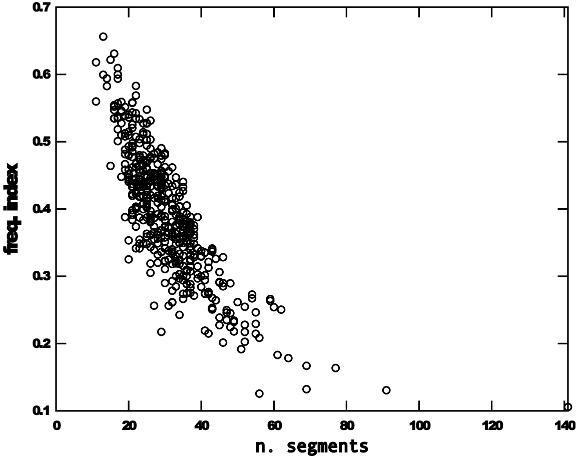


Figure: Relation between the frequency index and the number of phones in a language. (Data from UPSID)

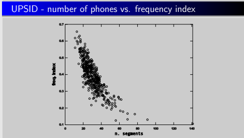


Figure: Relation between the frequency index and the number of phones in a language. (Data from UPSID)

Frequency index is the arithmetic average of the segment frequencies of a language. A language with mostly rare segments will have a low frequency index, whereas a language with mostly common sounds will have a high frequency index. A frequency index of 0.1 means that a language has many very rare segments; 0.7 means it has many common segments; the average frequency index of all languages is 0.39.

UPSID - number of phones vs. frequency index

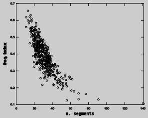


Figure: Relation between the frequency index and the number of phones in a language. (Data from UPSID)

Note that there is a relation between frequency index and number of segments in a language. That is, if a language has only few segments, it is likely that these are rather common in the languages in UPSID. On the other hand, a language with many segments will also have many segments that are uncommon in the UPSID database. This does not necessarily mean that certain sounds are more natural but it is a probabilistic effect: if you make a pot with many red marbles, few green marbles, and other marbles with different colors and you draw a small random sample (i.e. 10 marbles) you will have mostly red marbles. If you draw a large random sample (e.g. 100 marbles) you will have many single colored ones.

Frequency index: mean=0.39, max=0.66, min=0.11.

UPSID - cooccurrence of phones in a language

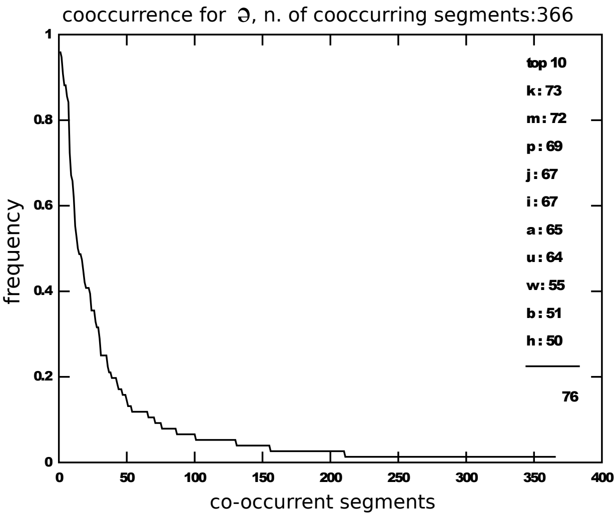
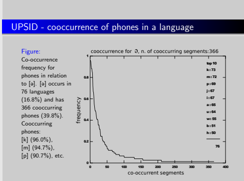


Figure: Co-occurrence frequency for phones in relation to [ə]. [ə] occurs in 76 languages (16.8%) and has 366 cooccurring phones (39.8%). Cooccurring phones: [k] (96.0%), [m] (94.7%), [p] (90.7%), etc.



This graph presents the co-occurrence frequency for phones in relation to [ə]. The phone [ə] occurs in 76 languages (16.8%). It has 366 cooccurring phones in UPSID, that means 39.8% of 919 phones. Among them, the phone [k] has a frequency of 96.0% (73/76). [m] follows with 94.7% (72/76) and [p] with 90.7% (69/76).

UPSID - cooccurrence of phones in a language

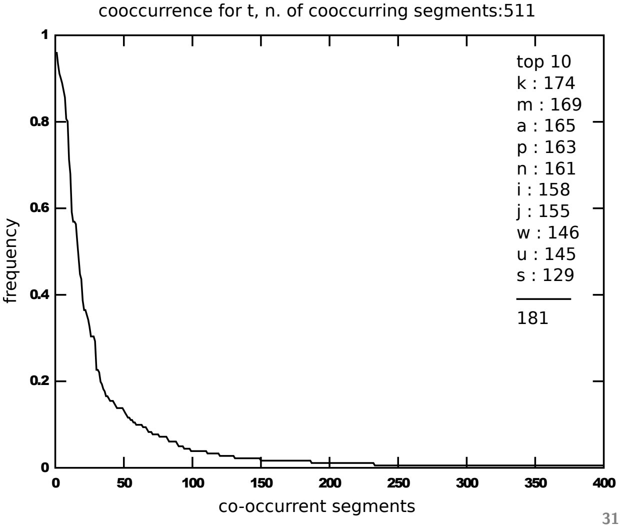
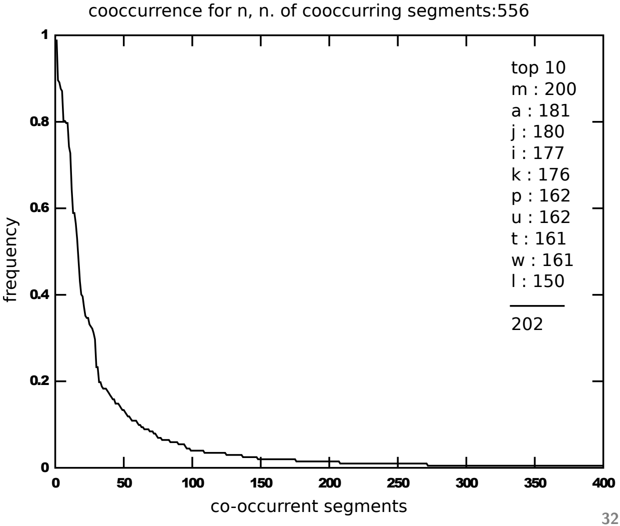


Figure: Co-occurrence frequency for phones in relation to [t]. [t] occurs in 181 languages (40.1%) and has 511 cooccurring phones (55.6%). Cooccurring phones: [k] (96.1%), [m] (93.3%), [a] (91.2%), etc.

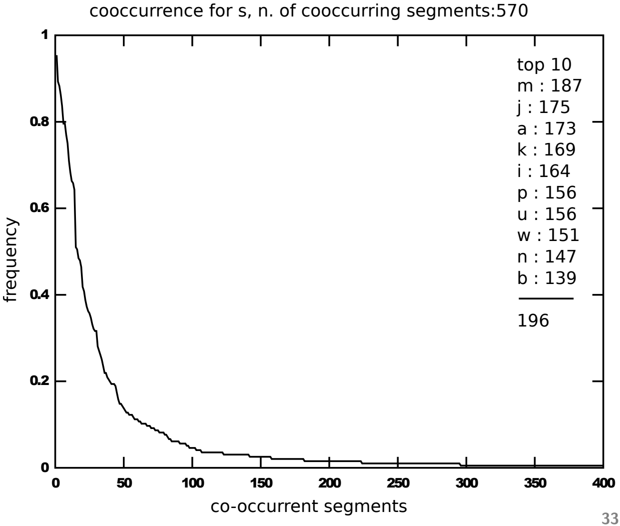
UPSID - cooccurrence of phones in a language

Figure:
Co-occurrence frequency for phones in relation to [n]. [n] occurs in 202 languages (44.7%) and has 556 cooccurring phones (60.5%). Cooccurring phones:
[m] (99.0%),
[a] (89.6%),
[j] (89.1%), etc.



UPSID - cooccurrence of phones in a language

Figure: Co-occurrence frequency for phones in relation to [s]. [s] occurs in 196 languages (43.5%) and has 570 cooccurring phones (62.0%). Cooccurring phones: [m] (95.4%), [j] (89.3%), [a] (88.3%), etc.



UPSID - cooccurrence of phones in a language

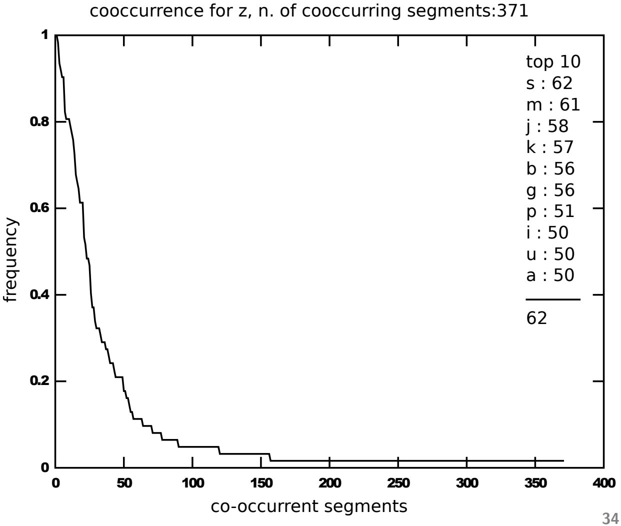
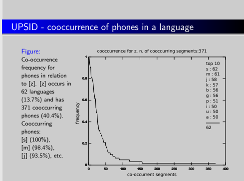


Figure: Co-occurrence frequency for phones in relation to [z]. [z] occurs in 62 languages (13.7%) and has 371 cooccurring phones (40.4%). Cooccurring phones: [s] (100%), [m] (98.4%), [j] (93.5%), etc.



Always when the phone [z] (voiced alveolar fricatives) is present, the phone [s] (voiceless alveolar fricatives) is also present.

UPSID - cooccurrence of phones in a language

Table: List of phones and their top 8 co-occurring pairs with their relative frequency of occurrence (data from UPSID).

phone	co-occurring phone with their respective relative frequency (%)							
ə	k : 96.1	m : 94.7	p : 90.8	j : 88.2	i : 88.2	a : 85.5	u : 84.2	w : 72.4
t	k : 96.1	m : 93.4	a : 91.2	p : 90.1	n : 89.0	i : 87.3	j : 85.6	w : 80.7
n	m : 99.0	a : 89.6	j : 89.1	i : 87.6	k : 87.1	p : 80.2	u : 80.2	t : 79.7
ɪ	m : 97.3	j : 91.9	k : 86.5	a : 83.8	p : 83.8	ʊ : 74.3	w : 71.6	b : 66.2
s	m : 95.4	j : 89.3	a : 88.3	k : 86.2	i : 83.7	p : 79.6	u : 79.6	w : 77.0
z	s : 100.0	m : 98.4	j : 93.5	k : 91.9	b : 90.3	g : 90.3	p : 82.3	i : 80.6
d	b : 96.7	m : 94.2	i : 91.7	a : 90.8	j : 90.0	n : 89.2	g : 87.5	u : 86.7
l	m : 98.9	j : 89.1	k : 86.8	n : 86.2	a : 86.2	i : 85.6	p : 81.0	w : 79.9
i	m : 93.9	u : 91.6	k : 89.6	a : 89.1	p : 82.7	j : 82.7	w : 73.8	b : 65.1
d̪	b : 97.5	m : 96.2	g : 93.8	j : 88.8	k : 85.0	t̪ : 83.8	i : 80.0	p : 76.2
m	k : 89.2	i : 86.8	a : 86.6	j : 85.2	p : 82.6	u : 81.6	w : 74.1	b : 64.2
n	m : 99.0	a : 89.6	j : 89.1	i : 87.6	k : 87.1	p : 80.2	u : 80.2	t : 79.7
k	m : 94.0	p : 91.3	i : 87.3	a : 86.6	j : 83.4	u : 82.1	w : 73.2	b : 62.8
g	b : 96.4	m : 95.3	i : 89.3	j : 87.4	k : 86.2	u : 84.2	a : 83.0	p : 76.7
p	k : 98.1	m : 93.6	a : 87.2	i : 86.7	j : 82.9	u : 81.3	w : 71.7	b : 60.5
b	m : 95.1	i : 89.2	k : 88.2	j : 86.8	g : 85.0	u : 84.3	a : 84.3	p : 79.1
ʃ	m : 94.7	j : 91.4	k : 88.8	i : 84.0	a : 82.9	p : 80.2	u : 78.1	w : 74.3
ʒ	ʃ : 95.1	m : 95.1	j : 90.2	k : 90.2	b : 82.0	g : 80.3	i : 78.7	p : 78.7

UPSID - cooccurrence of phones in a language

Table: List of phones and their top 8 co-occurring pairs with their relative frequency of occurrence (data from UPSID).

phone	co-occurring phone with their respective relative frequency (%)							
a	k : 96.1	m : 94.7	p : 90.8	j : 88.2	i : 88.2	a : 85.5	u : 84.2	w : 72.4
t	k : 96.1	m : 93.4	a : 91.2	p : 90.1	n : 89.0	i : 87.3	j : 85.6	w : 80.7
n	m : 99.0	a : 89.6	j : 89.1	i : 87.6	k : 87.1	p : 80.2	u : 80.2	t : 79.7
ɪ	m : 97.3	j : 91.9	k : 86.5	a : 83.8	p : 83.8	u : 74.3	w : 71.6	b : 66.2
s	m : 95.4	j : 89.3	a : 88.3	k : 86.2	i : 83.7	p : 79.6	u : 79.6	w : 77.0
z	s : 100.0	m : 98.4	j : 93.5	k : 91.9	b : 90.3	g : 90.3	p : 82.3	i : 80.6
d	b : 96.7	m : 94.2	i : 91.7	a : 90.8	j : 90.0	n : 89.2	g : 87.5	u : 86.7
l	m : 98.9	j : 89.1	k : 86.8	n : 86.2	a : 86.2	i : 85.6	p : 81.0	w : 79.9
i	m : 93.9	u : 91.6	k : 89.6	a : 89.1	p : 82.7	j : 82.7	w : 73.8	b : 65.1
d̪	b : 97.5	m : 96.2	g : 93.8	j : 88.8	k : 85.0	t̪ : 83.8	i : 80.0	p : 76.2
m	k : 89.2	i : 86.8	a : 86.6	j : 85.2	p : 82.6	u : 81.6	w : 74.1	b : 64.2
n	m : 99.0	a : 89.6	j : 89.1	i : 87.6	k : 87.1	p : 80.2	u : 80.2	t : 79.7
k	m : 94.0	p : 91.3	i : 87.3	a : 86.6	j : 83.4	u : 82.1	w : 73.2	b : 62.8
g	b : 96.4	m : 95.3	i : 89.3	j : 87.4	k : 86.2	u : 84.2	a : 83.0	p : 76.7
p	k : 98.1	m : 93.6	a : 87.2	i : 86.7	j : 82.9	u : 81.3	w : 71.7	b : 60.5
b	m : 95.1	i : 89.2	k : 88.2	j : 86.8	g : 85.0	u : 84.3	a : 84.3	p : 79.1
ʃ	m : 94.7	j : 91.4	k : 88.8	i : 84.0	a : 82.9	p : 80.2	u : 78.1	w : 74.3
ʒ	ʃ : 95.1	m : 95.1	j : 90.2	k : 90.2	b : 82.0	g : 80.3	i : 78.7	p : 78.7

As we take one bilabial consonants as reference, the others appear in the top-10 list, showing a strong adhesion of one bilabial consonant to another. Observing now the voiced alveolar sibilant fricative [z], its voiceless counterpart always appear, but doing the other way around analysis, taking [s] as a reference, its voiced counterpart [z] has a relative frequency of co-occurrence of 31.6%, figuring as the 29th in the list. Analyzing other pairs like [t]-[d] (alveolar plosive), [k]-[g] (velar plosive), [p]-[b] (bilabial plosive) and [ʒ]-[ʃ] (palato-alveolar fricative), it seems that the existence of the voiced counterpart subjects the existence of the voiceless much more emphatically than the other way around.

UPSID - cooccurrence of phones in a language

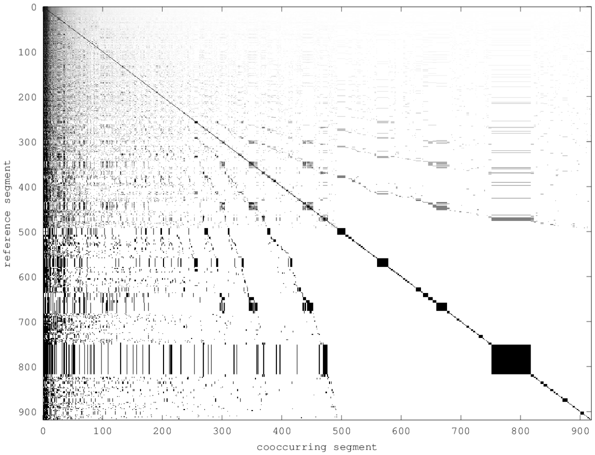


Figure: Number of cooccurring phones for each phone in UPSID. The ordered by frequency of occurrence in languages.

UPSID - cooccurrence of phones in a language

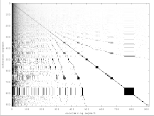
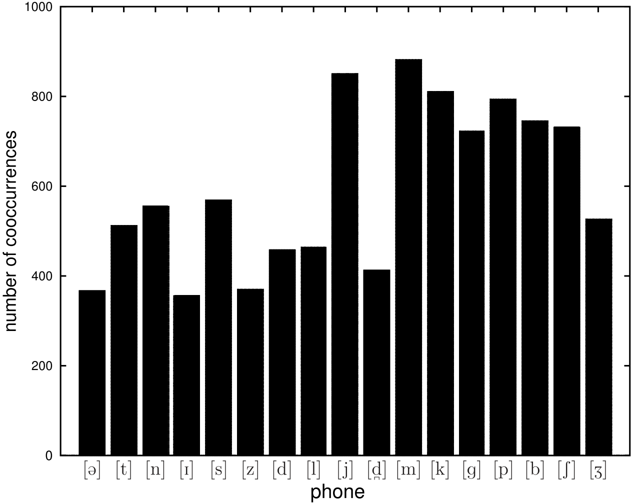


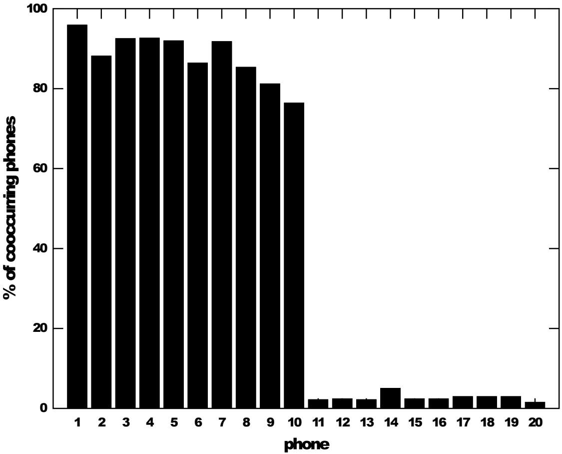
Figure: Number of cooccurring phones for each phone in UPSID. The ordered by frequency of occurrence in languages.

For each phone we listed the cooccurring ones and the number of times they cooccur (how many languages). We normalize by the number of languages in which the reference phone occurs. Observe that the most frequent phones in the database (left side) have a higher cooccurrence rate.

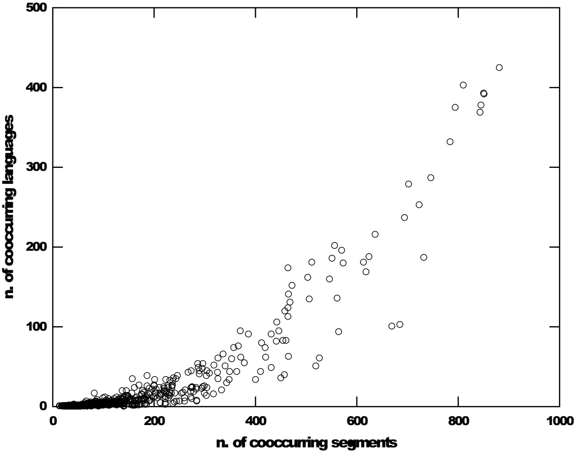
Cooccurrence for a given phone



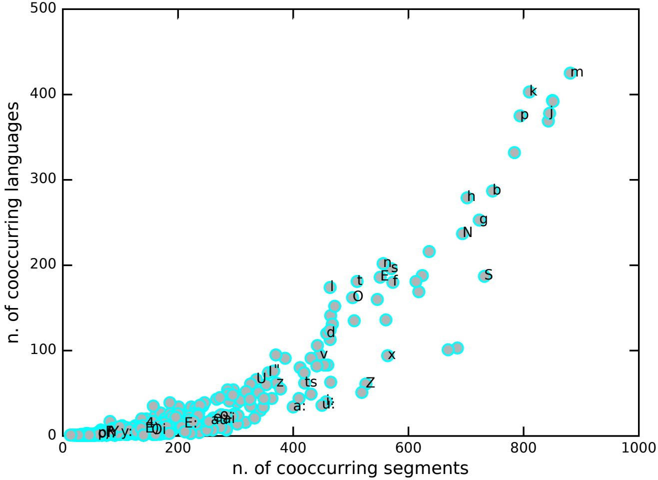
Cooccurrence: most frequent vs least frequent



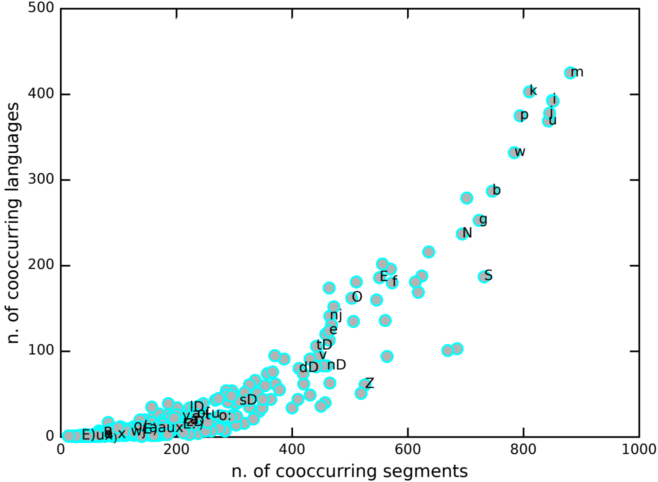
Cooccurrence : languages and segments



Cooccurrence : languages and segments - German



Cooccurrence : languages and segments - French



Outline

Approach

- synchrony
- written corpus
- pronouncing dictionary
- statistical analysis

Approach

- synchrony
- written corpus
- pronouncing dictionary
- statistical analysis

Although “languages are simultaneously products of history and entities existing at particular times”(Good, 2008) and both diachrony and synchrony aspects are important to determine what languages are, we focus here only on the synchrony aspects.
To carry out the analysis of it, we use large text databases, because it is much easier to be acquired, stored and handled.
A pronouncing dictionary is used to convert text into phones.
It is possible then to derive statistical analysis on this new database which statistically resembles a speech database.

Text Database

Project Gutenberg is a volunteer effort to provide free dgital access to cultural works. It was founded by Michael S. Hart in 1971 and is the oldest digital library. Most of its items are public domain books. Project Gutenberg claimed over 42,000 items in its collection (March 2013) (Project Gutenberg, 2013).

Google Ngram is a large database of n-grams (words combinations) based originally on 5.2 million books, published between 1500 and 2008, containing 500 billion words in American English, British English, French, German, Spanish, Russian, or Chinese (Michel et al., 2011b).

Project Gutenberg is a volunteer effort to digitize and archive cultural works, to encourage the creation and distribution of eBooks. Founded in 1971 by Michael S. Hart, it is the oldest digital library. Most of the items in its collection are the full texts of public domain books. The project tries to make these as free as possible, in long-lasting, open formats that can be used on almost any computer. As of December 2009, Project Gutenberg claimed over 33,000 items in its collection.
Today it is viable to have large databases and it is relatively easy do complile such databases and manage them. That could not be dreamt a few years ago. It would be extremily good to have large speech databases in order to study language as a spoken phenomena, investigate its usage and change, but it is a much harder task to acchieve this goal.

reCAPTCHA is a user-dialogue system originally developed by Luis von Ahn, Ben Maurer, Colin McMillen, David Abraham and Manuel Blum at Carnegie Mellon University's main Pittsburgh campus, and acquired by Google in September 2009.

Pronouncing Dictionary

Dictionary : Carnegie Mellon University Pronouncing Dictionary - is a machine-readable pronunciation dictionary for North American English that contains over 125,000 words and their transcriptions (Weide, 2008).

Phoneme set : 39 phonemes.

Symbols : ARPAbet.

Phoneme	IPA	Example	Translation	IPA Translation
AE	[æ]	at	AE T	[æt]
B	[b]	be	B IY	[bi]
S	[s]	sea	S IY	[si]

Pronouncing Dictionary

Dictionary : Carnegie Mellon University Pronouncing Dictionary - is a machine-readable pronunciation dictionary for North American English that contains over 125,000 words and their transcriptions (Weide, 2008).

Phoneme set : 39 phonemes.

Symbols : ARPAbet.

Phoneme	IPA	Example	Translation	IPA Translation
AE	[æ]	at	AE T	[æt]
B	[b]	be	B IY	[bi]
S	[s]	sea	S IY	[si]

To carry out the statistical analysis of a spoken language, we assume that a written language presents a very similar characteristics, and then the statistical analysis using a written text database and a text-to-phone transcription would be quite reasonable.

Arpabet is a phonetic transcription code developed by Advanced Research Projects Agency (ARPA) as a part of their Speech Understanding Project (1971-1976). It represents each phoneme of General American English with a distinct sequence of ASCII characters. Arpabet has been used in several speech synthesizers, including Computalker for the S-100 (Altair) system, SAM for the Commodore 64, SAY for the Amiga and TextAssist for the PC and Speakeasy from Intelligent Artefacts (see ST_Robotics) which used the Votrax SC01 speech synthesiser IC. It is also used in the CMU Pronouncing Dictionary.

Words Frequency

(1) the : 775911
(2) and : 471916
(3) of : 414499
(4) to : 350613
(5) a : 277321
(6) in : 226505
(7) i : 200689
(8) that : 173083
(9) he : 162183
(10) it : 145364
(11) was : 130804
(12) his : 129300
(13) you : 118473
(14) with : 114122
(15) is : 112640
(16) for : 107245
(17) as : 102009
(18) not : 96636
(19) be : 86896
(20) but : 81643
(21) had : 80327
(22) at : 76688
(23) her : 75761
(24) on : 75493
(25) my : 73879
(26) him : 72258
(27) have : 68463
(28) this : 67572
(29) all : 65960
(30) me : 64560
(31) by : 63944
(32) which : 63051
(33) she : 57839
(34) they : 57770
(35) from : 56128
(36) or : 52089
(37) so : 51617
(38) said : 50040
(39) no : 48930
(40) are : 45831
(41) one : 43822
(42) what : 41575
(43) them : 41320
(44) were : 40475
(45) will : 39733
(46) if : 38421
(47) there : 38209
(48) we : 37944
(49) when : 37385
(50) their : 36721
(51) who : 36109
(52) an : 35485
(53) your : 33401
(54) would : 32582
(55) do : 31225
(56) out : 30165
(57) then : 29682
(58) been : 29502
(59) up : 28860
...

Words Frequency and Zipf Law

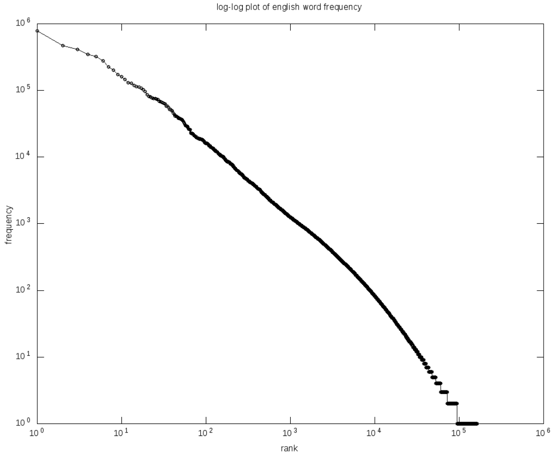


Figure: Log-log plot of words rank versus frequency of occurrence.

Letters Frequency

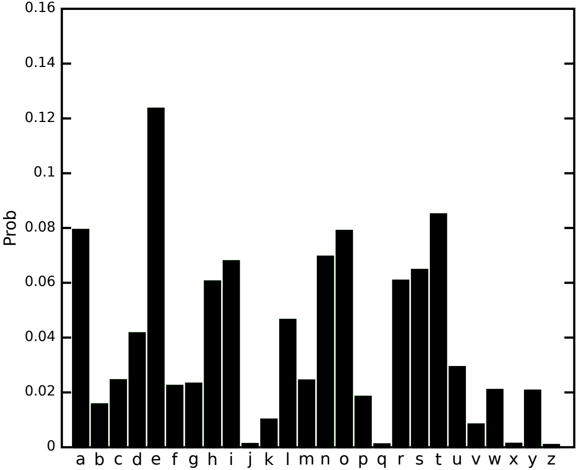


Figure: Relative frequency of letters in text.

Letters Frequency

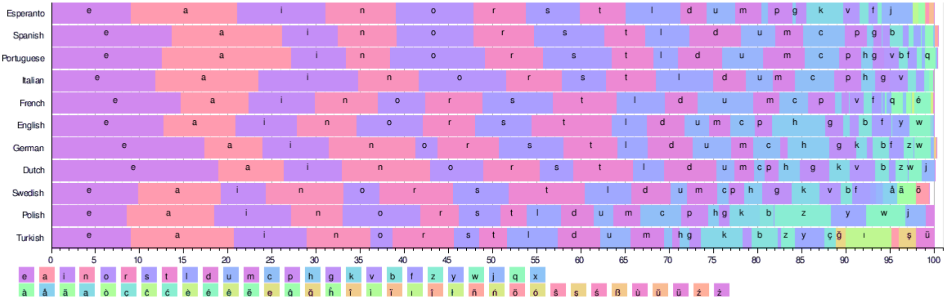


Figure: Frequency distributions of the 26 most common Latin letters across some languages (source: Wikipedia).

Words Frequency - Transcribed

(1) DH AH : 775911
(2) AH N D : 471916
(3) AH V : 414499
(4) T UW : 350613
(5) AH : 277321
(6) IH N : 226505
(7) AY : 200689
(8) DH AE T : 173083
(9) HH IY : 162183
(10) IH T : 145364
(11) W AA Z : 130804
(12) HH IH Z : 129300
(13) Y UW : 118473
(14) W IH DH : 114122
(15) IH Z : 112640
(16) F AO R : 107245
(17) AE Z : 102009
(18) N AA T : 96636
(19) B IY : 86896
(20) B AH T : 81643
(21) HH AE D : 80327
(22) AE T : 76688
(23) HH ER : 75761
(24) AA N : 75493
(25) M AY : 73879
(26) HH IH M : 72258
(27) HH AE V : 68463
(28) DH IH S : 67572
(29) AO L : 65960
(30) M IY : 64560
(31) B AY : 63944
(32) W IH CH : 63051
(33) SH IY : 57839
(34) DH EY : 57770
(35) F R AH M : 56128
(36) AO R : 52089
(37) S OW : 51617
(38) S EH D : 50040
(39) N OW : 48930
(40) AA R : 45831
(41) W AH N : 43822
(42) W AH T : 41575
(43) DH EH M : 41320
(44) W ER : 40475
(45) W IH L : 39733
(46) IH F : 38421
(47) DH EH R : 38209
(48) W IY : 37944
(49) W EH N : 37385
(50) DH EH R : 36721
(51) HH UW : 36109
(52) AE N : 35485
(53) Y AO R : 33401
(54) W UH D : 32582
(55) D UW : 31225
(56) AO L AW T : 30165
(57) DH EH N : 29682
(58) B IH N : 29502
(59) AH P : 28860
...

Phones Frequency

(1) ə : 44539
(2) t : 33131
(3) n : 31928
(4) ɹ : 28845
(5) s : 21928
(6) d : 20032
(7) r : 18563
(8) i : 16482
(9) l : 15816
(10) ɛ : 13896
(11) m : 13072
(12) ɝ : 12640
(13) k : 12308
(14) w : 11107
(15) z : 10744
(16) ð : 10720
(17) v : 10407
(18) h : 10009
(19) f : 9391
(20) ɑ : 8744
(21) æ : 8635
(22) b : 8390
(23) u : 7972
(24) p : 7501
(25) ɔ : 7429
(26) eɪ : 6196
(27) aɪ : 6148
(28) oɯ : 5283
(29) ʃ : 4915
(30) ŋ : 4861
(31) g : 3351
(32) tʃ : 2501
(33) j : 2462
(34) θ : 2309
(35) ʊ : 2276
(36) aʊ : 2242
(37) dʒ : 2100
(38) ɔɪ : 326
(39) ʒ : 314

Phones Frequency - Log-log plot

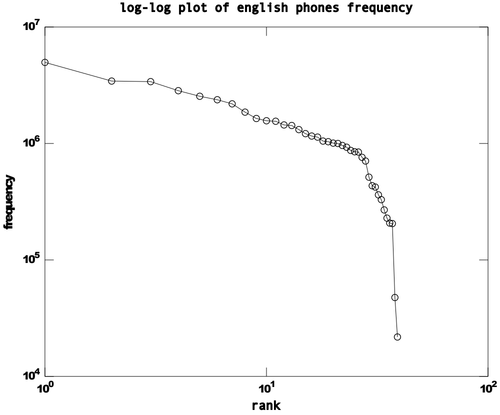


Figure: Log-log plot of phones rank versus frequency of occurrence.

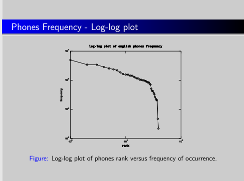


Figure: Log-log plot of phones rank versus frequency of occurrence.

This is not an example of distribution with large number of rare events ... where "rare events are common" We might say that there are no rare events. Zipf's law is found only on distributions with large number of rare events.

The simple "maximum likelihood" method for predicting estimate the probability of an event-type that has occurred r times in N trials as r/N. This generally works well if r is fairly large (and if the world doesn't change too much). But as r gets smaller, the maximum likelihood estimate gets worse. And if r is zero, it may still be quite unwise to bet that the event-type in question will never occur in the future. Even more important, the fraction of future events whose past counts are zero may be substantial.

One of the present time (2000) characteristics of the study of Zipf's law is the increase of interest in the more general concept of "distributions with a large number of rare events" (LNRE distributions), which was introduced and first systematically studied by (Khmaladze, 1987).

Probability of occurrence of [ə] across words rank

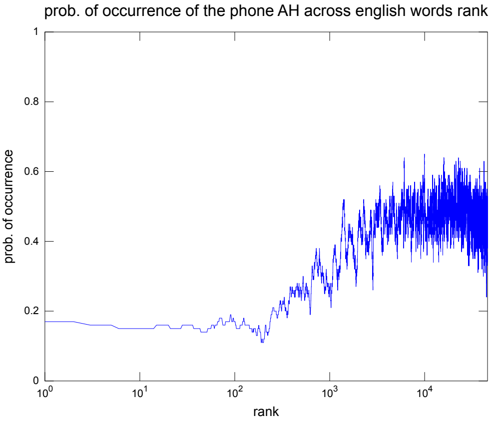


Figure: Probability of occurrence of [ə] in words versus words rank.

Probability of occurrence of [ə] across words rank

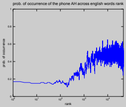


Figure: Probability of occurrence of [ə] in words versus words rank.

We might wonder whether some phones are more frequent then others because they are present in most frequent words, what would lead them along to a higher frequency of occurrence. The present graphic and the following ones shows this is a false hypothesis.

Probability of occurrence of [t] across words rank

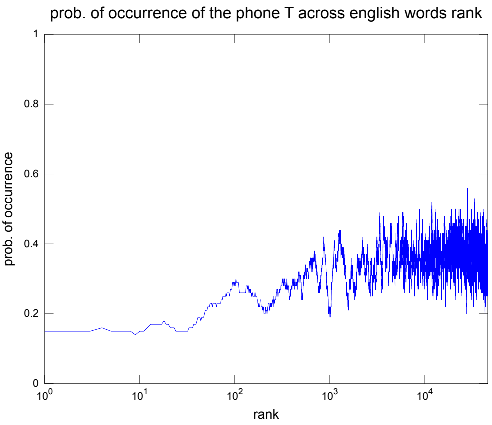


Figure: Probability of occurrence of [t] in words versus words rank.

Probability of occurrence of [n] across words rank

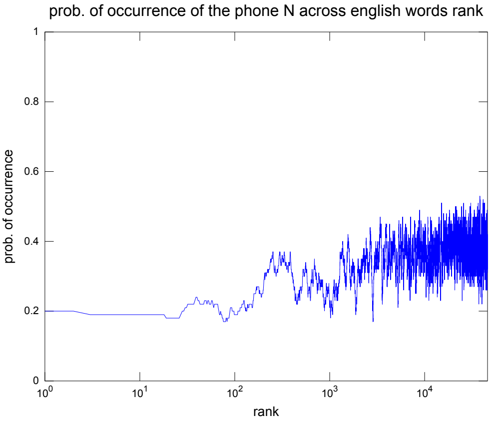


Figure: Probability of occurrence of [n] in words versus words rank.

Probability of occurrence of [ʊ] across words rank

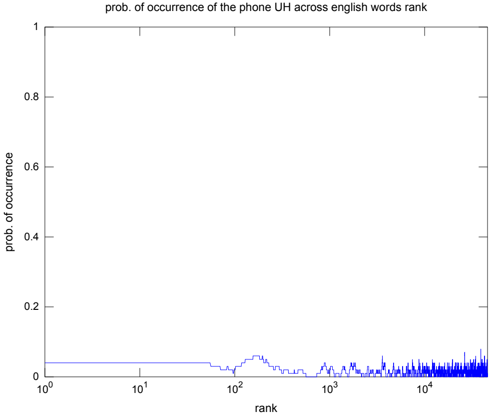


Figure: Probability of occurrence of [ʊ] in words versus words rank.

Probability of occurrence of [ɔɪ] across words rank

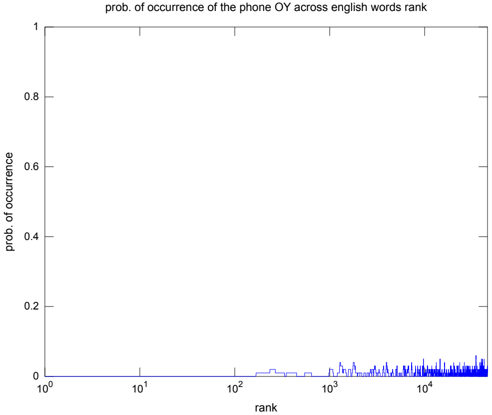


Figure: Probability of occurrence of [ɔɪ] in words versus words rank.

Probability of occurrence of [ʒ] across words rank

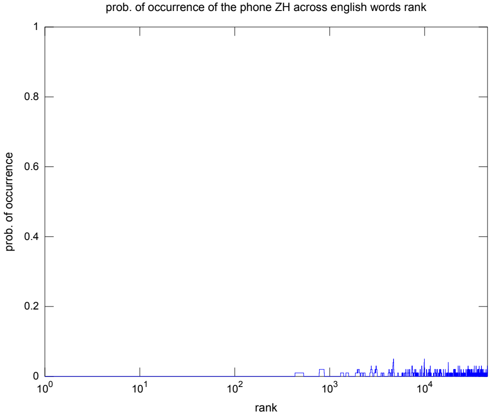


Figure: Probability of occurrence of [ʒ] in words versus words rank.

Diphones : Frequency of occurrence

(1) ən : 1296408
(2) ðə : 785354
(3) nd : 784028
(4) st : 651129
(5) əv : 489267
(6) ɔr : 472069
(7) ɪn : 470069
(8) tu : 425544
(9) tə : 420825
(10) ɪŋ : 387096
(11) ɪz : 380357
(12) əl : 372847
(13) ɛn : 349905
(14) nt : 337193
(15) æt : 284460
(16) wɪ : 282526
(17) ət : 265396
(18) ɛr : 264293
(19) rə : 261447
(20) ɪt : 260544
(21) əs : 246996
(22) ju : 245715
(23) hæ : 237724
(24) hɪ : 237490
(25) əm : 233793
(26) ri : 219997
(27) li : 219652
(28) ɑn : 213755
(29) æn : 213169
(30) sə : 210385
(31) ɪs : 208619
(32) ðæ : 194602
(33) əf : 193570
(34) ɑr : 193525

⋮

(1120) uaɪ : 1
(1121) ɔɪɑ : 1
(1122) bʃ : 1
(1123) pv : 1
(1124) iʊ : 1
(1125) ɛoʊ : 1

Diphones Frequency - Log-log plot

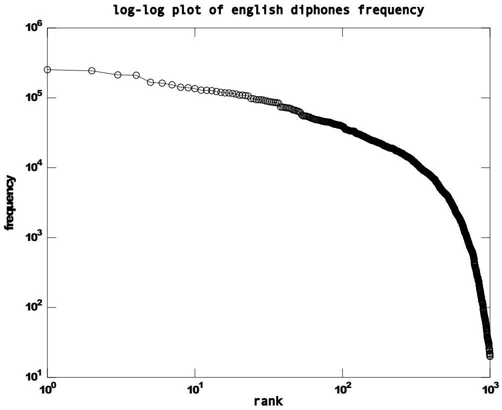


Figure: Log-log plot of the diphones frequency of occurrence versus their rank.

Diphones : phone-normalized frequency of occurrence

(1) ʒə
(2) ðə
(3) əv
(4) ʃə
(5) dʒə
(6) bə
(7) əl
(8) əf
(9) ju
(10) əm
(11) ɔr
(12) əb
(13) ətʃ
(14) nd
(15) kə
(16) əg
(17) əp
⋮
(1120) zs
(1121) uaɪ
(1122) aɪh
(1123) ɛoʊ
(1124) ddʒ
(1125) tv

Diphones Normalized Frequency - Log-log plot

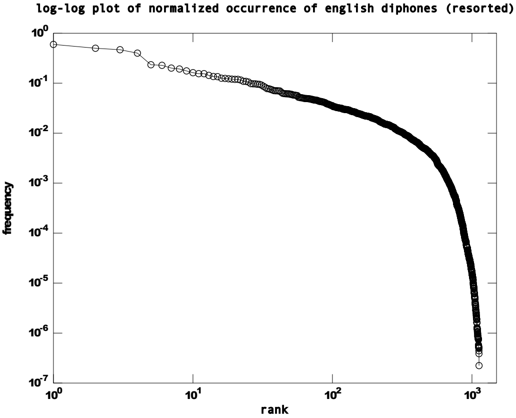


Figure: Log-log plot of the diphones normalized frequency of occurrence versus their rank. The normalization is made using the frequency of occurrence of each phone in the pair.

Diphones Normalized Frequency - Log-log plot

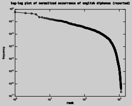


Figure: Log-log plot of the diphones normalized frequency of occurrence versus their rank. The normalization is made using the frequency of occurrence of each phone in the pair.

the normalization is total number of diphones divided by the occurrence of either one of the phones in the diphone pair

Diphones Conditional Probability

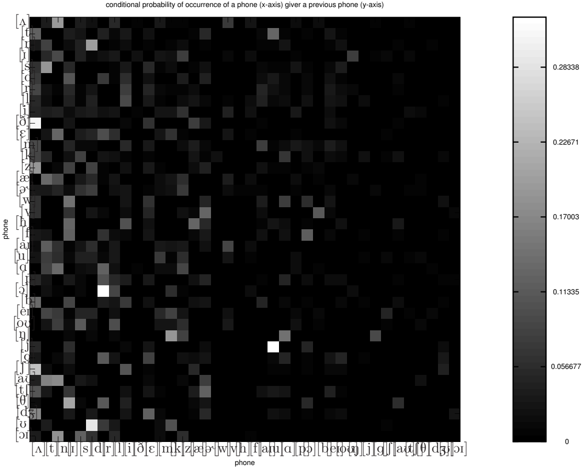
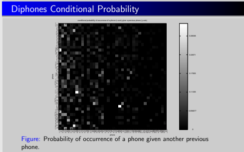


Figure: Probability of occurrence of a phone given another previous phone.



We observe in the figure that the most frequent phones (in the left part of the figure) also have an average higher conditional probability of occurrence regardless which the prior phone is.

(30,22) = [j u]
(25,7) = [ɔ r]
(37,6) = [ʊ d]
(10,1) = [ð ə]

Triphone Probabilities

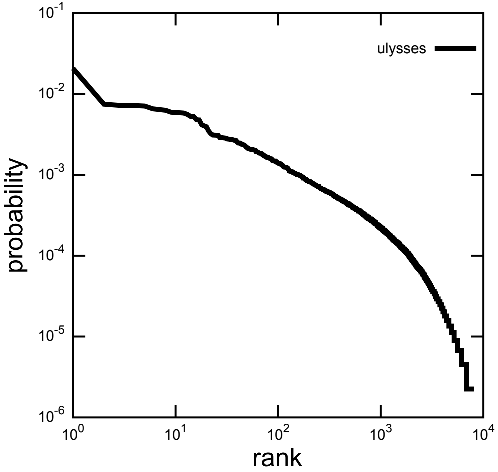


Figure: Triphone Probability (from *Ulysses*).

Syllable Probabilities

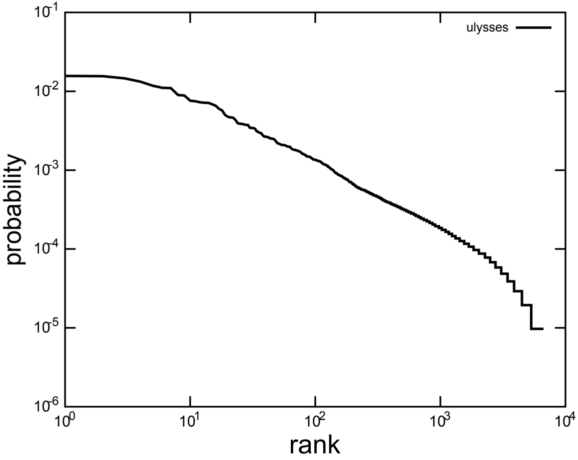


Figure: Syllable Probability (from *Ulysses*).

Words Length

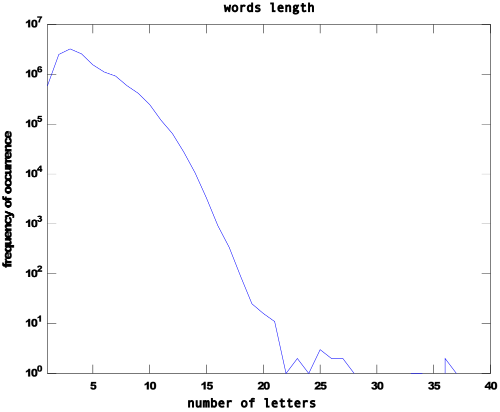


Figure: Frequency of occurrence of words of a given length (letters).

Words Phones Length

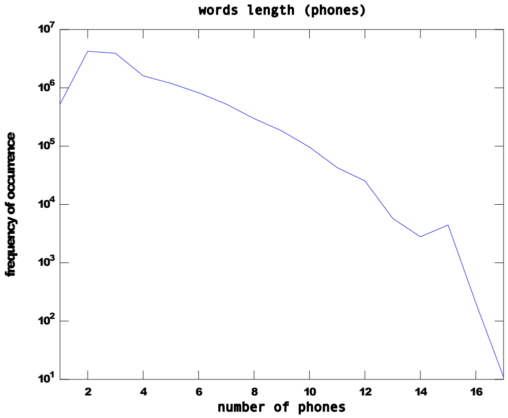


Figure: Frequency of occurrence of words of a given length (phones).

Words Phones Length

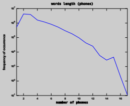


Figure: Frequency of occurrence of words of a given length (phones).

Fucks (e.g. 1955, 1956) demonstrated both theoretically and empirically that word length, measured in terms of the number of syllables, follows a displaced Poisson distribution.

Frequent discrepancies between the data and this model led GROTJAHN (1982) to consider the parameter of the Poisson distribution as a random variable following a gamma distribution. This approach yielded a displaced negative binomial distribution which provided a much better fit than the original Poisson distribution, being merely a limiting form of the negative binomial.

Words Phones Length

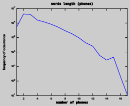


Figure: Frequency of occurrence of words of a given length (phones).

As long as the empirical data are consistent with the model there is no need for improvement; this is common practice in all sciences. However, since it can be shown that a great part of the published data does not fit the above models closely enough, it is advisable to reconsider the whole problem of modelling the distribution of word length.

Words Phones Length

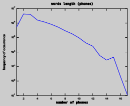


Figure: Frequency of occurrence of words of a given length (phones).

In addition to the purely linguistic difficulty of defining the concept of word and of delimiting wordlike units for example in texts, we encounter six practical and theoretical problems in modelling the distribution of word length:

(i) Up to now word length has been primarily measured in terms of letters, syllables and morphemes, although other possibilities are conceivable. This is the fundamental *problem of the unit of measurement*.

(ii) There are a number of factors which affect the distribution of word length in a specific population. This phenomenon will be referred to as the *population problem*.

(iii) Different boundary conditions might lead to different models and hence to different laws rather than one general law, which would be the ideal state of affairs. This problem will be referred to as the *modelling problem*.

Words Phones Length

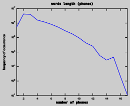


Figure: Frequency of occurrence of words of a given length (phones).

(iv) The usual criterion for the goodness of fit of a probabilistic model to the data is Pearson's chi-square test. This test – as well as others based on the divergence of the observed frequencies from the expected frequencies – is however strongly dependent on the sample size. In linguistics, where word length can be measured on the basis of texts, dictionaries etc., sample size is often so large that Pearson's chi-square test - as well as other power divergence tests – necessarily lead to a falsification of the model. This is the *goodness-of-fit problem*.

(v) Word length plays a role in the self-regulation of language and is thus partially dependent on other linguistic properties. This is the problem of the *interrelation- ship of linguistic properties*.

Words Phones Length

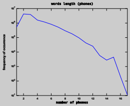


Figure: Frequency of occurrence of words of a given length (phones).

(vi) Knowledge of the interrelationship of word length with other linguistic phenomena as well as of the mechanisms involved in generating the word length patterns in various languages, helps us in establishing a valid explanation of the nature of word length. This is the *problem of explanation*.

(Grotjan and Altmann, 2007)

Average word length (letters)

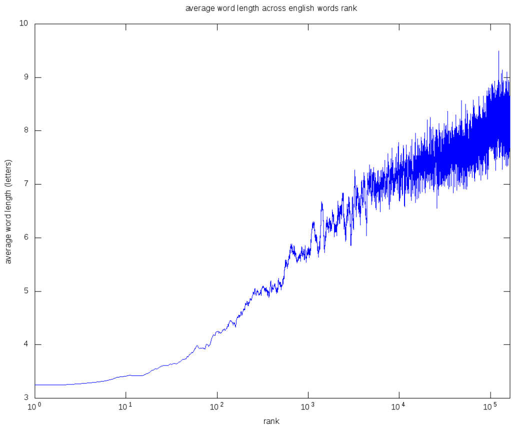


Figure: Average word length (letters) across word rank.

Average word length (phones)

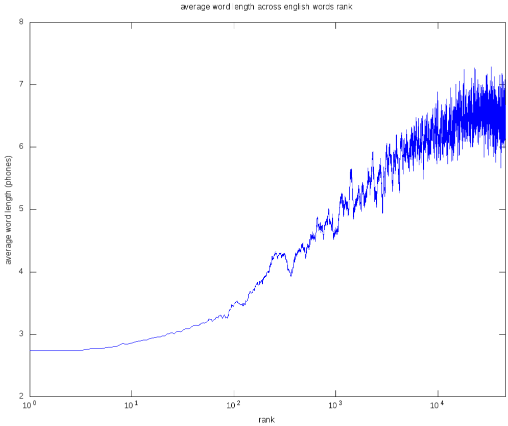


Figure: Average word length (phones) across word rank.

Frequency of occurrence and frequency index

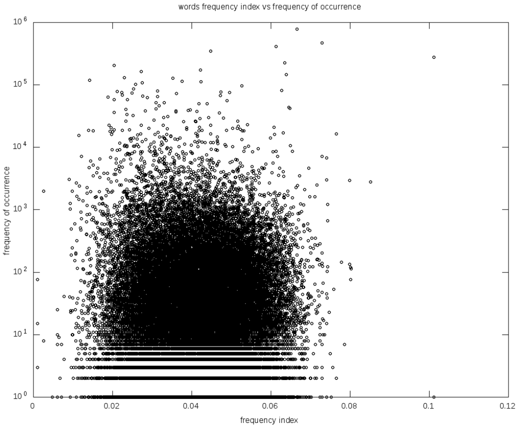


Figure: Each spot represents a word with a given frequency of occurrence and a frequency index. For a better visualization the frequency of occurrence is displayed in a logarithm scale.

Frequency of occurrence vs frequency index (density plot)

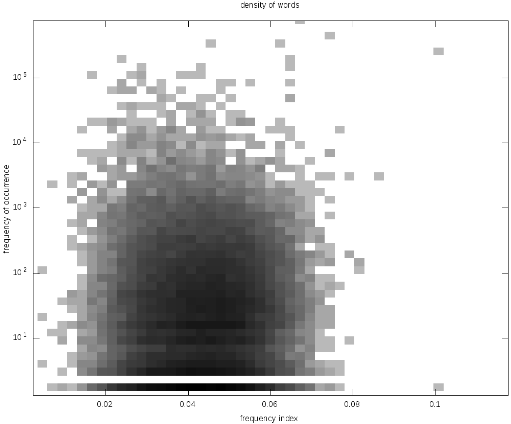


Figure: This picture is the density of words in each partition on the frequency of occurrence vs. frequency index space. The largest number is displayed in black and it refers to 578 words. White represents no word found in a spot.

Analysis of Smaller Units

Word is considered the smallest free form that can be uttered or written and carries a meaning (Bloomfield, 1926).

Basic units of speech perception:

- phones (Pisoni, 1982)
- diphones (Klatt, 1979)
- triphones (speech synthesis)
- syllables (Studdert-Kennedy, 1976)
- demisyllables (Fujimura and Lovins, 1978)

Distinctive Features

Figure: English Distinctive Features Table.

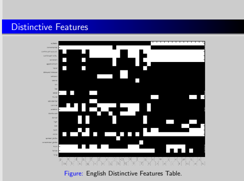


Figure: English Distinctive Features Table.

The theory proposes the existence of a small finite set of features that may be used to analyze speech sounds, in such way that the description of each speech sound in terms of these features is unique. The analysis into distinctive features is a unambiguous process to associate speech sounds into arrays of features that describe these sounds. The features proposed are a conjunction of perceptual and articulatory factors. The classical statement of distinctive features was brought by Chomsky and Halle (1968).

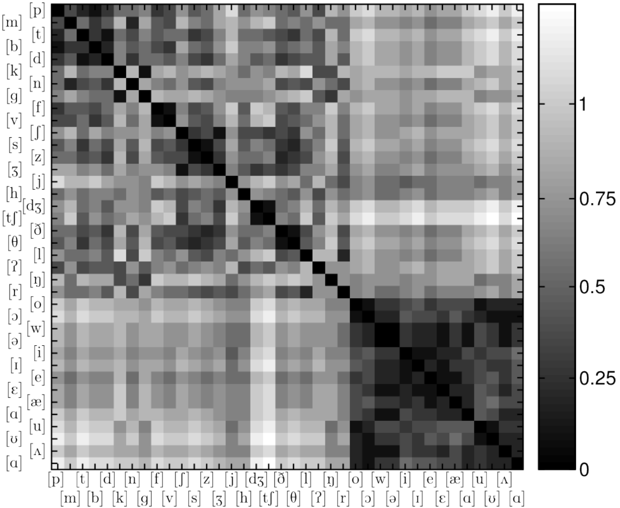
The most important contribution to phonology from the distinctive feature theory is that a set of segments may be analyzed into some features and it is possible to identify classes of segments in rules, creating the notion of natural class, as a set of sounds that has certain phonetic features in common and is affected in the same way in the same environment.

Distinctive Feature Distance

definition: distance measure between two segments

we define a distance measure of two segments as the number of features not shared between them (that dissimilarity definition resembles the natural class definition of Flemming (2005))

Dissimilarity Matrix



Multidimensional Scaling (MDS)

	cph	aar	ode	aal
cph	0	93	82	133
aar	93	0	52	60
ode	82	52	0	111
aal	133	60	111	0

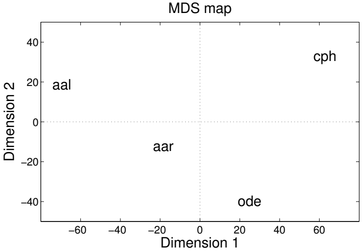


Figure: Simple Multidimensional Scaling Example.

MDS - Vowels

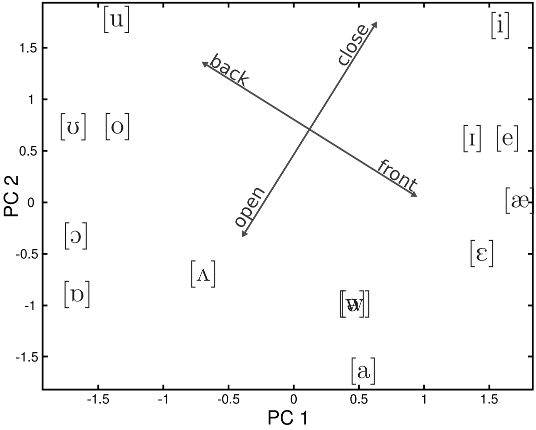
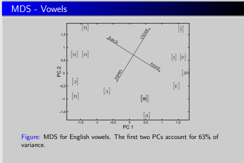


Figure: MDS for English vowels. The first two PCs account for 63% of variance.



Multidimensional scaling (MDS) is a means of visualizing the level of similarity of individual cases of a dataset. An MDS algorithm aims to place each object in N-dimensional space such that the between-object distances are preserved as well as possible. Each object is then assigned coordinates in each of the N dimensions. The number of dimensions of an MDS plot N can exceed 2 and are specified a priori. Choosing N=2 optimizes the object locations for a two-dimensional scatterplot.

MDS - Consonants

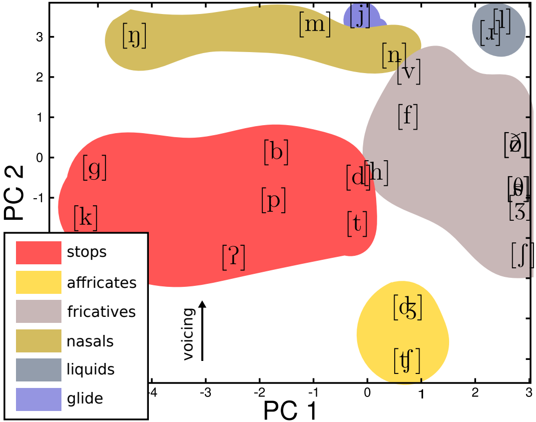


Figure: MDS for English consonants. The first two PCs account for 46% of variance.

Intradistances - triphones

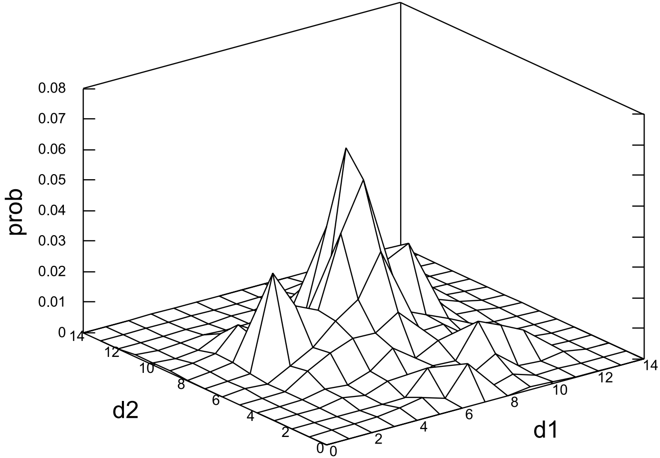


Figure: Probability of triphones given its intradistances.

Frequency of occurrence of words in logarithmic scale

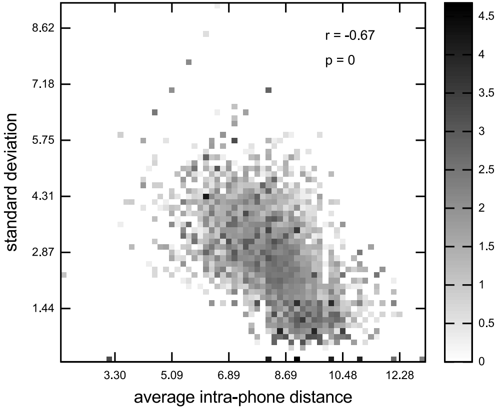


Figure: Relation in words between frequency of occurrence, average intra-phone distances and standard deviation of these distances.

Words Frequency and Zipf Law

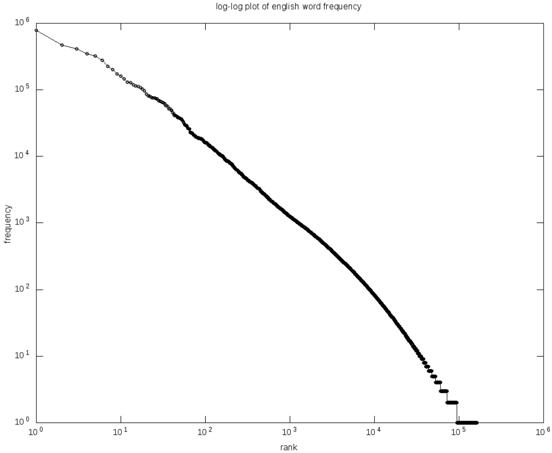


Figure: Log-log plot of words rank versus frequency of occurrence.

Zipf's Law

Zipf law states that there is a relationship between the word's frequency of appearance in texts and its rank, the product of them is roughly a constant.
Power law relation:

$$f(k; s, N) = Ck^{-s} = \frac{k^{-s}}{\sum_{n=1}^{N} n^{-s}} \tag{1}$$

f frequency

k rank

N number of elements in the set

s characterizing exponent

Zipf's Law

Zipf law states that there is a relationship between the word's frequency of appearance in texts and its rank, the product of them is roughly a constant.
Power law relation:

$$f(k; s, N) = Ck^{-s} = \frac{k^{-s}}{\sum_{n=1}^{N} n^{-s}} \tag{1}$$

- f frequency
- k rank
- N number of elements in the set
- s characterizing exponent

As already pointed out by Li (1992), "Zipf's law is not a deep law in natural language as one might first have thought. It is very much related the to particular representation one chooses, i.e., rank as the independent variable." Li (1992) showed that random texts also exhibit Zipf's law-like curves.

Zipf's Law

The normalizing constant C in Zipf's law might also be written as

$$C = \frac{1}{H_{N,s}} \tag{2}$$

where $H_{N,s}$ is known as the generalized harmonic number

$$H_{N,s} = \sum_{n=1}^{N} \frac{1}{n^s} \ . \tag{3}$$

(as $N \to \infty$, converges for $s > 1$)

Zipf's exponent

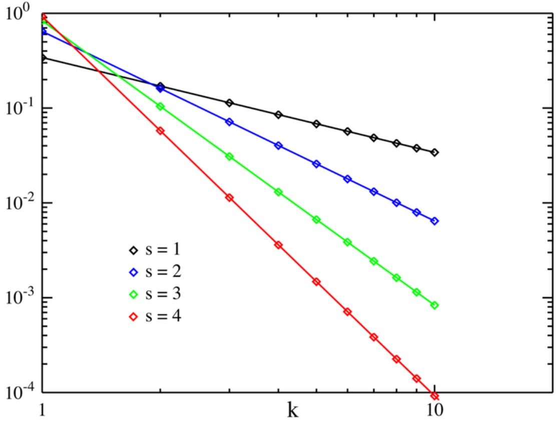


Figure: Probability mass function with $N = 10$. (Wikipedia)

The ubiquity of Zipf's law

This seemingly universal law is present in various phenomena:

(1) earthquakes
(2) avalanches
(3) populations
(4) firms
(5) stocks
(6) game of life
(7) lifespan of genera
(8) etc

Gutenberg-Richter law for earthquakes

Relation between magnitude m of a earthquake and the number N of earthquakes with magnitude greater or equal to m:

$$\log(N) = -bm + a \ . \quad (4)$$

$$m = \log_{10}\left(\frac{A}{A_0(\delta)}\right) \quad (5)$$

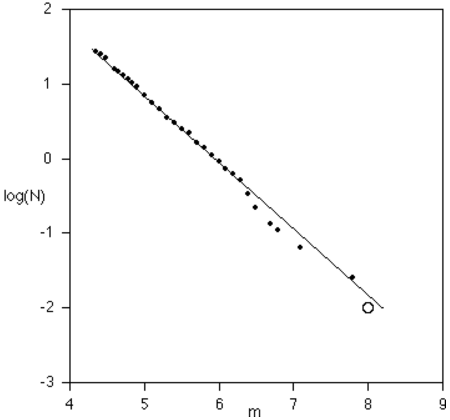


Figure: Data from the South of California from 1932 to 1972 (Turcotte, 1989).

Gutenberg-Richter law for earthquakes

Relation between magnitude m of a earthquake and the number N of earthquakes with magnitude greater or equal to m:

$$\log(N) = -bm + a \ . \quad (4)$$

$$m = \log_{10}\left(\frac{A}{A_0(\delta)}\right) \quad (5)$$

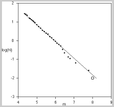


Figure: Data from the South of California from 1932 to 1972 (Turcotte, 1989).

The Richter magnitude of an earthquake is determined from the logarithm of the amplitude of waves recorded by seismographs (adjustments are included to compensate for the variation in the distance between the various seismographs and the epicenter of the earthquake). The original formula is: $M_{\mathrm{L}} = \log_{10} A - \log_{10} A_0(\delta) = \log_{10}[A/A_0(\delta)]$, where A is the maximum excursion of the Wood-Anderson seismograph, the empirical function A_0 depends only on the epicentral distance of the station, δ. In practice, readings from all observing stations are averaged after adjustment with station-specific corrections to obtain the M_L value.

Sandpile

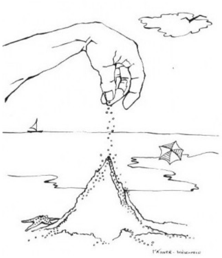


Figure: Sandpile avalanches (Bak, 1999).

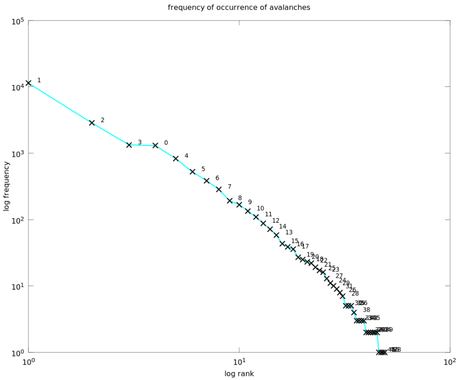
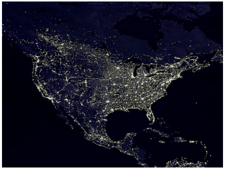


Figure: Frequency of avalanches.

Population



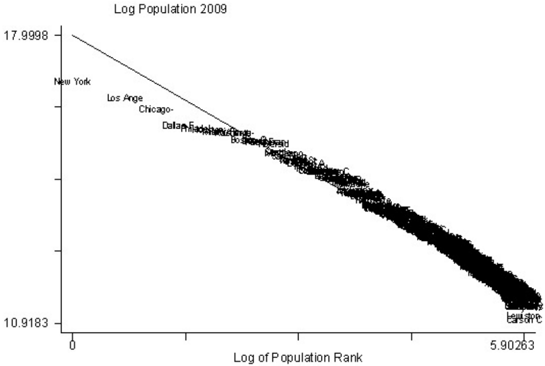
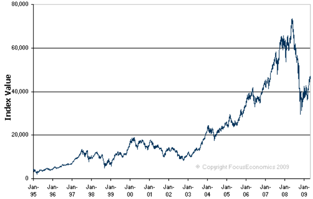


Figure: Population of American metropolitan areas (Edward L. Glaeser).

Stocks



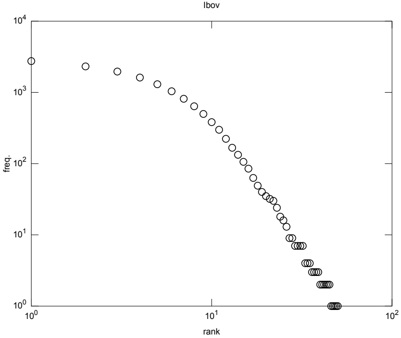


Figure: Variations on the Ibov index, using the idea presented by Mandelbrot (1963).

Game of Life

- cellular automata (Stanisław Ulam, John von Neumann, 1940)

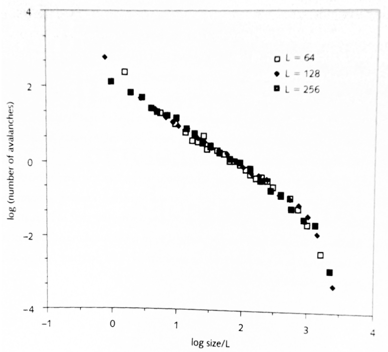
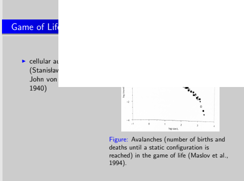
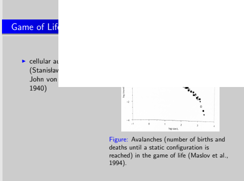


Figure: Avalanches (number of births and deaths until a static configuration is reached) in the game of life (Maslov et al., 1994).



A cellular automaton (pl. cellular automata) is a discrete model studied in computability theory, mathematics, physics, complexity science, theoretical biology and microstructure modeling. A cellular automaton consists of a regular grid of cells, each in one of a finite number of states, such as on and off. The grid can be in any finite number of dimensions. For each cell, a set of cells called its neighborhood is defined relative to the specified cell. An initial state (time t=0) is selected by assigning a state for each cell. A new generation is created (advancing t by 1), according to some fixed rule (generally, a mathematical function) that determines the new state of each cell in terms of the current state of the cell and the states of the cells in its neighborhood. Typically, the rule for updating the state of cells is the same for each cell and does not change over time, and is applied to the whole grid simultaneously.



The concept was originally discovered in the 1940s by Stanislaw Ulam and John von Neumann while they were contemporaries at Los Alamos National Laboratory. While studied by some throughout the 1950s and 1960s, it was not until the 1970s and Conway's Game of Life, a two-dimensional cellular automaton, that interest in the subject expanded beyond academia. In the 1980s, Stephen Wolfram engaged in a systematic study of one-dimensional cellular automata, or what he calls elementary cellular automata; his research assistant Matthew Cook showed that one of these rules is Turing-complete. Wolfram published A New Kind of Science in 2002, claiming that cellular automata have applications in many fields of science. These include computer processors and cryptography.

Lifespans of Genera

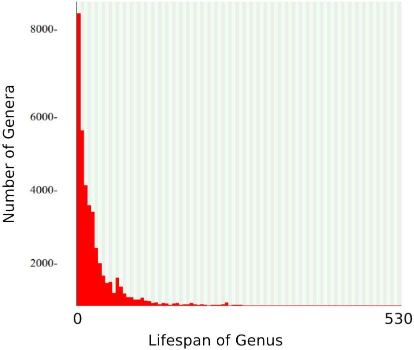


Figure: Distribution of length of lifespans of genera (in millions of years).

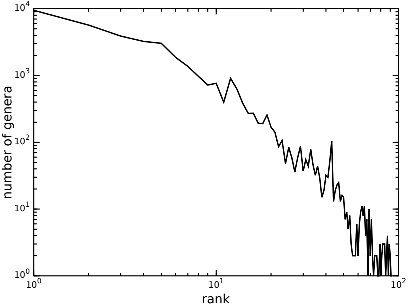


Figure: Loglog plot of the lifespans (data from Sepkoski Fossil Marine Genera Data Warehouse).

Random Texts

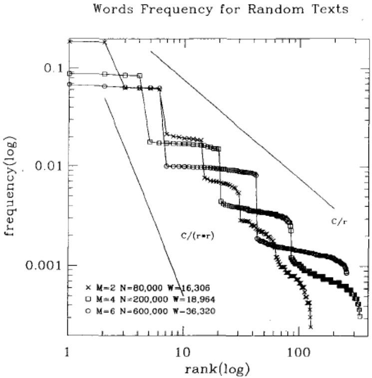


Figure: Word frequency vs *rank* for random generated symbols equally distributed. Symbol set size: $M = 2$, 4, and 6. Reference Zipf's law for $s = 1$ and 2 are displayed. (Li, 1992; Miller, 1957)

Random Text vs Ulysses

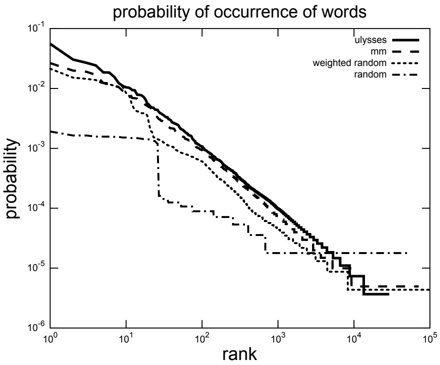


Figure: Random text and *Ulysses* compared.

Random Texts vs Ulysses: word length

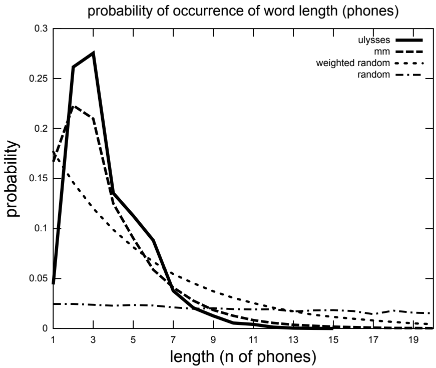


Figure: Word probability for a given length.

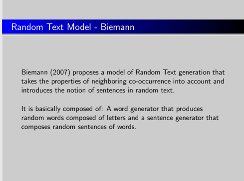
Markov Processes and Zipf's law

Kanter and Kessler (1995) shows that a 2-parameter random Markov process constructed with N states and biased random transitions gives rise to a stationary distribution where the probabilities of occurrence of the states exhibit a rank-ordered frequencies of occurrence of words given by Zipf's law.

Random Text Model - Biemann

Biemann (2007) proposes a model of Random Text generation that takes the properties of neighboring co-occurrence into account and introduces the notion of sentences in random text.

It is basically composed of: A word generator that produces random words composed of letters and a sentence generator that composes random sentences of words.



What distinguishes Biemann's proposal is that it is a hierarchical random text generator. In a higher level there is the random text generator that creates sentences, and in a lower level there is the random sentence generator which will create a random selection of words to built a sentence.

Random Text Model - Biemann

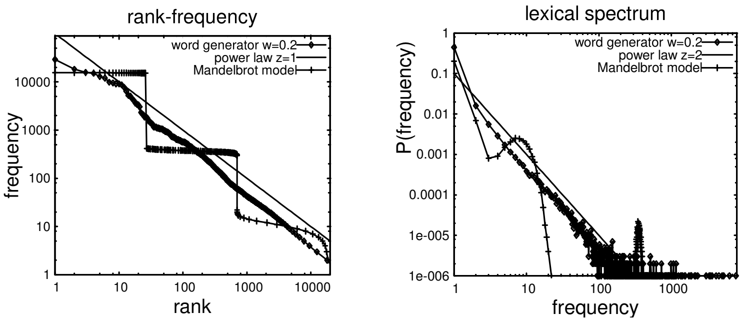


Figure: Rank-frequency distribution and lexical spectrum for the word generator in comparison to the Mandelbrot model (Biemann, 2007).

Random Text Model - Biemann

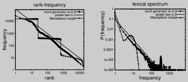


Figure: Rank-frequency distribution and lexical spectrum for the word generator in comparison to the Mandelbrot model (Biemann, 2007).

The creation of this hierarchy in the random text generation makes the resulted text much more similar to a true Zippfian distributed process. That is observed in the Zipf's plot and also on the lexical spectrum plot (inverse Zipf plot).

Zipf Exponent

The Zipf's exponent characterizes the source's distribution.

- natural phenomena $1 \leq s \leq 2$ (Baek et al., 2011)
- natural languages $s \approx 1$ (Piotrovskii et al., 1994)
- children's speech $s \approx 1.66$ (Piotrovskii et al., 1994)
- military combat text $s \approx 1.42$ (Kolguškin, 1960)
- adult and infant dolphins $s \approx 1.1$ and $s \approx 0.87$, respectively (McCowan et al., 1999)
- authorship attribution (Havlin, 1995)
- noncoding DNA sequence $s \approx 0.36$ and coding DNA sequence $s \approx 0.20$ (Mantegna et al., 1994)

Zipf Fit

Probability density function for a Zipf distribution:

$$f(k|s,N) = \frac{1/k^s}{\sum_{n=1}^{N} n^{-s}} = \frac{1/k^s}{H_{N,s}}, \tag{6}$$

Maximum likelihood estimation (MLE).

$$\begin{aligned} L(s|k_1,\cdots,k_M,N) &= f(k=(k_1,k_2,\cdots,k_M)|s,N) \\ &= \prod_{m=1}^{M} f(k_m|s,N) \\ &= \left(\frac{1}{H_{N,s}}\right)^M \prod_{m=1}^{M} \frac{1}{k_m^s}. \end{aligned} \tag{7}$$

Zipf Fit

The logarithm of the likelihood function:

$$\ln L(s|k_1, \cdots, k_M, N) = -M \ln H_{N,s} - s \sum_{m=1}^{M} \ln k_m. \tag{8}$$

Solution: the parameter s such that

$$\frac{\partial \ln L(s|k, N)}{\partial s} = 0 \ , \tag{9}$$

if

$$\frac{\partial^2 \ln L(s|k, N)}{\partial s^2} < 0 \ . \tag{10}$$

Zipf Fit

We need to solve

$$\frac{\partial \ln L(s|k_1,\cdots,k_M,N)}{\partial s} = -M\frac{G_{N,s}}{H_{N,s}} - \sum_{m=1}^{M} \ln k_m = 0 \ , \tag{11}$$

given that

$$\frac{\partial^2}{\partial s^2} \ln L(s|k_1,\cdots,k_M,N) = M\frac{G_{N,s}^2 - I_{N,s}H_{N,s}}{H_{N,s}^2} < 0 \ . \tag{12}$$

where

$$G_{N,s} = -\sum_{n=1}^{N} n^{-s} \ln n \ , \tag{13}$$

$$I_{N,s} = \sum_{n=1}^{N} n^{-s} \ln^2 n \ . \tag{14}$$

Zipf Fit

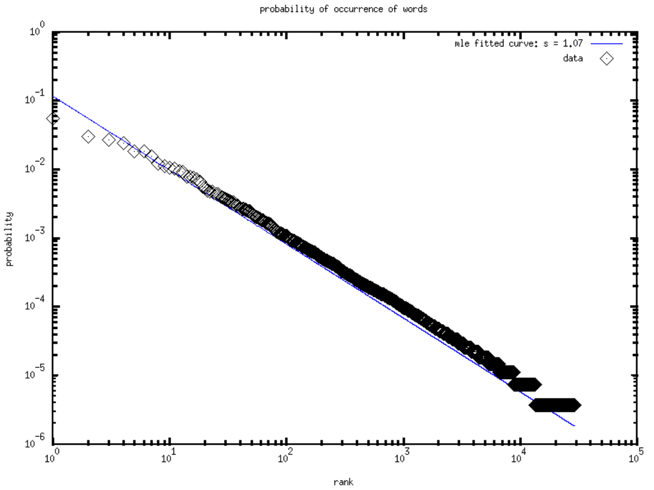


Figure: Zipf fit to *Ulysses* words ($s_{mle} = 1.0738$).

Zipf-Mandelbrot Fit

The Zipf-Mandelbrot distribution is given by

$$f(k|s,q,N) = \frac{1/(k+q)^s}{\sum_{n=1}^{N}(n+q)^{-s}} = \frac{1/(k+q)^s}{H_{N,s,q}} \ . \tag{15}$$

The logarithm of the likelihood is

$$\ln L(s|k_1,\cdots,k_M,N) = -M\ln H_{N,s,q} - s\sum_{m=1}^{M}\ln(k_m+q) \ . \tag{16}$$

Zipf-Mandelbrot Fit

Now we need to solve:

$$\frac{\partial \ln L(s|k_1, \cdots, k_M, N)}{\partial s} = -M\frac{G_{N,s,q}}{H_{N,s,q}} - \sum_{m=1}^{M} \ln(k_m + q) = 0 \ , \quad (17)$$

$$\frac{\partial^2 \ln L(s|k_1, \cdots, k_M, N)}{\partial s^2} = M\frac{G^2_{N,s,q} - I_{N,s,q}H_{N,s,q}}{H^2_{N,s,q}} < 0 \ . \quad (18)$$

$$\frac{\partial \ln L(s|k_1, \cdots, k_M, N)}{\partial q} = sM\frac{H_{N,s+1,q}}{H_{N,s,q}} - sH_{N,1,q} = 0 \ , \quad (19)$$

$$\begin{aligned}\frac{\partial^2 \ln L(s|k_1, \cdots, k_M, N)}{\partial q^2} \quad = \quad & s^2M\frac{H^2_{N,s+1,q}}{H^2_{N,s,q}} - s(s+1)M\frac{H_{N,s+2,q}}{H_{N,s,q}} + \\ & +sH_{N,2,q} < 0 \ . \qquad (20)\end{aligned}$$

Zipf-Mandelbrot Fit

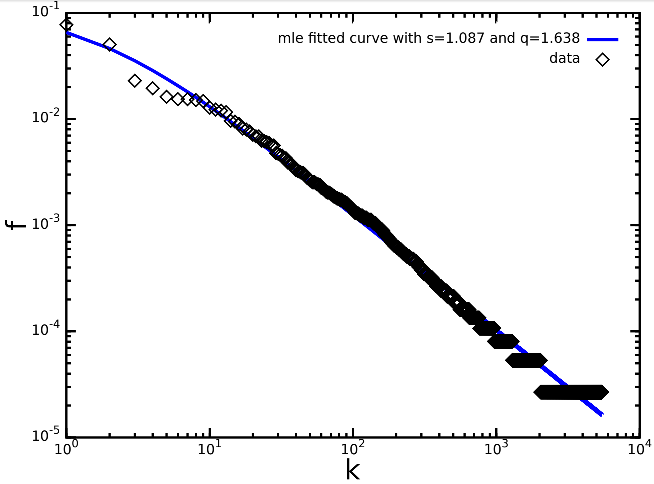


Figure: Zipf-Mandelbrot fit to *Hamlet* words ($s_{mle} = 1.087$ $q_{mle} = 1.638$).

Lexical Spectrum or Inverse Zipf

"Understanding the origins and evolution of language requires an appropriate identification of its universal features. One of the most obvious is the statistical distribution of word abundances. The second form is called the lexical spectrum or the inverse Zipf's distribution"(Ferrer-i-Cancho and Solé, 2002).

Inverse Zipf

The number of words for a given frequency of occurrence

$$N(f) = af^{-\beta} \tag{21}$$

where

$N(f)$	number of words with a given frequency f
f	frequency of occurrence
a and β	parameters

The exponents are related by

$$\beta = \frac{1}{s} + 1\ . \tag{22}$$

Inverse Zipf

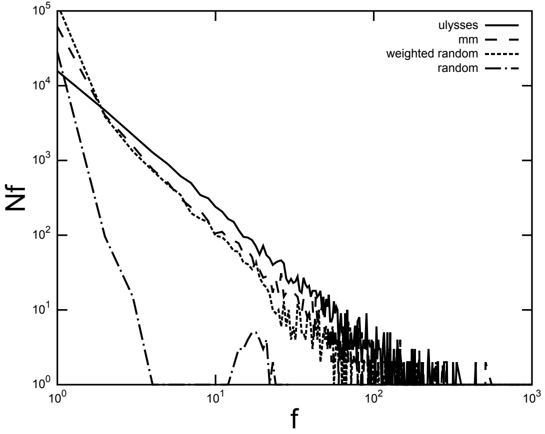


Figure: Inverse Zipf. Natural and Artificial Texts compared.

Smoothing

The maximum likelihood estimator predicts that the probability of a word not seen in the corpus is zero.

Smoothing is used to overcome this problem.

(1) *Add-one* estimator (Laplace, 1902)
(2) Good-Turing estimator (Good, 1953)
(3) Simple Good-Turing estimator (Gale and Sampson, 1995)

Good-Turing

The Good-Turing method states that $N_0 = N_1$ (the total probability of all unseen events is equal to the sum of probabilities of all events that occur only once) and the number of observations are adjusted by

$$f^* = (f+1)\frac{E[N_{f+1}]}{E[N_f]} \tag{23}$$

- $E[\cdot]$ expectation of a random variable
- f^* adjusted number of observations
- f observed frequency of occurrence
- N_f number of types observed f times

Simple Good-Turing

Gale and Sampson (1995) proposes a simple way to do a Good-Turing estimation by choosing $E[\cdot]$ so that

$$E[N_{f+1}] = E[N_f]\left(\frac{f}{f+1}\right)\left(1 - \frac{E[N_1]}{N}\right) \tag{24}$$

leading to

$$p_f^* = p_f\left(1 - \frac{E[N_1]}{N}\right) \ . \tag{25}$$

Simple Good-Turing

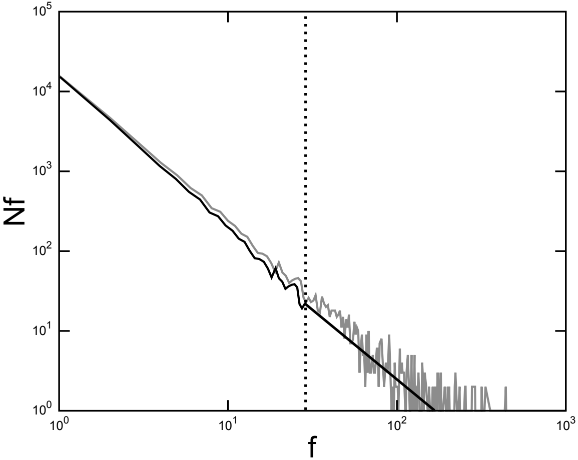


Figure: Simple Good-Turing.

How Zipf's law emerges

According to Ferrer-i-Cancho and Solé (2003); Ferrer-i-Cancho (2005a), Zipf's law is a manifestation of a complex system operating between order and disorder.

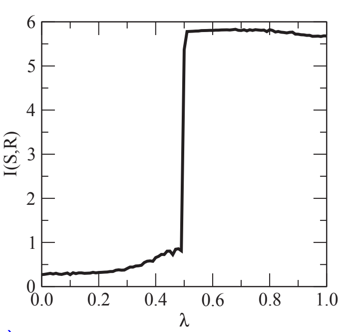
A communication system should minimize Ω

$$\Omega(\lambda) = -\lambda I(S,R) + (1-\lambda)H(S)\ , \tag{26}$$

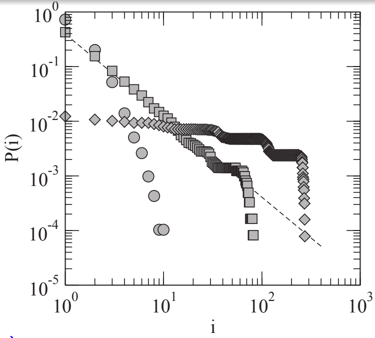
where

- S set of signs (words)
- R set of estimula
- I(S,R) transmitted information
- H(S) cost (entropy of words)

How Zipf's law emerges



(a) $I(S,R)$, the information transfer between words and meanings, versus λ, the parameter regulating the balance between maximizing $I(S,R)$ and minimizing the entropy of words.



(b) $P(i)$, the probability of the i-th most likely word in the system for $\lambda = 0.49$ (circles), $\lambda = 0.498$ (squares) and $\lambda = 0.5$ (diamonds). The dashed line contains the theoretical curve for $\lambda = 0.498$.

Figure: Some computational results on the model where meaning probabilities are governed by the internal structure of the communication system. The size of the system is $n = m = 400$ (i.e. 400 words and meanings). Figure reproduced from Ferrer-i-Cancho (2005b).

Entropy of a Zipfian Source

Entropy of a Zipfian distributed source:

$$\begin{aligned}\overline{H} &= -\frac{1}{\ln 2}\sum_{k=1}^{N} Ck^{-s}\ln(Ck^{-s}) \\ &= \frac{sC}{\ln 2}\sum_{k=1}^{N}\frac{\ln k}{k^s} - \frac{\ln C}{\ln 2}\,. \end{aligned} \tag{27}$$

We propose lower and upper bounds to estimate the entropy

$$B_l \leq \sum_{k=1}^{N} k^{-s}\ln k \leq B_u. \tag{28}$$

Entropy of a Zipfian Source

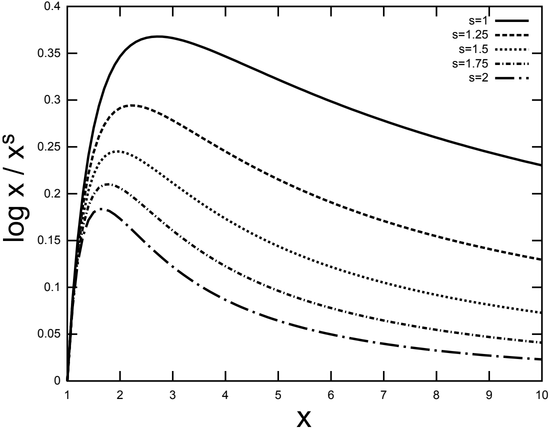


Figure: Function $f(x) = \ln x / x^s$ for different values of s.

Entropy of a Zipfian Source

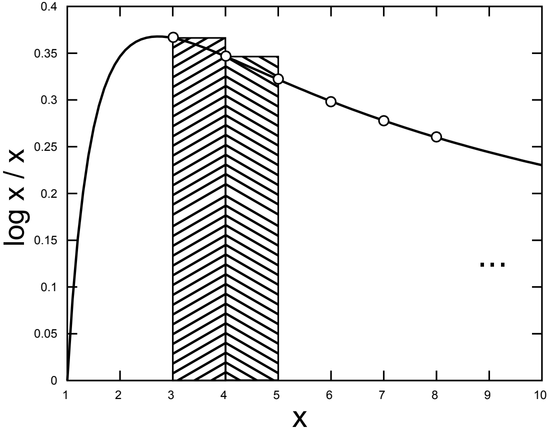


Figure: Left Riemann sum approximation of the integral.

Entropy of a Zipfian Source

The bounds are given by

$$
\begin{aligned}
B_l &= \int_3^N \frac{\ln x}{x^s} dx + \frac{\ln 2}{2^s} + \frac{\ln N}{N^s} \\
&\leq \sum_{n=1}^{N} \frac{\ln n}{n^s} \\
&\leq \int_3^{N-1} \frac{\ln x}{x^s} dx + \frac{\ln 3}{3^s} + \frac{\ln 2}{2^s} + \frac{\ln N}{N^s} = B_u \ , \qquad (29)
\end{aligned}
$$

and the estimated Entropy is bounded by

$$
\frac{sC}{\ln 2} B_l - \frac{\ln C}{\ln 2} = H_l \leq \overline{H} \leq H_u = \frac{sC}{\ln 2} B_u - \frac{\ln C}{\ln 2} \ . \qquad (30)
$$

Entropy of a Zipfian Source

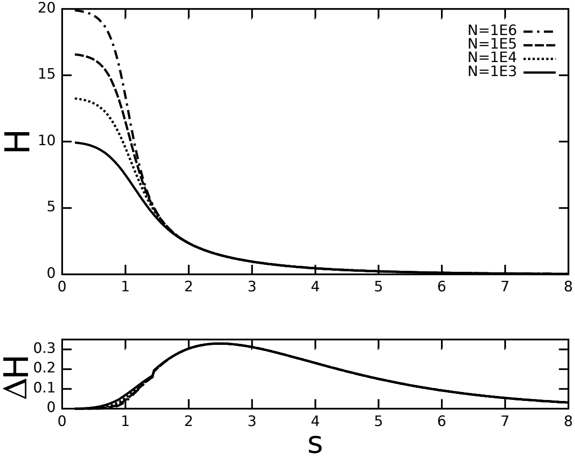


Figure: Entropy H (in bits) as a function of the Zipf exponent s and the number of types N. The upper plot presents the average Entropy estimated and the lower plot presents the difference between the upper and lower bounds of the entropy estimated.

Entropy of a Zipfian Source

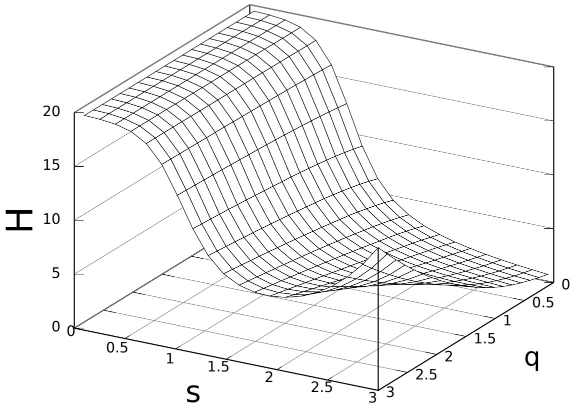


Figure: Effect of the parameters s and q on the entropy H (in bits) for a Zipf-Mandelbrot distribution.

Entropy of a Zipfian Source

Table: Entropy of real texts (bits), with and without SGT smoothing, compared with the estimated entropy (bits) using the parameter N (number of types) found in the text, parameter s (Zipf exponent) found by a Maximum Likelihood Estimation (MLE) and the flattening parameter q, also found by MLE.

source	N	estimated parameters			entropy		estimated entropy	
		Zipf	Zipf-Mandelbrot		normal	sgt	Zipf	Zipf-Mandelbrot
		s	s	q				
Alice	3016	0.992	1.172	3.27	8.49	8.79	8.55	8.73
Hamlet	5447	0.991	1.087	1.64	9.04	9.08	9.09	9.13
Macbeth	4017	0.969	1.009	0.56	9.00	9.00	9.02	9.04
Shakespeare	29847	1.060	1.172	2.33	9.52	9.57	9.60	9.69
Ulysses	34391	1.025	1.085	1.18	10.19	10.25	10.22	10.25

Heaps' law (Herdan's law)

There is a relation between

T text size

V_T vocabulary length

given by

$$V_T \propto T^\alpha \ , \quad \alpha < 1 \ . \tag{31}$$

Leijenhorst et al. (2005) shows that a Zipfian distributed source presents this relation between text size and vocabulary length.

Heaps' law (Herdan's law)

The expected number of types is monotonically increasing with the sample size. It converges to the length of the underlying lexicon.

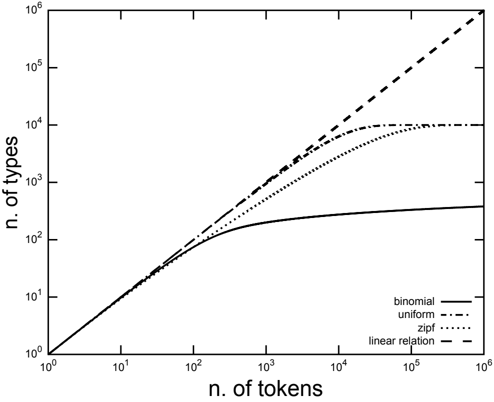


Figure: The recurrence equation is used to estimate the expected number of types for a sample with a certain number of tokens.

Heaps' law (Herdan's law)

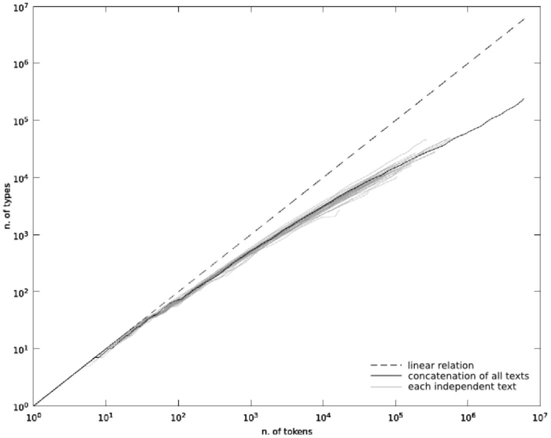


Figure: The relation between the number of tokens and types in 35 books from Gutenberg Database is presented in gray.

Menzerath's law

The longer the whole, the smaller the parts.
Menzerath (1928) observed the decreasing relation between syllable's length and length of a word (number of syllables).

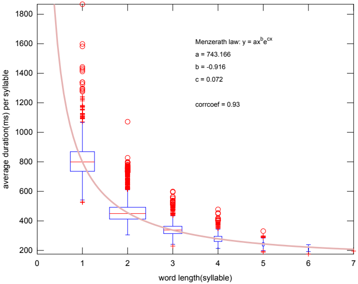
The change relative rate on the length of the constituent is inversely proportional to the length of the construct.

$$\frac{y'}{y} = \frac{b}{x}. \tag{32}$$

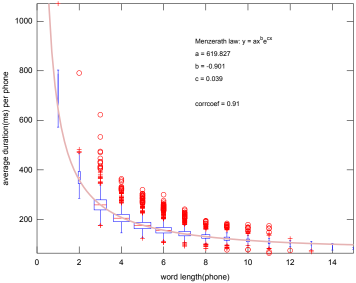
- y constituent length
- x construct length

Solution: $y = ax^b$.

Menzerath's law



(a) Average syllable duration as words get longer (number of syllables).



(b) Average phone duration as words get longer (number of phones).

Menzerath's law

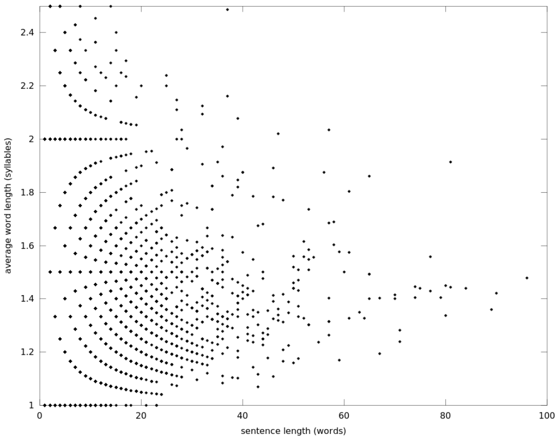


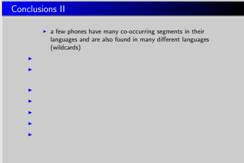
Figure: Relation between sentence length (number of words) and average words length (number of syllables) in a sentence.

Conclusions I

- usage-based driven language (Bybee, 2001, 2003, 2006; Ellis, 2002)
- humans track co-occurrence patterns and statistical regularities (Saffran et al., 1996a, 1999; Saffran and Wilson, 2003)
- regularities among several languages
  - some speech segments are very common (example: [m] appear in 94.2% of the languages)
  - 427 out of 919 speech sounds appear just once among all 425 languages
  - 90% of the languages use at most 6 times more consonants than vowels

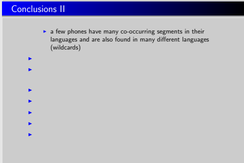
Conclusions II

  - a few phones have many co-occurring segments in their languages and are also found in many different languages (wildcards)
- ubiquity of scaling laws in nature
- a compared study of ‘intermittent silence’, Markov model and natural text
- compare different sources using the Zipf-Maldelbrot model
- mle method to fit the best model
- entropy estimation estimation of a Zipf-Maldelbrot source
- Menzerath’s law and Heap’s law
- distance measure between phones using Feature Theory

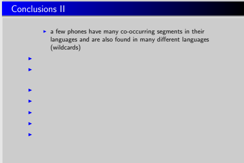


It is important to note that the 425 languages in the UPSID are chosen so that they represent well all family branches in the language tree. We adopt here a usage-based theory of language in which its cognitive organization is based directly on one's experience with it, rather than being an abstract set of rules or structures that are only indirectly related to experience with language. We see grammar as a network built up from the categorized instances of language use.

Since the grammar is based on usage, it contains many details of co-occurrence as well as a record of the probabilities of occurrence and co-occurrence. The evidence for the impact of usage on cognitive organization includes the fact that language users are aware of specific instances of constructions that are conventionalized and the multiple ways in which frequency of use has an impact on structure.



The latter include speed of access related to token frequency, priming, morphological and phonetic properties of high and low frequency words, as well as the importance of usage to the process of grammaticalization. (In linguistics, grammaticalization is a process by which words representing objects and actions (i.e. nouns and verbs) transform through sound change and language migration to become grammatical markers (affixes, prepositions, etc.).)



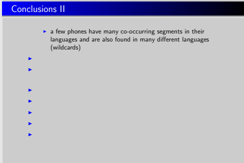
A number of recent experimental studies (Saffran et al., 1996b, 1999; Saffran and Wilson, 2003) show that both infants and adults track co-occurrence patterns and statistical regularities in artificial grammars. Such studies indicate that subjects learn patterns even when the grammar corresponds to no meaning or communicative intentions. It is thus not surprising that in actual communicative settings, the co-occurrence of words has an impact on cognitive representation. There is indeed evidence from multiple sources that such cognitive changes occur, and contribute to the shape of grammar.

The cognitive representations underlying language use are built up by the categorization of utterances into exemplars and exemplar clusters based on their linguistic form as well as their meaning and the context in which they have been experienced (Pierrehumbert, 2001).

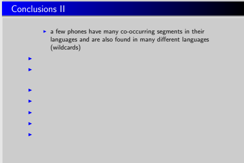
The fact that proposed 'universals of grammar' appear to be highly general can be seen as a by-product of this process: grammars are emergent phenomena dependent on idiosyncratic localized interactions.

- a few phones have many co-occurring segments in their languages and are also found in many different languages (wildcards)

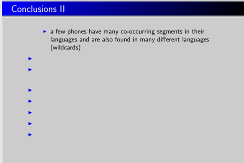
The ubiquity of scaling laws in nature suggests that complex systems organize themselves, they are not just a collection of random independent realizations, but exhibit similar patterns and relations that are observed in different scales. Complex systems are more than mere collections of independent random variables, they have interdependent effects and behaviours.



Only the Markov model was able to result in similar patterns to those observed in natural texts. Frequency of 1-gram, 2-grams, 3-grams, ‘word’ length, occurrence of hapax legomenon, Zipf and inverse Zipf (lexical spectrum) relations.

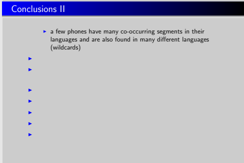


The exponent parameter in the Zipf distribution is important to determine the statistical properties of that distribution. We observe that different type of sources present different exponent values and we argue that this exponent is related to the balance between order and disorder in a Complex System, the amount of information produced by the source and the underlying lexicon size (and the possibility of being virtually infinite). In order to compare data that follow a power law relation, we need to find the best fit to the data. We have presented the usage of the maximum likelihood approach to determine the exponent parameters that gives the best fit. We have derived the equations which lead to a root finding problem, for which the solution is the maximum likelihood estimated parameter. The same procedure is used to find the exponent and flattening parameters of a Zipf-Mandelbrot model. This procedure was tested using real text data and the results show a good performance.



Menzerath's law states a relation on the usage of tiles to create a whole. According to this law, the length of the whole is inversely proportional to the length of the parts.

Heap's law: sublinear relation between the text length and the vocabulary size. Zipf's law implies Heap's law, but any distribution would lead to a convergent sublinear lexical growth in relation to the text length.



The Feature Theory might be providential since it give us a direct and formal way to quantify a phone characteristics making it possible to create a distance measure between two phones. We propose to measure the distance between two phones by the number of distinctive features not shared by them. We believe this measure might be a meaningful choice. It could be used alongside frequency of occurrence to draw better explanations to phonological process and cognitive representation.

Outline

Abe, S. and Suzuki, N. (2005). Scale-free statistics of time interval between successive earthquakes. *Physica A: Statistical Mechanics and its Applications*, 350(2-4):588–596.

Adamic, L. A. and Huberman, B. A. (2002). Zipf's law and the internet. *Glottometrics*, 3:143–150.

Aldrich, J. (1997). R. A. Fisher and the Making of Maximum Likelihood 1912-1922. *Statistical Science*, 12(3):162–176.

Alexander, L., Johnson, R., and Weiss, J. (1998). Exploring zipf's law. *Teaching Mathematics Applications*, 17(4):155–158.

Altmann, G. (1980). Prolegomena to menzerath's law. *Glottometrika*, 2:1–10.

Altmann, G. and Arens, H. (1983). *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik. Festschrift für Peter Hartmann*, chapter ,,Verborgene Ordnung" und das Menzerathsche Gesetz, pages 31–39. Narr, Tübingen.

Altmann, G., Schwibbe, M., and Kaumanns, W. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Olms.

Andrews, S. (1989). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(5):802–814.

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., and Yamada, T. (2004). Perceived phonetic dissimilarity and l2 speech learning: the case of japanese /r/ and english /l/ and /r/. *Journal of Phonetics*, 32:233–250.

Apostel, L., Mandelbrot, B., and Morf, A. (1957). *Logique, Langage et Théorie de L'Information*. Presses Universitaires de France, Paris.

Baayen, R. H. (2001). *Word Frequency Distributions*. Kluwer, Dordrecht.

Baayen, R. H. (2003). *Probabilistic Linguistics*, chapter Probabilistic Approaches to Morphology. MIT Press.

Baddeley, A. D., Thomson, N., and Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 14(6):575–589.

Baek, S. K., Bernhardsson, S., and Minnhagen, P. (2011). Zipf's law unzipped. *New Journal of Physics*, 13(4).

Bak, P. (1999). *How Nature Works: The Science of Self-organized Criticality*. Copernicus Series. Springer.

Balasubrahmanyana, V. and Naranan, S. (1996). Quantitative linguistics and complex system studies. *Journal of Quantitative Linguistics*, 3(3):177–228.

Biemann, C. (2007). A random text model for the generation of statistical language invariants. In *Proceedings of HLT-NAACL-07*, Rochester, NY, USA.

Bisani, M. and Ney, H. (2003). Multigram-based grapheme-to-phoneme conversion for lvcsr. In *In Proc. Eurospeech*, pages 933–936.

Bloomfield, L. (1926). A set of postulates for the science of language. *Language 2*.

Boas, F. (1911). *Handbook of American Indian languages*. Government PrintingOffice.

Bod, R., Hay, J., and Jannedy, S. (2003). *Probabilistic Linguistics*. MIT Press.

Bond, Z. S. and Garnes, S. (1980). *Perception and Production of Fluent Speech*, chapter Misperception of fluent speech. Erlbaum, N.J.

Bornstein, M. H. and Korda, N. O. (1984). Discrimination and matching within and between hues measured by reaction times: some implications for categorical perception and levels of information processing. *Psychological research*, 46(3):207–222.

Brent, M. R. and Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61:93–125.

Brown, G. D. A. and Neath, I. (2007). A temporal ratio model of memory. *Psychological Review*, 114(3):539–576.

Buk, S. and Rovenchak, A. A. (2007). Menzerath-altmann law for syntactic structures in ukrainian. *CoRR*.

Bybee, J. L. (2001). *Phonology and Language Use*. Cambridge University Press.

Bybee, J. L. (2003). *The evolution of language out of pre-language*, chapter Sequentiality as the basis of constituent structure, pages 109–132. John Benjamins.

Bybee, J. L. (2006). *Frequency of Use And the Organization of Language*. Oxford University Press.

Capek, M. J. (1983). Phoneme theory and umlaut: A note on the creation of knowledge. *Monatshefte*, 75(2):126–130.

Charoenpornsawat, P. and Schultz, T. (2006). Example-based grapheme-to-phoneme conversion for thai. In *INTERSPEECH 2006*.

Chen, S. F. and Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. Technical report, Computer Science Group, Harvard University.

Chomsky, N. (1957). *Syntactic Structures*. Mouton.

Chomsky, N. (1968). *Language and Mind*. Harcourt, Brace and World, New York.

Chomsky, N. (1969). *Aspects of the Theory of Syntax*. The MIT Press.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. Harper & Row.

Church, K. W. and Gale, W. A. (1991). A comparison of the enhanced good-turing and deleted estimation methods for estimating probabilities of english bigrams. *Computer Speech and Language*, 5:19–54.

Cole, R. A. and Jakimik, J. (1980). *Perception and Production of Fluent Speech*, chapter A model of speech perception, pages 133–163. Lawrence Erlbaum Associates, Hillsdale, NJ.

Coleman, J. (2002). *Phonetics, phonology and cognition*, chapter Phonetic representations in the mental lexicon, pages 96–130. Oxford University Press.

Copelli, M. and Campos, P. R. A. (2007). Excitable scale free networks. *The European Physical Journal B: Condensed Matter and Complex Systems*, 56:273–278.

Coulmas, F. (2003). *Writing Systems: An introduction to their linguistic analysis*. Cambridge University Press.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley-Interscience, New York, USA.

Cristelli, M., Batty, M., and Pietronero, L. (2012). There is more than a power law in Zipf. *Scientific Reports*, 2.

Cristófaro-Silva, T., de Almeida, L. S., and Fraga, T. (2005). Aspa: a formulacao de um banco de dados de referência da estrutura sonora do português contemporâneo. In *XXV Congresso da Sociedade Brasileira de Computação*, pages 2268–2277.

Davidson, D. (2005). *Truth, Language, and History*. Oxford University Press.

de Boer, B. (2000). Self-organization in vowel systems. *Journal of Phonetics*, 28(4):441–465.

de Saussure, F. (1916). *Cours de linguistique générale*. Payot.

Deacon, T. (1997). *The Symbolic Species: The Co-Evolution of Language and the Brain*. Science: Linguistics. W.W. Norton.

Disner, S. F. (1983). *Vowel Quality: The relation between Universal and Language-specific Factors*. PhD thesis, UCLA: Department of Linguistics.

Dresher, B. E. (2011). *The Blackwell Companion to Phonology*, volume 1, chapter The Phoneme, pages 241–266. John Wiley & Sons.

Düring, B., Matthes, D., and Toscani, G. (2008). A boltzmann-type approach to the formation of wealth distribution curves. Technical Report 08-05, Center of Finance and Econometrics, University of Konstanz.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24(2):143–188.

Etcoff, N. L. and Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, 44:227–240.

Everett, D. (2005). Cultural constraints on grammar and cognition in pirahã: Another look at the design features of human language. *Current Anthropology*, 46:621–646.

Ferrer-i-Cancho, R. (2005a). The variation of zipf's law in human language. *European Physical Journal B*, 44(2):249–257.

Ferrer-i-Cancho, R. (2005b). The variation of zipf's law in human language. *European Physical Journal B*, 44(2):249–257.

Ferrer-i-Cancho, R. (2006). *Exact methods in the study of language and text. In honor of Gabriel Altmann*, chapter On the universality of Zipf's law for word frequencies, pages 131–140. Gruyter.

Ferrer-i-Cancho, R. and Elvevåg, B. (2010). Random texts do not exhibit the real zipf's law-like rank distribution. *PLoS ONE*, 5(3):e9411+.

Ferrer-i-Cancho, R. and Gavaldà, R. (2009). The frequency spectrum of finite samples from the intermittent silence process. *Journal of the American Society for Information Science and Technology*, 60(4):837–843.

Ferrer-i-Cancho, R. and Solé, R. V. (2002). Zipf's law and random texts. *Advances in Complex Systems*, 5(1):1–6.

Ferrer-i-Cancho, R. and Solé, R. V. (2003). Least effort and the origins of scaling in human language. *PNAS*, 100(3):788–791.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, 222:309–368.

Flemming, E. (2005). Deriving natural classes in phonology. *Lingua*, 115(3):287–309.

Frank, M. C., Everett, D. L., Fedorenko, E., and Gibson, E. (2008). Number as a cognitive technology: Evidence from pirahã language and cognition. *Cognition*.

Freedman, D. A. and Wang, W. (1996). Language polygenesis: A probabilistic model. *Anthropological Science*, 104(2):131–137.

Frisch, S. (1996). *Similarity and frequency in phonology*. PhD thesis, Northwestern University.

Fujimura, O. and Lovins, J. (1978). *Syllables and segments*, chapter Syllables as concatenative phonetic elements, pages 107–120. North Holland, Amsterdam.

Fujiwara, Y. (2004). Zipf law in firms bankruptcy. *Physica A: Statistical and Theoretical Physics*, 337(1-2):219–230.

Fukada, T. and Sagisaka, Y. (1997). Automatic generation of a pronunciation dictionary based on a pronunciation network. In *European Conference on Speech Communication and Technology (EuroSpeech'97*, pages 2471–2474.

Gabaix, X. (1999). Zipf's law for cities: An explanation. *Quarterly Journal of Economics*, 114(3):739–67.

Gale, W. (1994). Good-turing smoothing without tears. *Journal of Quantitative Linguistics*, 2.

Gale, W. A. and Church, K. W. (1994). What's wrong with adding one. In *Corpus-Based Research into Language. Rodolpi*.

Gale, W. A. and Sampson, G. (1995). Good-turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2:217–237.

Gernsbacher, M. A. (1994). *Handbook of Psycholinguistics*. Academic Press.

Givón, T. (1979). *On understanding grammar*. Academic Press.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264.

Good, J. (2008). *Linguistic Universals and Language Change*. Oxford University Press.

Grainger, J. (1990). Word Frequency and Neighborhood Frequency Effects in Lexical Decision and Naming. *Journal of Memory and Language*, 29:228–244.

Greenberg, J. H., Osgood, C. E., and Jenkins, J. J. (1966). *Universals of Language*, chapter Memorandum concerning language universals. M.I.T. Press.

Gregory, M. L., Raymond, W. D., Bell, A., Fosler-lussier, E., and Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production.

Grignetti, M. (1964). A note on the entropy of words in printed english. *Information and Control*, 7:304–306.

Grotjan, R. and Altmann, G. (2007). *Contributions to Quantitative Linguistics*, chapter Modelling the Distribution of Word Length: Some Methodological Problems, pages 141–153. Springer Publishing Company, Incorporated.

Grzybek, P., Stadlober, E., and Kelih, E. (2006). The relationship of word length and sentence length: The inter-textual perspective. pages 611–618.

Hagège, C. (1986). *La Structure des langues*. Presses universitaires de France.

Halle, M. (1985). Speculations about the representation of words in memory. pages 101–114, Orlando, Florida. Academic Press.

Hancock, J. T., Curry, L. E., Goorha, S., and Woodworth, M. T. (2004). Lies in conversation: An examination of deception using automated linguistic analysis. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, volume 26, pages 534–539, Mahwah, NJ: LEA.

Harris, C. M. and Wolpert, D. M. (1988). Signal-dependent noise determines motor planning. *Nature*, 394:780–784.

Hartley, R. V. L. (1928). Transmission of information. *Bell System Technical Journal*.

Havlin, S. (1995). The distance between zipf plots. *Physica A: Statistical Mechanics and its Applications*, 216(1):148–150.

Hockett, C. F. (1960). *A course in modern linguistics*. Macmillan.

Hopper, P. J. and Thompson, S. A. (1980). Transitivity in grammar and discourse. *Language*, 56(2):281–299.

Hopper, P. J. and Thompson, S. A. (1984). The discourse basis for lexical categories in universal grammar. *Language*, 60(4):703–752.

Hualde, J. I. (2004). Quasi-phonemic contrasts in spanish. In Chand, V., Kelleher, A., Rodríguez, A. J., and Schmeiser, B., editors, *WCCFL 23: Proceedings of the 23rd West Coast Conference on Formal Linguistics*, pages 374–398, Somerville, MA. Cascadilla Press.

Hurford, J. (1989). Biological evolution of the saussurean sign as a component of the language acquisition device. *Lingua*, 77(2):187–222.

Iverson, P. and Kuhl, P. K. (2000). Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism? *Perception & Psychophysics*, 62(4):874–886.

Jakobson, R. (1942). *On Language*, chapter The Concept of Phoneme, pages 218–241. Harvard University Press. reprinted.

Jakobson, R. and Waugh, L. R. (2002). *The Sound Shape of Language*. Mouton de Gruyter.

Jeffreys, H. (1939). *Theory of probability*. International series of monographs on physics. The Clarendon press.

Johnson, W. (1932). Probability: deductive and inductive problems. *Mind*, 41:421–423.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20:137–194.

Jurafsky, D. (2003). *Probabilistic Linguistics*, chapter Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production. MIT Press.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. D. (2001). *In Frequency and the emergence of linguistic structure*, chapter Probabilistic relations between words: Evidence from reduction in lexical production, pages 229–254. John Benjamins.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall Series in Artificial Intelligence. Pearson Prentice Hall.

Kanter, I. and Kessler, D. A. (1995). Markov processes: Linguistics and zipf's law. *Phys. Rev. Lett.*, 74(22):4559–4562.

Kay, P. and Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences*, 100:9085–9089.

Kello, C. T., Brown, G. D., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., and Orden, G. C. V. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, 14(5):223–232.

Khmaladze, E. V. (1987). The statistical analysis of large number of rare events. Technical report, Dept. Math. Statist. CWI, Amsterdam.

Klatt, D. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, 7:279–312.

Klatt, D. H. (1977). Review of the arpa speech understanding project. *The Journal of the Acoustical Society of America*, 62(6):1345–1366.

Köhler, R. (1986). *Zur Linguistischen Synergetik: Struktur und Dynamik der Lexik*. Brockmeyer.

Köhler, R., Altmann, G., and Piotrovskiĭ, R. (2005). *Quantitative Linguistik /Quantitative Linguistics: Ein Internationales Handbuch /An International Handbook*. De Gruyter.

Kolguškin, A. N. (1960). *Linguistic and Engineering Studies in Automatic Language Translation of Scientific Russian Into English: Technical Report Phase II*. University of Washington Press.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*. Sage.

Laboissiere, R. (2010). preprint.

Labov, W., Ash, S., and Boberg, C. (2006). *The Atlas of North American English: Phonetics, Phonology, and Sound Change : a Multimedia Reference Tool*. Number v.1 in Topics in English Linguistic Series. Mouton De Gruyter.

Lacquaniti, F., Terzuolo, C., and Viviani, P. (1983). The law relating the kinematic and figural aspects of drawing movements. *Acta Psychologica*, 54:115–130.

Ladd, D. R. (2002). Distinctive phones in surface representation. In *8th Conference on Laboratory Phonology*, New Haven, Connecticut.

Ladefoged, P. and Maddieson, I. (1996). *The Sounds of the World's Languages*. Wiley-Blackwell.

Lamel, L. and Adda, G. (1996). On designing pronunciation lexicons for large vocabulary continuous speech recognition. In *ICSLP 96. Proceedings*, volume 1, pages 6–9.

Laplace, P. S. (1902). *A philosophical essay on probabilities*. John Wiley & Sons.

Leijenhorst, D. C. v., Weide, T. P. v. d., and Grootjen, F. (2005). A formal derivation of heaps' law. *Information Sciences*, 170(2-4):263–272.

Li, W. (1992). Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38(6):1842–1845.

Liberman, A. M. (1970). The grammars of speech and language. *Cognitive Psychology*, 1(4):301–323.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74:431–461.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1972). *Human Communications: A Unified View*, chapter Perception of the Speech Code. McGraw-Hill.

Liberman, A. M., Harris, K., Hoffman, H., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54:358–368.

Lidstone, G. J. (1920). Note on the general case of the bayes-laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8:182–192.

Lü, L., Zhang, Z.-K., and Zhou, T. (2010). Zipf's law leads to heaps' law: Analyzing their relation in finite-size systems. *PLoS ONE*, 5(12):e14139.

MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language*, 32:692–715.

Maddieson, I. (1984). *Patterns of Sounds*. Cambridge University Press.

Mandelbrot, B. (1954). Structure formelle des textes et communications. *Word*, 10:1–27.

Mandelbrot, B. (1963). The variation of certain speculative prices. *The Journal of Business*, 36(4):394–419.

Mandelbrot, B. B. (1953). An informational theory of the statistical structure of languages. In Jackson, B. W., editor, *Communication Theory*, pages 486–502.

Mandelbrot, B. B. (1965). Information theory and psycholinguistics. In Wolman, B. B. and Nagel, E. N., editors, *Scientific Psychology: Principles and Approaches*, pages 550–562. Basic Books Publishing.

Mandelbrot, B. B. (1982). *The Fractal geometry of Nature*. Freeman, New York.

Mandelbrot, B. B. (1999). *Multifractals and 1/f noise: wild self-affinity in physics (1963-1976)*. Selected works of Benoit B. Mandelbrot. Springer.

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

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.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Mit Press.

Mantegna, R. N., Buldyrev, S., Goldberger, A., Havlin, S., Peng, C., Simons, M., and Stanley, H. (1994). Linguistic features of noncoding dna sequences. *Physical review letters*, 73(23):3169–3172.

Martínez-Celdrán, E. (2004). Problems in the classification of approximants. *Journal of the International Phonetic Association*, 34:201–210.

Maslov, S., Paczuski, M., and Bak, P. (1994). Avalanches and 1/f noise in evolution and growth models. *Physical Review Letters*, 73:2162–2165.

McCowan, B., Hanser, S. F., and Doyle, L. R. (1999). Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Animal Behaviour*, 57:409–419.

Mendenhall, T. C. (1887). The characteristic curves of composition. *Science*, 9(214):237–249.

Mendenhall, T. C. (1901). A mechanical solution of a literary problem. *Popular Science*, 9:97–105.

Menzerath, P. (1928). über einige phonetische probleme. In *Actes du premier Congres International de Linguistes*, pages 104–105, Sijthoff: Leiden.

Menzerath, P. (1954). *Die Architektonik des deutschen Wortschatzes*. F. Dümmler.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M., and Lieberman-Aiden, E. (2011a). Quantitative analysis of culture using millions of digitized books. *Science*, 331:176–182.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., and Aiden, E. L. (2011b). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Mielke, J. (2005). Modeling distinctive feature emergence. In *West Coast Conference on Formal Linguistics XXIV*.

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

Miller, G. A. (1957). Some effects of intermittent silence. *The American Journal of Psychology*, 70(2):311–314.

Miller, G. A. (1965). *The Psycho-biology of language: an introduction to dynamic philology*, chapter Introduction. The MIT Press.

Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, 27(2):338–352.

Miller, G. A. and Taylor, W. G. (1948). The perception of repeated bursts of noise. *The Journal of the Acoustical Society of America*.

Montemurro, M. A. and Zanette, D. H. (2011). Universal entropy of word ordering across linguistic families. *PLoS ONE*, 6(5):e19875.

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

Nádas, A. (1985). On turing's formula for word probabilities. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(6):1414–1416.

Nairne, J. S. (1988). A framework for interpreting recency effects in immediate serial recall. *Memory & Cognition*, 16:343–352.

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, 18:251–269.

Neath, I. and Nairne, J. S. (1995). Word-length effects in immediate memory: Overwriting trace decay theory. *Psychonomic Bulletin & Review*, 2(4):429–441.

Nowak, M. A., Komarova, N. L., and Niyogi, P. (2002). Computational and evolutionary aspects of language. *Nature*, 417:611–617.

Nowak, M. A., Plotkin, J. B., and Jansen, V. A. A. (2000). The evolution of syntactic communication. *Nature*, 404:495–498.

Odden, D. (2005). *Introducing Phonology*. Cambridge University Press.

Olson, D. (1994). *The World on Paper*. Cambridge University Press, Cambridge.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2):175–184.

Pierpont, W. G. (2002). *The Art and Skill of Radio-Telegraphy*. NOHFF.

Pierrehumbert, J. (1994). *Papers in laboratory phonology, vol. 3: Phonological structure and phonetic*, chapter Syllable structure and word structure: A study of triconsonantal clusters in English, pages 168–190. Cambridge University Press, Cambridge.

Pierrehumbert, J. (2001). *Frequency and the emergence of linguistic structure*, chapter Exemplar dynamics: Word frequency, lenition, and contrast, pages 137–157. John Benjamins, Amsterdam.

Pierrehumbert, J. B. (2003). *Probabilistic Linguistics*, chapter Probabilistic Phonology: Discrimination and Robustness. MIT Press.

Pinker, S. (2003). *The Language Instinct: The New Science of Language and Mind*. Penguin Books Limited.

Piotrovskii, R. G., Pashkovskii, V. E., and Piotrovskii, V. R. (1994). Psychiatric linguistics and automatic text processing. *Nauchno-Tekhnicheskaya Informatsiya, Seriya 2*, 28(11):21–25.

Pisoni, D. B. (1982). *Project SCAMP 1981: Acoustic Phonetics and Speech Modeling*, chapter In defense of segmental representations in speech processing. Institute for Defense Analyses, Communications Research Division, Princeton, NJ.

Pollack, I. (1952). The informationof elementaryauditorydisplays. *The Journal of the Acoustical Society of America*, 24(6):745–749.

Pombo, E. L. (2002). Visual construction of writing in the medieval book. *Diogenes*, 49(196):31–40.

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.

Port, R. (2007). How are words stored in memory? beyond phones and phonemes. *New Ideas in Psychology*, 25:143–170.

Port, R. and Leary, A. (2005). Against formal phonology. *Language*.

Project Gutenberg (2013). Project Gutenberg. http://www.gutenberg.org/.

Pulvermüller, F. (2003). *The Neuroscience of Language: On Brain Circuits of Words and Serial Order*. Cambridge University Press.

Reetz, H. (2010). Simple upsid interface. [online] `http://web.phonetik.uni-frankfurt.de/upsid.html`.

Saenger, P. (1991). *Literacy and Orality*, chapter The separation of words and the physiology of reading. Cambridge University Press.

Saenger, P. H. (1997). *Space between words: the origins of silent reading*. Stanford University Press.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science*, 274:1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., and Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70:27–52.

Saffran, J. R., Newport, E. L., and Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of memory and language*, 35:606–621.

Saffran, J. R. and Wilson, D. P. (2003). From syllables to syntax: Multi-level statistical learning by 12-month-old infants. *Infancy*, 4(2):273–284.

Samuelsson, C. (1996). Relating turing's formula and zipf's law. *CoRR*, cmp-lg/9606013.

Sapir, E. (1929). The status of linguistics as a science. *Language*, 5(4):207–214.

Schneier, B. (1996). *Applied cryptography: protocols, algorithms, and source code in C*. Wiley.

Sedelow, S. Y. and Sedelow, W. A. (1966). *The computer and literary style*, chapter A preface to computational stylistics, pages 1–13. Kent State University Press.

Sendlmeier, W. F. (1995). Featere, phoneme, syllable or word: How is speech mentally represented. *Phonetica*, 52:131–143.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*.

Shannon, C. E. (1951). Prediction and entropy of printed english. Technical Report 30, The Bell System Technical Journal.

Shoup, J. E. (1980). Phonological aspects of speech recognition. In Lea, W. A., editor, *Trends in Speech Recognition*, pages 125–138. Prentice Hall, Englewood Cliffs.

Silva, D. C., de Lima, A. A., Maia, R., Braga, D., ao F. de Moraes, J., ao A. de Moraes, J., and Jr., F. G. V. R. (2006). A rule-based grapheme-phone converter and stress determination for brazilian portuguese natural language processing. In *International Telecommunications Symposium*, pages 550–554.

Siravenha, A. C., Neto, N., Macedo, V., and Klautau, A. (2008). Uso de regras fonológicas com determinação de vogal tônica para conversão grafema-fone em português brasileiro. In *7th International Information and Telecommunication Technologies Symposium*.

Steels, L. (1997). Synthesising the origins of language and meaning using co-evolution, self-organisation and level formation.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, 64:153–181.

Studdert-Kennedy, M. (1976). *Contemporary issues in experimental phonetics*, chapter Speech perception, pages 243–293. Academic Press, New York.

Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie*. Vandenhoeck und Ruprecht.

Tuldava, J. (1996). The frequency spectrum of text and vocabulary. *Journal of Quantitative Linguistics*, pages 38–50.

Tuldava, J. (2004). The development of statistical stylistics (a survey). *Journal of Quantitative Linguistics*, 11(1–2):141–151.

Turcotte, D. (1989). Fractals in geology and geophysics. In Scholz, C. and Mandelbrot, B., editors, *Fractals in Geophysics*, pages 171–196. Birkhauser Verlag, Basel.

von Humboldt, W. (1836). *The Heterogeneity of Language and its Influence on the Intellectual Development of Mankind (Über die Verschiedenheit des menschlichen Sprachbaus und ihren Einfluss auf die geistige Entwicklung des Menschengeschlechts)*. Cambridge University Press.

Vygotsky, L. S. (1934). *Thought and Language*. MIT Press, Cambridge, MA.

Wang, W. S.-Y. (1991). *Explorations in Language*. Pyramid Press.

Wang, W. S.-Y., Ke, J., and Minett, J. W. (2004). Computational studies of language evolution. In Huang, C. and Lenders, W., editors, *Computational linguistics and Beyond*. Academia Sinica: Institute of Linguistics.

Weide, R. L. (2008). Carnegie mellon pronouncing dictionary, release 0.7a. `http://www.speech.cs.cmu.edu/`.

Whaley, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior*, 17:143–154.

Whorf, B. (1940/1956). *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*, chapter Science and linguistics, pages 207–219. MIT Press, Cambridge, MA.

Wickelgren, W. A. (1969). *Information Processing in the Nervous System*, chapter Context-sensitive coding in speech recognition, articulation, and development, pages 85–95. Springer-Verlag, New York.

Wilden, A. (2001). *System and Structure: Essays in Communication and Exchange*. International Behavioural and Social Sciences Classics from the Tavistock Press, 96. Routledge.

Yoshida, K., Watanabe, T., and Koga, S. (1989). Large vocabulary word recognition based on demi-syllable hidden markov model using small amount of training data. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 1–4.

Young, G. and Householder, A. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3:19–22.

Zipf, G. K. (1935). *The Psycho-biology of language: an introduction to dynamic philology*. The MIT Press.

Zipf, G. K. (1942). Children's speech. *Science*, 96:344–345.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*. Hafner Pub. Co.

Ziv, J. and Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536.