On the ranking property and underlying dynamics of complex systems

Weibing Deng

Université du Maine

Ecole Doctorale 3MPL

Thèse de Doctorat

Spécialité: Physique

Rélisée au Laboratoire de Physique Statistique et Systèmes Complexes de l'ISMANS

par

Weibing DENG

pour l'obtention du grade de Docteur en Sciences

On the Ranking Property and Underlying Dynamics of Complex Systems

Présentée le 21 Juin 2013

Acceptée sur proposition du jury:

Prof. Q. Alexandre Wang (Directeur de thèse)

Prof. Alain Bulou (Co-Directeur)

Prof. Yicheng Zhang (Rapporteur)

Prof. Eduard V. Vakarin (Rapporteur)

Prof. Wei Li

Prof. Paul Bourgine

Prof. Rene Doursat

Prof. François Tsobnang

Acknowledgements

Two and a half years slid away so quickly, during this important period of my PhD studies in ISMANS, I was greatly supported by my supervisors, colleagues, friends and family members. At this very moment, I would like to express my great thanks to all of them.

Foremost, I deeply appreciate my supervisor, Prof. Q. Alexandre Wang, for giving me the precious opportunity to study in ISMANS. He introduced me to a more international research environment, and gave me the chance to get in touch with the senior experts of our research area. His valuable ideas, suggestions and guidance are very helpful and beneficial to me. I am so inspired by his enthusiasm and tremendous interests in doing research. I am also very much indebted to his great help for my daily life here.

I am profoundly grateful to my supervisor, Prof. Xu Cai, in Central China Normal University. He provided me with insights into the research subjects, and encouraged me to work independently. He taught me, how to cultivate the positive values of life, how to communicate with people, how to handle affairs, etc. His guidance, encouragement and endless support always prompt me to move forward, and to struggle vigorously to be better.

I wish to express the warm and sincere thanks to my another supervisor, Prof. Wei Li, in Central China Normal University. He led me to the interesting research subject of complexity science 6 years ago. He taught me how to do the empiric and analytic research, how to write research papers, how to respond to the referees' comments, etc. He is more like a good friend in my daily life, who always give me a hand when I am in difficulties.

I owe the particular thanks to Prof. Armen E. Allahverdyan for his guidance and great help. He taught me how to do the theoretical calculations, he trained me how to give the report, he encouraged me to be strong in face of difficulties, etc. I appreciate all his contributions of time, ideas and helpful discussions.

I am thankful to my associated supervisor, Prof. Alain Bulou, the CST scientific members, Prof. Pascal Picart, Prof. Wenping Bi, for their useful reports and comments. Thanks should be also given to Prof. Alain Le Méhauté, Prof. Abe Sumiyoshi, Prof. Alberto Robledo and Prof. Christian Beck for their helpful discussions.

I would like to thank the director of ISMANS, Prof. Mouad Lamrani for his support and encouragement. I appreciate the valuable discussions with Dr. Aziz El Kaabouchi, Dr. Laurent Nivanen, Dr. Benoît Minisini and Dr. Cyril Pujos. I would like to thank Prof. François Tsobnang, Prof. Jean-Charles Craveur, Dr. Dominique Marceau, Dr. Gilles Brement, Mr. Perry Ngabo, Mrs. Sylvie Bacle, Mrs. Sandra Dugué, Miss Zélia Aveline, Miss Susanne Vilchez and all the other staffs in ISMANS for their personal kind help during my stay here.

I would like to take this opportunity to thank my previous teachers in Institute of Particle Physics of Central China Normal University, Prof. Enke Wang, Prof. Yadong Yang, Prof. Daicui Zhou, Prof. Feng Liu, Prof. Chunbin Yang, Prof. Benwei Zhang, Prof. Zhongbao Yin, Prof. Daimei Zhou, Prof. Liping Chi, Prof. Yaping Wang, Prof. Mingmei Xu, and all the other professors and staffs, they helped me a lot during my undergraduate and postgraduate.

Many thanks are given to my colleagues and friends here, Ru Wang, Jian Jiang, Kai Zhao, Tongling Lin, Zi Hui, Xiaozheng Zhang and Wenxuan Zhang, they made my life funny and enjoyable. Grateful thanks are also given to all the previous and current members in Complexity Science Center of Central China Normal University for their support, help and fruitful discussions.

I acknowledge the financial support of China Scholarship Council, Natural Science Foundation of China, Ministry of Education of China.

Lastly, I would like to express the deepest gratitude to my parents, sisters, brothers, nephews and nieces, thank you for giving me a warm and beloved family. Your constant support and encouragement strengthened my courage to go through these years. Special thanks are given to my fiancee, Ms. Dujuan Wang, for her selfless supporting, encouraging, and understanding. I am also deeply indebted to the continuous care of her family members.

Le Mans, France

March, 2013

Abstract

Ranking procedures are widely used to describe the phenomena in many different fields of social and natural sciences, e.g., sociology, economics, linguistics, demography, physics, biology, etc. In this dissertation, we dedicated to study the ranking properties and underlying dynamics embedded in complex systems. In particular, we focused on the scores/prizes ranking in sports systems and the words/characters usage ranking in human languages. The aim is to understand the mechanisms behind these issues by using the methods of statistical physics, Bayesian statistics and agent-based modeling. The concrete results concern the following aspects.

We took up an interesting topic on the scores/prizes ranking in sports systems, and analyzed 40 data samples in 12 different sports fields. We found the striking similarities in different sports, i.e., the distributions of scores/prizes follow the universal power laws. We also showed that the data yielded the Pareto principle extensively observed in many social systems: 20% of the players accumulate 80% of the scores and money. For the tennis head-to-head data, we revealed that when two players compete, the probability that the higher-ranked player will win is related to the rank difference of the two opponents. In order to understand the origins of the universal scaling, we proposed an agent-based model, which can simulate the competitions of players in different matches, and results from our simulations are consistent with the empirical findings. Extensive simulation studies indicate that the model is quite robust with respect to the modifications of some parameters.

Zipf's law is the major regularity of statistical linguistics that served as a prototype for the rank-frequency relations and scaling laws in natural sciences. We investigated several English texts, clarified the valid range of Zipf's law, and found this valid range increases upon mixing different texts. Based on the latent semantic analysis, we proposed a probabilistic model, in which we assumed that the words are drawn into the text with random probabilities, while their apriori density relates, via Bayesian statistics, to the general features of mental lexicon of the author who produced the text. Our model explained the Zipf's law together with the limits of its validity, its generalization to high and low frequencies and hapax legomena.

In another work, we specified the rank-frequency relations for Chinese characters.

We chose to study the short texts first, since for the sake of the rank-frequency analysis, long texts are just mixtures of shorter, thematically homogenous pieces. Our results showed that the Zipf's law for Chinese characters perfectly holds for sufficiently short texts (few thousand different characters), and the scenario of its validity is similar to that for short English texts. We argued long Chinese texts display a two-layer, hierarchic structure: power-law rank-frequency characters (first layer) and the exponential ones (second layer). The previous results on the invalidity of the Zipf's law for long texts are accounted for by showing that in between of the Zipfian range and the region of very rare characters (hapax legomena) there emerges a range of ranks, where the rank-frequency relation is approximately exponential. From comparative analysis of rank-frequency relations for Chinese and English, we suggested the characters play for Chinese writers the same role as the words for those writing within alphabetical systems.

Keywords: Ranking systems, Power laws, Pareto principle, Zipf's law
Sports ranking, Human languages, Prior probability

Résumé

Des procédures de classement sont largement utilisées pour décrire les phénomènes observés dans de nombreux domaines des sciences sociales et naturelles, par exemple la sociologie, l'économie, la linguistique, la démographie, la physique, la biologie, etc. Dans cette thèse, nous nous sommes attachés à l'étude des propriétés de classement et des dynamiques sous-jacentes intégrées dans les systèmes complexes. En particulier, nous nous sommes concentrés sur les classements par score ou par prix dans les systèmes sportifs et les classements d'utilisation des mots ou caractères dans les langues humaines. Le but est de comprendre les mécanismes sous-jacents à ces questions en utilisant les méthodes de la physique statistique, de la statistique bayésienne et de la modélisation multi-agents. Les résultats concrets concernent les aspects suivants.

Nous avons tout d'abord traité une étude sur les classements par score/prix dans les systèmes sportifs et analysé 40 échantillons de données dans 12 disciplines sportives différentes. Nous avons trouvé des similitudes frappantes dans différents sports, à savoir le fait que la répartition des résultats/prix suit les lois puissance universelles. Nous avons également montré que le principe de Pareto est largement respecté dans de nombreux systèmes sociaux: ainsi 20% des joueurs accumulent 80% des scores et de l'argent. Les données concernant les matchs de tennis en individuels nous ont révélé que lorsque deux joueurs s'affrontent, la probabilité que le joueur de rang supérieur gagne est liée à la différence de rang des deux adversaires. Afin de comprendre les origines de la mise à l'échelle universelle, nous avons proposé un modèle multi-agents, qui peut simuler les matchs de joueurs à travers différentes compétitions. Les résultats de nos simulations sont cohérents avec les résultats empiriques. L'extension du domaine d'étude de la simulation indique que le modèle est assez robuste par rapport aux modifications de certains paramètres.

La loi de Zipf est le comportement le plus régulièrement observé dans la linguistique statistique. Elle a dès lors servi de prototype pour les relations entre rang d'apparitions et fréquence d'apparitions (relations rang-fréquence dans la suite du texte) et les lois d'échelle dans les sciences naturelles. Nous avons étudié plusieurs textes, précisé le domaine de validité de la loi de Zipf, et trouvé que la plage de validité augmente lors du mélange de différents textes. Basé sur l'analyse sémantique latente, nous avons

proposé un modèle probabiliste, dans lequel nous avons supposé que les mots sont ajoutés au texte avec des probabilités aléatoires, tandis que leur densité a priori est liée, via la statistique bayésienne, aux caractéristiques générales du lexique mental de l'auteur de ce même texte. Notre modèle explique la loi de Zipf ainsi que ses limites de validité, et la généralise aux hautes et basses fréquences et au hapax legomena.

Dans une autre étude, nous avons précisé les relations rang-fréquence pour les caractères chinois. Nous avons choisi d'étudier des textes courts en premier, car pour le bien de l'analyse rang fréquence, les longs textes ne sont que des mélanges de textes plus courts, thématiquement homogènes. Nos résultats ont montré que la loi de Zipf appliqués aux caractères chinois tient parfaitement pour des textes assez courts (quelques milliers de caractères différents). Le même domaine de validité est observé pour les textes courts anglais. Nous avons soutenu que les longs textes chinois montrent une structure hiérarchique à deux couches: des caractères dont la fréquence d'apparition suit une loi puissance (première couche) et des caractères dont l'apparition suit une loi exponentielle (deuxième couche). Les résultats antérieurs sur la nullité de la loi de Zipf pour les textes longs sont comptabilisés en montrant qu'entre l'intervalle de la gamme de Zipf et la région de caractères très rares (hapax legomena), il se dégage une gamme de rangs, pour laquelle la relation rang-fréquence est approximativement exponentielle. À partir de l'analyse comparative des relations rang-fréquence pour le chinois et l'anglais, nous suggérons que les caractères jouent pour les écrivains chinois le mêmer? Le que les mots pour ceux qui écrivent dans un système alphabétique.

Mots-clefs: Systèmes de classement, Lois puissance, Principe de Pareto,
Loi de Zipf, Classement sportif, Langues humaines, Probabilité a priori

Contents

Chapter 1

Introduction

1.1 Social physics and stylized facts

We never stop the pace of uncovering the structures, dynamics, evolutions, and functions of our human society [1–9]. Although the complexity of its nature, e.g., the interactions are between sophisticated human beings with cognitive capabilities; each individual interacts only with a limited number of peers, while this number is normally negligible compared to the total number of individuals in the system, etc, on the macro-level, human societies are characterized by the amazing global regularities [10]. For instance, there are collective "emergent" behaviors [11–13], like food riots, revolutions, ethnic violence, urban health, panics, etc. There are self-organization phenomena [14, 15], like critical mass, herd behavior, groupthink, etc. There are transitions from disorder to order [16], like the spontaneous formation of a common language/culture or the emergence of consensus on a specific issue. There are examples of scaling and universality [17, 18]. All these macroscopic phenomena spontaneously call for a natural science approach to study the social behaviors [19].

The father of sociology, Auguste Comte[1] put forward the idea of "social physics" nearly 200 years ago, he hoped that the puzzles of social systems could be revealed by the natural science (physics, mathematics, computer science, etc) approaches, that is, to use the concepts, principles and methods of natural science to explore, simulate, and understand the social behavior rules [10].

[1]Auguste Comte (1798 – 1857), the French philosopher, he is traditionally considered as the "father of sociology" – first used the term "sociology" in 1838 to refer to the scientific study of society.

During the last century, great progress has been achieved in this field, there are many stylized facts which have been found with a surprisingly large range of validity:

1. The Gravity Law [20–22], which is employed to describe the distribution of trade flows and migration.

2. The Pareto Principle (80/20 rule) [23], according to which roughly 80% of an effect comes from about 20% of the causes.

3. The Fisher Equation [24] for financial mathematics, which determines the relationship between nominal and real interest rates under inflation.

4. The Zipf's Law [25], which has been widely found in the rank-frequency relations for city population, human wealth, word usage, webpage visit, scientific citation, and other physical phenomena [26–29].

5. The Fat-tailed Distributions [30–32], which exhibit extremely large skewness or kurtosis, have been observed in economics, physics, earth science, etc.

6. The Matthew Effect [33], i.e., the rich-gets-richer effect [34], or the accumulation of capital in economics [35], the preferential attachment in networks [36], etc.

7. The Goodhart's Law [37, 38], according to which any observed statistical regularity breaks down once pressure is placed upon it for control purposes.

8. The Dunbar's Number [39], which is a suggested cognitive limit to the number of people with whom one can maintain stable social relationships.

9. The Scaling Law or Power Laws [40, 41], that is, when measuring the probability of a particular value of some quantity, if it varies inversely as a power of that value, then the quantity is said to follow a power law. It appears widely in physics, biology, computer science, earth and planetary sciences, economics and finance, demography and the social sciences, etc.

.....

1.2 General framework of methods

Being a rather interdisciplinary field, there is a natural tendency to appreciate different perspectives and methods, to welcome innovations and new ideas. Here, some commonly used methods are briefly reviewed as follows.

A. Statistical analysis and Data mining

Statistical analysis [42] concerns the study of collection, organization, analysis and interpretation of the data. The descriptive quantities of the data include the mean, standard deviation, skewness, kurtosis, etc. While the generally employed methods are time series analysis [43], regression analysis [44], statistical hypothesis testing [45], etc.

Data mining is more related to the purpose of inference statistics [46], it is the computational process of discovering patterns in large data sets, which involves different methods from the artificial intelligence, machine learning, statistics, and database systems. The overall goal is to extract information from a data set and transform it into an understandable structure for further use.

B. Network perspective

A network is a representation of a set of nodes or vertices, where some nodes are connected by links or edges. [47, 48]. An extensively wide range of systems in nature and society take the form of networks, for examples, the cell could be considered as a network of chemicals linked by chemical reactions, and the internet is a network of routers and computers connected by physical links, etc.

The last decade has witnessed the tremendous progress in the research of networks [49–51], which was largely inspired by the empirical study of real-world networks, e.g., the social, biological, and technological networks. Examples [2, 47, 48, 50–53] include the internet, the world wide web, social friendship networks, networks of business relations between companies, movie actor collaboration network, neural networks, metabolic networks, ecological networks, scientific citation networks, networks in linguistics, telephone call network, transportation networks, and many others.

The subjects studied include topology, dynamics, formation and function of networks. For instance, the general structural properties [2, 47, 48, 54–56] considered are, degree distributions, clustering, shortest path length, small-world effect, assortativity or disassortativity among nodes, community structure, hierarchical structure, etc; the dynamical processes taking place on networks [2, 47–51], such as information or epidemic spreading, emergence theory of evolving networks, network's robustness against failures and attacks, etc; the network models [40, 41, 47, 48, 51, 57, 58], for example, random graph models, models of network growth and preferential attachment, constructions of small-world network, temporal networks, geographical networks, etc.

C. Probabilistic model

Probabilistic model [59, 60] is widely used in the uncertainty analysis of social systems. It works by showing that if one randomly chooses objects from a specified class, the probability that the result belongs to the prescribed kind is more than zero. In probabilistic approach, uncertainties are characterized by the probabilities associated with events. While the probability of an event can be interpreted in terms of the frequency of occurrence of that event, when a large number of samples or experiments are considered, the probability of an event is defined as the ratio of the number of times the event occurs to the total number of samples or experiments (the law of large numbers).

In social systems, many problems are complicated that they cannot be solved accurately by using the simple and deterministic rules. However, if we introduce the stochastic mechanisms into the solution, it is possible to find the good approximate answers to these problems [61, 62]. Moreover, many natural and social phenomena are characterized by a variety of randomness, and probabilistic model is of fundamental importance to the show the randomness of the phenomena [63]. For instance, probabilistic model for languages [64], probabilistic model for speech recognition [65], or probabilistic model for machine perception [66], etc.

D. Agent-based model

Agent-based model (ABM) [67,68] is a powerful simulation modeling technique that has seen a number of applications in life sciences, ecological sciences and social sciences in the last few years [69]. It is a class of computational models for simulating the actions and interactions of autonomous agents (both individual or collective entities such as organizations or groups). Agents assess their situations and make decisions on the basis of a set of rules independently, they may execute various behaviors appropriate for the system they represent, such as producing, consuming, or selling. It combines the elements of game theory, complex systems, emergence, computational sociology, multi-agent systems, evolutionary programming and monte carlo methods [70].

ABM has many advantages over other modeling techniques [68], For instance, it makes the model closer to reality, it provides a natural description and simulation of the system composed of "behavioral" entities. It is suited not only to reflect interactions between different individuals, it also allows one to determine the implications of

different hypotheses [70].

Agent-based simulations acquire a very important role in the modeling of complex systems, and are proving successful in a number of areas [71–73], ranging from structure formation in biological systems, pedestrian traffic to the simulation of urban aggregation, opinion formation processes, competition-driven systems, etc.

E. Dynamical systems approach

Physicists introduce the methods and tools from theory of dynamical systems [74], non-linear dynamics [75], or chaos [76] to study the social systems, namely social dynamics, it refers to a systematic approach for mathematical modeling of social systems. It studies not only the behavior of groups that results from individuals' interactions, but also the relationship between individual interactions and group level behaviors [77]. It concerns with changes over time and emphasizes the role of feedbacks [78].

Research in social dynamics typically takes a behavioral approach [79,80], assuming that individuals are rational and act on local information. On the one hand, mathematical and computational modeling are important tools, since it focuses on individual level behavior, and recognizes the importance of heterogeneity across individuals [81]. On the other, the approximation techniques, such as mean field approximations from statistical physics, or averaging methods from computer simulation, are often used to understand the behaviors of the system that changes over time [82, 83].

F. Critical phenomena

Critical phenomena [10, 84, 85] is the collective name associated with the physics of critical points, it includes scaling relations among different quantities, self-organized criticality effects, universality, fractal behavior, finite size effects, etc. The compact combination of social systems, with the feature of small diameters, and their complex architectures result in a variety of critical effects [10].

One common theme is the understanding of transition from an initial disordered state to an ordered one (emergence of consensus in opinion dynamics, collective patterns of behavior in social systems, etc) [10, 86]. In order to explain the origins of these phenomena, we shall employ the Ising model as a pedagogical example [87]. It is an extremely simplified mathematical model for describing the spontaneous emergence of order. Despite its simplicity, it is valuable for verification of general theories and

assumptions [88]. Moreover, we should consider the finite size effects [10], the very concept of order-disorder phase-transitions is rigorously defined only in the limit of a system with an infinite number of particles (thermodynamic limit), because only in that limit truly singular behavior can arise.

While critical phenomena in networks include a wide range of issues [10, 89–91]: structural changes in networks, the emergence of critical scale-free network architectures, various percolation phenomena, epidemic thresholds, phase transitions in cooperative models defined on networks, critical points of diverse optimization problems, transitions between different regimes in processes taking place on networks, equilibrium and growing networks including the birth of the giant connected component, critical phenomena in spin models placed on networks, synchronization, and self-organized criticality effects in interacting systems on networks, etc.

1.3 Research motivation and thesis overview

Ranking is an effective technique skill to structure our perceptions of the real-world. It is a kind of evaluation and organization of information according to certain criteria, which shows the relationship between a set of items such that, for any two items, one is either 'ranked higher than', 'ranked lower than' or 'ranked equal to' the other. Ranking procedures have been widely used in almost every corner of our society:

- Politics: Rankings of the governance performance, human power/influence, national comprehensive strength, democracy index, etc.

- Economics: Rankings of the world's richest people, world's largest corporations, world's most valuable brands, countries's GDP or CPI, etc.

- Culture: Rankings of the oldest languages, world's most cultured cities, words usage in human language texts, etc.

- Science: Rankings of the countries' academy, impact factors of scientific journals, citations of scientific papers, etc.

- Technology: Rankings of the international patent filings, world's most efficient power plants, webpages' visits, blogs' click through rates, etc

- Education: Rankings of the world's best universities, best business schools, best high schools, etc.

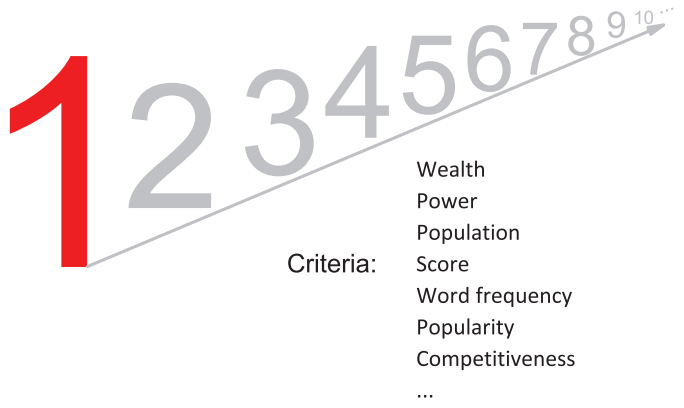

Fig. 1.1: (Color online) Schematic diagram of rankings. In different complex systems, either social, economical, or technological systems, according to certain criteria, for examples, wealth, power, population, etc, individuals or elements in the systems are ranked from 1 to N, N is the system size.

- Sports: Rankings of the players' or teams' scores or prize money, countries' medals in Olympic games, field goal shooting records of NBA players, etc.

.....

By studying all these ranking architectures, we shall be able to detect the empirical features and uncover the underlying dynamics in real-world complex systems.

Ranking means competition, all the agents in the system compete with each other and struggle to occupy the higher ranks. When some ones enter into higher ranks, then the former ones will fall off to lower ranks, and this process repeats constantly during the evolution of the system. Therefore, the ranking structure should be able to reflect an efficient organization of the system.

In this thesis, I am motivated to study two specific kinds of rankings, namely, the scores/prizes ranking in sports systems and the words/characters usage ranking in human languages. I would like to explore:

In sports, whether the ranking architectures agree with those stylized facts found in other social systems, such as the Pareto principle, the winner-take-all rule, the rich-get-richer effect, the scaling and universality, etc. If it does, how shall we explain that for sports systems?

In human languages, Zipf's law is the most remarkable regularity for the rank-

frequency relations. Does it image the cognitive capabilities of human beings' efficient organizations of information? Is the Zipf' law informative, or it is just reducible to any trivial statistical rule.

Thesis overview:

Sports ranking and agent-based model

We found the striking similarities in different sports, i.e., the distributions of scores/ prizes follow the universal scaling law, and uncovered the data yielded the Pareto principle which was extensively observed in many social systems. We also proposed an agent-based model which can simulate the competition process in sports, and generally produce the trend conveyed by empirical data.

Zipf's Law and probabilistic model

We clarified the valid range of the Zipf's law, and proposed a probabilistic model, in which the words are drawn into the text with random probabilities, while their apriori density relates to the stable and efficient organization of the author's mental lexicon.

Chinese characters: Zipf's Law and beyond

We specified the rank-frequency relations for Chinese characters. Our results showed the Zipf's law for Chinese characters perfectly holds for sufficiently short texts (few thousand different characters). While long Chinese texts display a two-layer, hierarchic structure: power-law rank-frequency characters (first layer) and the exponential ones (second layer).

Chapter 2

Universal Scaling in Sports Ranking

2.1 Research motivation in sports ranking

As is well known, ranking is a very ubiquitous phenomenon in social, economical, or technological systems. Our motivation is whether there are some common patterns in the vastly different ranking systems. Moreover, if yes, can we understand the formalism of such patterns, and unravel some properties of such competition driven systems and human dynamics [92–95]? In order to facilitate our study we choose a specific kind of ranking system, sports ranking, in which data are easily accessible and more suitable for quantitative analysis. Here players' performances in different matches will be used as the basis of their respective rankings, in terms of scores and/or prize money.

To understand how a certain sports ranking system works [96–98], let us take tennis as an example. ATP (Association of Tennis Professionals) and WTA (Women's Tennis Association) are world's most successful tennis associations for male and female professionals, respectively. To appear on the ranking systems of ATP or WTA, the number of tournaments a player has to play each year should reach a minimum, say 10. Tournaments have been divided into several categories, such as grand slams, premier tournaments, international tournaments and year-ending tour championships, mainly based on the scale of prize money. For the most important tournaments such as grand slams, the main draw only consists of 128 players. The entry rule is that if you are higher ranked, then you have more chances to be accepted. On the other hand, players' good performance will improve their rankings which will in turn entitle them more chances to play tournaments. Since there are so many tournaments each year, for

both ATP and WTA, the ranking list of scores and/or prize money may change very frequently. Here we are not interested in which specific player is world No.1 in certain sports, but instead the statistical distribution of performance, measured by scores and prize money, of all the players. What is the form of such a distribution? Is it stable over different time periods? Is it universal?

2.2 Empirical results of sports ranking

2.2.1 Database of sports systems

Our data sets cover 12 different sports fields, they are tennis, golf, table tennis, volleyball, football, snooker, badminton, basketball, baseball, hockey, handball and fencing, in those sports fields competitions are pairwise (i.e., among two players or teams). We collected the data of the scores or prize money of players or teams on the official web pages of those sports, all the data are updated up to February 2011 [99].

2.2.2 Cumulative distributions of scores

A player's score or prize money is a direct measure of his/her performance in different matches. The higher the score, the better the performance. The statistical distribution of scores or prize money reflects the profile of the performance of all the members belonging to the same association. Every sports field has its own scoring system, hence the orders of magnitude of scores are usually different. In order to make the distributions of scores or prize money comparable for different sports fields, we rescale the quantities of interest. That is,

$$R_S = S/S_{max}, \tag{2.1}$$

where S denotes the values of quantities considered, e.g., scores or prize money, and S_{max} is the maximum value of S in the sample, which pertains to the No. 1 player in the ranking list by using S. We adopt the cumulative distribution due to small system sizes.

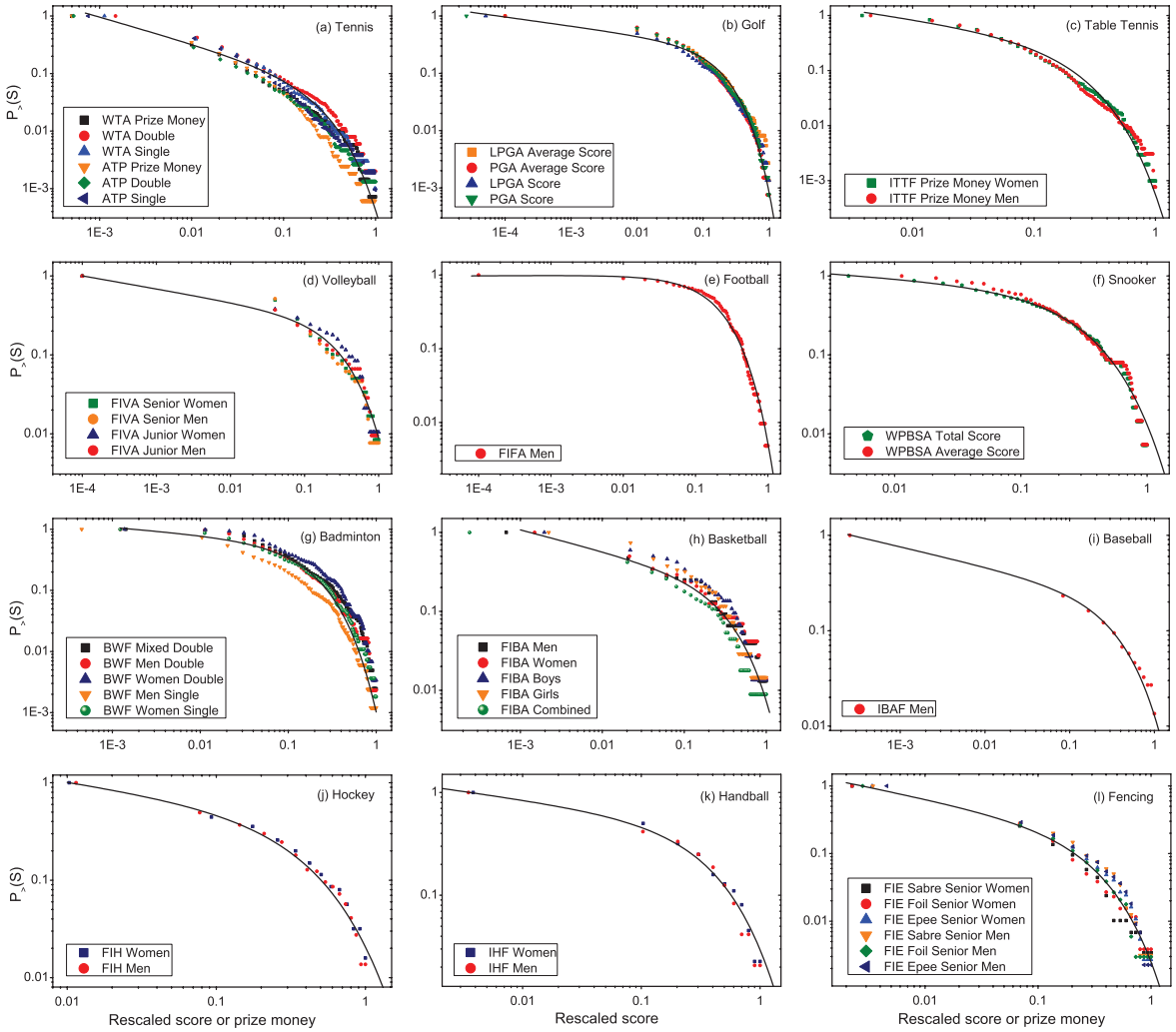

Fig. 2.1: (Color online) Cumulative distributions of scores and/or prize money for 12 different sports fields. (a) Tennis: Association of Tennis Professionals (ATP) and Women's Tennis Association (WTA). (b) Golf: Professional Golfers' Association (PGA) and Ladies Professional Golf Association (LPGA). (c) Table tennis: International Table Tennis Federation (ITTF). (d) Volleyball: International Federation of Volleyball (FIVB). (e) Football: International Federation of Football Association, commonly known as FIFA. (f) Snooker: World Professional Billiards and Snooker Association (WPBSA). (g) Badminton: Badminton World Federation (BWF). (h) Basketball: International Basketball Federation, more commonly known as FIBA. (i) Baseball: International Baseball Federation (IBAF). (j) Hockey: International Field Hockey Federation (FIH). (k) Handball: International Handball Federation (IHF). (l) Fencing: International Fencing Federation (FIE). All the black solid curves in the Figs are the power laws with exponential decay, $P_>(S) \propto S^{-\tau} \exp(-S/S_c)$, where τ is the power law exponent and S_c corresponds to the characteristic turning point of the exponential decay. The values of τ and S_c for different sports fields are provided in Table 2.1.

Table 2.1: System sizes of 40 samples in the 12 different sports ranking systems, p-values of goodness-of-fit tests [41], with hypothesized distribution being the power law with exponential decay distribution (p_1), the exponential distribution (p_2), the stretched exponential distribution (p_3), and the log-normal distribution (p_4), values of the exponents τ and S_c in the power law with exponential decay, and values of the ratio for the test of Pareto principle.

Sports ranking systems	Sizes	p_1	p_2	p_3	p_4	τ	S_c	ratio
ATP Single	1763	0.65	0.00	0.03	0.06	0.31	0.12	0.79
ATP Double	1516	0.52	0.01	0.02	0.03	0.32	0.18	0.78
ATP Prize Money	1636	0.56	0.02	0.05	0.03	0.33	0.13	0.79
WTA Single	1523	0.62	0.00	0.23	0.18	0.39	0.15	0.78
WTA Double	1028	0.75	0.00	0.18	0.20	0.38	0.19	0.80
WTA Prize Money	1388	0.81	0.00	0.21	0.16	0.39	0.12	0.81
PGA Score	1323	0.85	0.00	0.00	0.00	0.16	0.18	0.82
LPGA Score	734	0.82	0.00	0.00	0.00	0.18	0.19	0.78
PGA Average Score	1323	0.76	0.00	0.00	0.00	0.16	0.19	0.79
LPGA Average Score	734	0.82	0.00	0.00	0.00	0.17	0.20	0.82
ITTF Prize Money Men	1717	0.85	0.00	0.02	0.03	0.32	0.17	0.83
ITTF Prize Money Women	1288	0.73	0.00	0.01	0.02	0.32	0.18	0.82
FIVA Junior Men	105	0.86	0.00	0.00	0.00	0.16	0.21	0.76
FIVA Junior Women	95	0.68	0.00	0.00	0.00	0.14	0.20	0.79
FIVA Senior Men	138	0.69	0.01	0.00	0.00	0.13	0.16	0.78
FIVA Senior Women	127	0.92	0.00	0.00	0.00	0.11	0.18	0.82
FIFA Men	209	0.59	0.01	0.00	0.00	0.01	0.19	0.77
WPBSA Total Score	97	0.69	0.00	0.00	0.00	0.11	0.27	0.83
WPBSA Average Score	97	0.58	0.00	0.00	0.00	0.13	0.25	0.78
BWF Women Single	548	0.68	0.00	0.00	0.00	0.12	0.16	0.80
BWF Women Double	295	0.53	0.00	0.00	0.00	0.13	0.18	0.78
BWF Men Single	833	0.62	0.00	0.00	0.00	0.06	0.17	0.82
BWF Men Double	429	0.75	0.00	0.00	0.00	0.08	0.13	0.81
BWF Mixed Double	407	0.63	0.00	0.00	0.00	0.07	0.14	0.79
FIBA Men	79	0.86	0.00	0.00	0.00	0.19	0.20	0.81
FIBA Women	72	0.98	0.00	0.00	0.00	0.18	0.21	0.83
FIBA Boys	77	0.62	0.00	0.00	0.00	0.18	0.23	0.82
FIBA Girls	72	0.85	0.00	0.01	0.01	0.26	0.22	0.76
FIBA Combined	115	0.52	0.01	0.00	0.00	0.23	0.20	0.81
IBAF Men	78	0.96	0.00	0.00	0.00	0.20	0.28	0.79
FIH Men	73	0.86	0.00	0.00	0.00	0.23	0.26	0.78
FIH Women	68	0.83	0.00	0.00	0.00	0.21	0.27	0.81
IHF Men	52	0.68	0.00	0.00	0.00	0.16	0.25	0.79
IHF Women	46	0.69	0.00	0.00	0.00	0.15	0.27	0.76
FIE Sabre Senior Women	371	0.56	0.00	0.12	0.08	0.34	0.25	0.81
FIE Foil Senior Women	260	0.65	0.00	0.03	0.00	0.32	0.23	0.78
FIE Epee Senior Women	293	0.53	0.01	0.16	0.17	0.36	0.24	0.83
FIE Sabre Senior Men	319	0.67	0.00	0.00	0.02	0.32	0.23	0.78
FIE Foil Senior Men	337	0.56	0.00	0.00	0.00	0.30	0.21	0.82
FIE Epee Senior Men	442	0.72	0.00	0.01	0.00	0.28	0.25	0.81

2.2.3 Methods of goodness-of-fit tests

Cumulative distributions of players' scores or prize money have been shown in Fig. 2.1 for 40 data samples of 12 different sports ranking systems. Amazingly, all the distributions share very similar trend, and it should also be noticed that for the same field, all the curves nearly collapse with each other. Therefore now, the main task is to determine which statistical distribution is favored over the others, or equivalently, which statistical distribution is ruled out by the observed data, while the others are not.

There are several common statistical distributions [41], such as the power law with exponential decay distribution, $p(x) \sim x^{-\alpha}e^{-\lambda x}$, the exponential distribution, $p(x) \sim e^{-\lambda x}$, the stretched exponential distribution, $p(x) \sim x^{\beta-1}e^{-\lambda x^{\beta}}$, and the log-normal distribution, $p(x) \sim \frac{1}{x}\exp[-\frac{(\ln x-\mu)^2}{2\sigma^2}]$, etc. Here, we employ the methods of goodness-of-fit tests in Ref. [41] to quantify which hypothesis distribution is favored over the others in fitting the data. To do this, we would first determine the least square fitting to the data. Secondly, we calculate the corresponding Kolmogorov-Smirnov (KS) statistics for the goodness-of-fit test of the best-fit hypothesis distribution, then repeat the calculation of the KS statistics for a large number of synthetic data sets. Lastly, we calculate the p-value as the fraction of the KS statistics for the synthetic data sets whose value exceeds the KS statistic for the real data. If the p-value is sufficiently small (say $p < 0.1$), then the hypothesis distribution can be ruled out.

The p-values of the goodness-of-fit tests for the above hypothesis distributions are given in Tab. 2.1. As one could find, with hypothesis distribution being the power law with exponential decay, the p-values are all much larger than 0.1. Whereas for the exponential distribution, the p-values are all smaller than 0.1, so the exponential distribution is ruled out. While for the stretched exponential distribution and the log-normal distribution, the majority of p-values are smaller than 0.1, yet few of them are a little bit larger than 0.1, which implies these two alternative distributions are just good fits in the very rare cases. Therefore, we can conclude, the case of the power law with exponential decay in its favor is strengthened. With the form

$$P_{>}(S) \propto S^{-\tau}\exp(-S/S_c), \tag{2.2}$$

where τ and S_c are exponents of the power law and the exponential decay, respectively, values of them are shown in Table 2.1, with $0.01 \leq \tau \leq 0.39$ and $0.12 \leq S_c \leq 0.28$.

Therefore, by using the goodness-of-fit test and checking values of the fitting parameters, we could observe the shared feature in the sports systems. The evidence of the power-laws in the sports ranking indicates that there is still significant probability to have superman such as Roger Federer in tennis or Tiger Woods in golf. But the prevalent probability is still the players who do not play in the top form. Unlike the human height system, it seems there is no typical player who plays with average level.

2.2.4 Comparisons with the Random Group Formation

Such kind of distributions have also been widely found in a number of different systems, such as the distribution of richness, city-sizes, word-frequencies, family names, species, and degrees of metabolic networks, etc. In Refs. [100–102], it is proposed that the shared feature in these systems could be well characterized by the Random Group Formation (RGF), from which a Bayesian estimate is obtained based on the minimal information cost, given the sole a priori knowledge of the total number of elements, groups and the number of elements in the largest group. This estimate predicts a unique distribution of the system, with the form

$$P(k) = A\frac{\exp(-bk)}{k^{\gamma}}, \tag{2.3}$$

where k denotes the elements of the system, and values of A, b and γ are obtained directly from a set of self-consistent equations, while γ usually takes the values in the range of $1 \leq \gamma \leq 2$ [101]. According to the detailed explanations and calculation processes in Ref. [101], we applied the RGF predictions to the sports systems, with a priori knowledge being the total scores of the system M, the number of players N and the highest scores in the system k_{max}. Tab. 2.2 gives the values of M, N and k_{max} of 19 sports systems described above, which are needed for uniquely determining the RGF prediction for each case. By using the same calculation method in Ref. [101], we could obtain the values of A, b and γ of the RGF predictions for each sports system (Tab. 2.2).

Now, we employ the Kolmogorov-Smirnov test [103] (KS test) to compare the RGF predictions with the original probability distributions of scores in sports systems, in order to quantify whether the RGF prediction could characterize the sports data. With null hypothesis being the sports data follows the RGF prediction, we calculated the

Table 2.2: Basic quantities in the Random Group Formation (RGF) predictions of the sports systems. M: the total score of the system; N: the total number of players; k_{max}: the highest score; k_0: the lowest score. A, γ, b, and k_c refer to four parameters in the procedure of the RGF prediction [101]. D and p denote the maximum differences D and p-values in the KS tests, while "BWF M" and "BWF W" mean "BWF Men" and "BWF Women", respectively.

Sports	M	N	k_{max}	k_0	A	γ	b	k_c	D	p
ATP	227193	1763	7965	6	1.044	1.44	3.91E-4	6138	0.5098	0.000
WTA	276279	1523	8835	10	1.261	1.42	3.68E-4	6864	0.6667	0.000
BWF M	4777609	548	81706	110	0.0018	0.363	7.72E-5	69345	0.9048	0.000
BWF W	5191108	833	89002	40	0.0169	0.661	6.37E-5	74812	0.7619	0.000
PGA	30690	1323	384	1	0.158	0.793	0.0149	325	0.6190	0.000
LPGA	22779	734	590	1	0.174	0.919	0.0079	483	0.7143	0.000
ITTF M	195176	1717	2706	20	3.045	1.501	0.0015	2193	0.8095	0.000
ITTF W	180106	1288	2728	23	1.9421	1.373	0.0015	2225	0.7097	0.000
FIVA M	2626	138	210	1	0.180	0.826	0.0176	163	0.5238	0.004
FIVA W	2411	127	200	1	0.174	0.803	0.0185	155	0.4516	0.002
FIBA M	6921	79	892	1	0.090	0.755	0.0037	665	0.4762	0.011
FIBA W	6976	72	940	1	0.082	0.733	0.0035	699	0.5161	0.000
FIH M	36964	73	2620	30	0.0030	0.058	0.0020	2122	0.6153	0.000
FIH W	36079	68	2700	35	0.0029	0.065	0.0019	2180	0.8571	0.000
IHF M	2600	52	286	1	0.0381	0.265	0.0152	224	0.7095	0.000
IHF W	2326	46	261	1	0.0283	0.141	0.0173	205	0.8182	0.000
FIE M	8593	319	290	1	0.1137	0.622	0.0174	239	0.5806	0.000
FIE W	9149	371	294	1	0.1336	0.696	0.0168	242	0.6364	0.012
IBAF M	7377	78	986	1	0.091	0.771	0.0033	731	0.7273	0.000

maximum differences D and p-values in the KS tests for the 19 sports systems. From Tab. 2.2, one could find all the p-values are much smaller than 0.05, which suggests all the KS tests reject the null hypothesis at the 5% significance level. Therefore, we can draw the conclusion that the RGF predictions could not be used to characterize the sports data.

The possible reason is that the data samples of the sports systems are quite small, which might lead to large uncertainty. We also conjecture that the differences between the two kinds of systems might be caused by different mechanisms of formation. For sports systems, the competition is the main driven force. Whether a player's rank will be upped or lowered, depends not only on his own performance but also on other's. In sports there is not much "rich-gets-richer" mechanism, which is dominant in city sizes, human wealth and etc, however.

2.3 Pareto principle

The Pareto principle [104], also well known as the 80-20 rule, states that, for many events, roughly 80% of the effects comes from 20% of the causes. Pareto noticed that, 80% of Italy's land was owned by 20% of the population. He carried out such surveys on a variety of other countries further, and to his surprise, the rule was also fulfilled.

The 80-20 rule has also been used to attribute the widening economic inequality, which showed that, the distribution of global income to be very uneven, with the richest 20% of the world's population controlling 82.7% of the world's income. The 80-20 rule could be applied to many systems, from the science of management to the physical world. The 80-20 rule seems to be almost an universal truth and can be applied to practically all aspects of management and even to our personal lives. When used correctly, Pareto analysis is a powerful and an effective tool for making continuous improvement and in problem solving. Continued application of this rule will greatly improve productivity, quality and profitability.

We also check this rule in the sports ranking systems. It is interesting to find that, 20% players indeed possess approximately 80% scores or prize money of the whole system. The ratios obtained from different sports ranking systems are shown in Tab. 2.1, values of the ratios being all very close to 0.8. This suggests the imbalance in the sports systems, exactly how this rule emerges in sports with different rules, governing bodies and tournament structures is something of a puzzle. However, it means there is certain predictability in the outcome of events in which two players are pitted against each other.

2.4 Dependence of win probability on Δ rank

Here we employ the concept of "win probability" to describe the chances that a player or a team will win when encountering an opponent. For instance, what is the odds that a No.1 player will top a No.100 player? What is again her chance against No.2? Theoretically, the chance is much higher in the former case than is in the latter one. But the result of a competition is not unknown until it is over, which mainly depends on how the player performs at that specific match. However, the win probability could be solely based on the previous performance of a player against a certain opponent,

which then can used to predict her future performance against the same opponent. This might have some applications in betting the result of a match. To simplify the case without loss of generality, we relate the win probability solely to the rank difference of a pair of players. Suppose we now have two players A and B, with A having a higher rank. We will then need to know how likely A can beat B when they meet. This quantity is related to but different from the win percentage we usually refer to. The win percentage depicts the percentage of win of a player over all previous encounters. We assume that the win probability only depends on the rank difference between two players. This means, the probability that No.1 beats No.100 is the same as the one that No.100 beats No.200. Hence, we have the following definition,

$$P_{win}(\Delta r) = \frac{N_{win}(\Delta r)}{N_{total}(\Delta r)}, \tag{2.4}$$

where Δr denotes the rank difference (integer), $N_{win}(\Delta r)$ is the total number of win for the higher-ranked players when the rank difference is Δr, and $N_{total}(\Delta r)$ is the total number of matches in which the rank difference between the pair is Δr. We here emphasize again that the win probability is the probability that the higher-ranked player will win when two players meet. When Δr is small, say 1, it is difficult to judge which player will win, and in this case P_{win} might approximately equal 0.5. When Δr is large, for instance 100, P_{win} might approach 1, which means the higher-ranked player is very likely to win.

By using the *Head to Head* records of ATP and WTA, we find that the dependence of P_{win} on Δr can be well characterized by the Bradley-Terry model [105] for paired comparisons as follows,

$$P_{win} = \frac{1}{1 + \exp\left(-a * \Delta r\right)}, \tag{2.5}$$

where a is a parameter dependent on the specific systems. For ATP and WTA, a is 0.021 and 0.032, respectively (Fig. 2.2). The existence of fluctuations is quite natural since even Roger Federer will not win all the matches. The value of a can still tell us some information about how competitive that certain sports is. The smaller a is, the more competitive the sports will be. Let us take WTA and ATP as two examples. When Δr is 30, the win probability for WTA is nearly 0.7, while the counterpart for ATP is 0.65. This means the game is more unpredictable in ATP than in WTA, it is not strange since men's game is more competitive than women's.

The competitiveness parameter a plays a key role in both empirical analysis of the

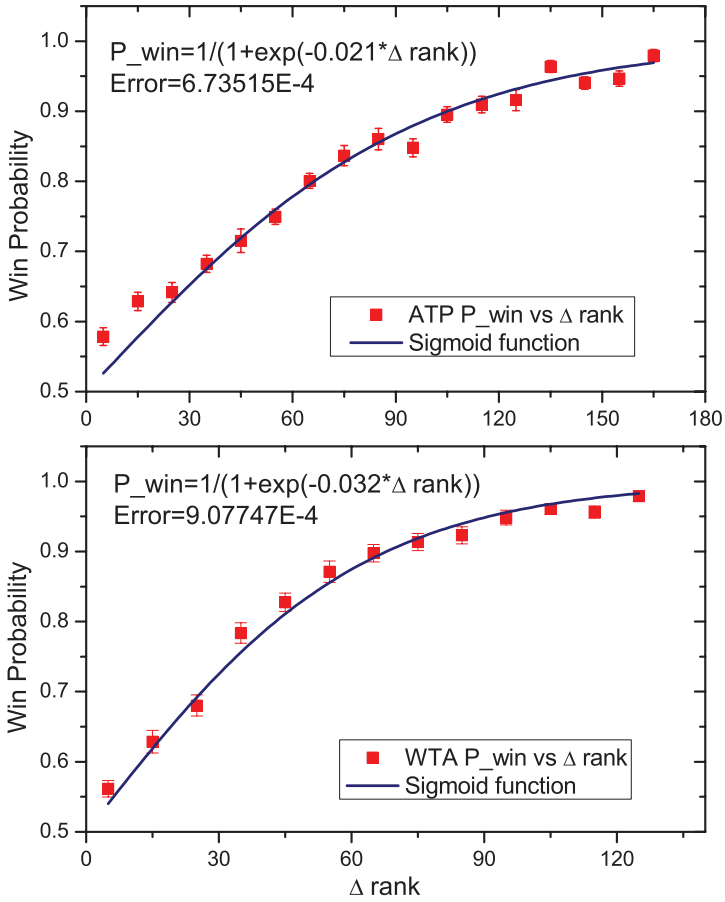

Fig. 2.2: (Color online) Dependence of win probability on Δr of the players for ATP and WTA, which can be well fitted to the sigmoidal function $P_{win} = 1/(1 + \exp(-a * \Delta r))$, with a=0.021 and 0.032 for ATP and WTA, respectively.

win probability and the simulations of the toy model, so we explain the differences between different systems in two respects.

For the empirical part, we really wish to test the empirical finding by checking data from different sports fields, other than in the tennis field of ATP and WTA. The problem is that the data source of the *Head to Head* records is very limited in other sports fields, to our best knowledge. Alternatively we present here the trend of the functional form of the win probability in Fig. 2.3, in which, $a = 0.01, 0.015$ and 0.03, may correspond to three different sports systems. As one can see, for the same $\Delta rank$, the competitions become stronger when a gets smaller.

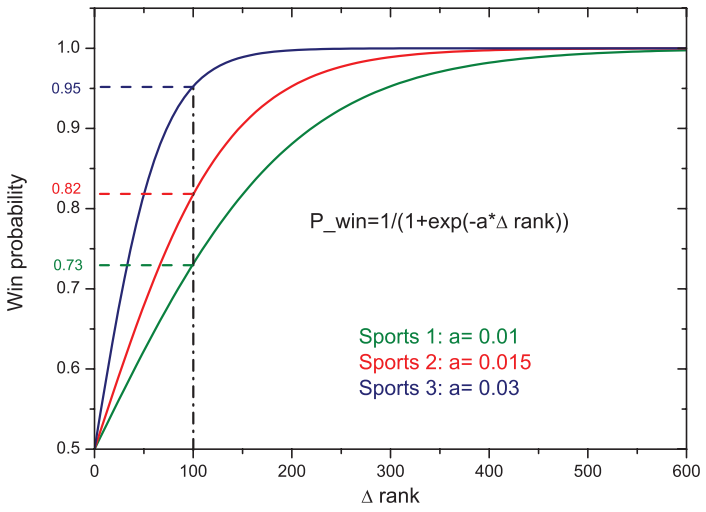

Fig. 2.3: (Color online) Theoretical curves of the win probability formula, $P_{win} = 1/(1 + \exp(-a * \Delta r))$, with a=0.01, 0.015 and 0.03 for three different sports systems, respectively.

2.5 An agent-based model for sports systems

What is the origin of the universal scaling in different sports systems? Of course, there have been so many approaches which can explain the origin of power-laws. Some mechanisms or theories are elegant, e.g., random walks [40], and self-organized criticality (SOC) [106,107], etc. It is, however, difficult to try to apply these frameworks to sports ranking systems. We propose an agent-based model, inspired by tennis. Of course, the model may not suit any sports field but does have some general implications. Most importantly, our model can reproduce robust power-laws without having to introduce additional parameters.

2.5.1 Mechanisms of the model

The rules of the model are defined in the following way (Fig. 2.4),

(1) 2^N players are ranked from 1 to 2^N, being assigned random scores drawn from a Gaussian distribution.

(2) For each tournament, all the players have entry permission. Therefore the draw will include 2^N players and in total N rounds. At each round, half of the players will be eliminated when they lose. The rest will enter the next round. The losers at round

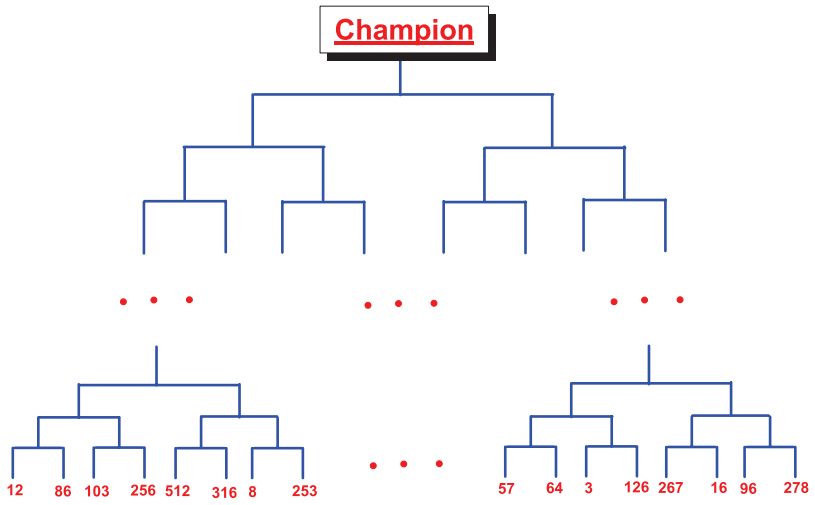

Fig. 2.4: (Color online) A cartoon of a draw sample. After each round, half players will be eliminated, the numbers "12, 86 ..." denote the ranks of the players.

n will gain score $2^{(n-1)}$. The final champion wins score 2^N.

(3) The key mechanism is to decide which one will lose for a given pair of players. Here our empirical finding will be employed. Namely, when two players meet, the probability that the higher-ranked player will top the lower-ranked opponent is given by $1/(1+\exp(-a*\Delta r))$, where Δr is their rank difference, as before.

(4) A new tournament opens up and a new draw is made.

In principle, there is only one parameter in our model, that is a. We can simply call it competitiveness parameter. Of course, there are some shortcomings in the model. First, in the actual tournaments not all the players will be accepted. In grand slams there are only 128 players. Second, tournaments can be divided into many categories and may consist of different players. Third, the scoring systems for different tournaments are a little different. For grand slams the scores and prize money are much higher than other tournaments, if the players are eliminated at the same round. We certainly can add these issues into our model in order to test the resilience of the model. At the moment we do not wish to complicate the model by introducing additional parameters. What we need here is a skeleton which may allow us to understand some key features of the specific systems. Namely, if the power-laws with exponential decay can be reproduced through our model, then it is a feasible model.

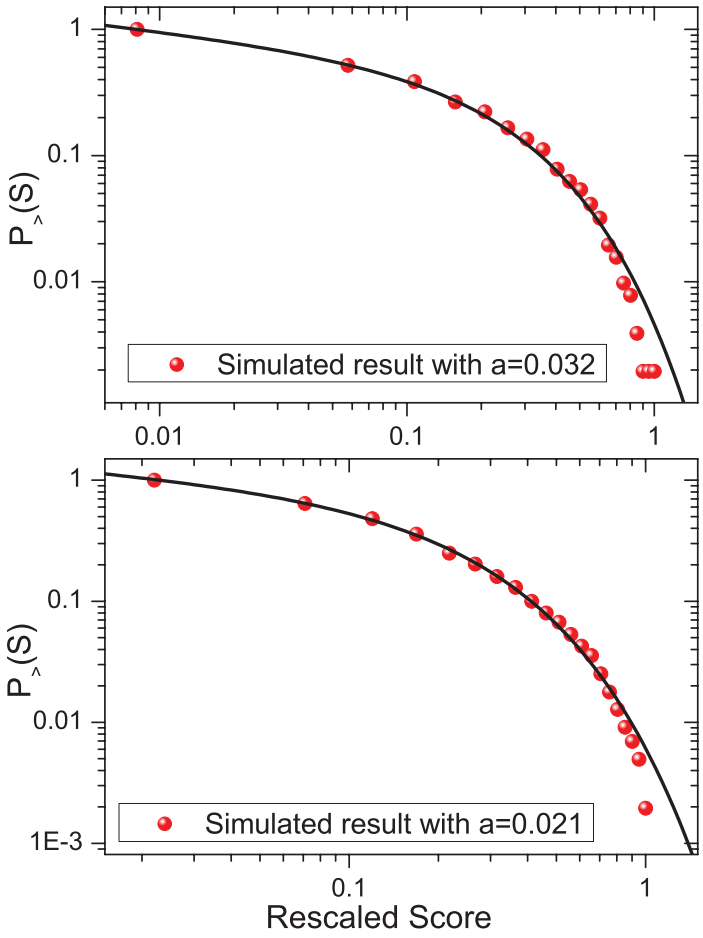

Fig. 2.5: (Color online) Cumulative distribution of scores from the simulation. For these two samples, number of players $N_p = 2048$, and number of total tournaments $N_t = 128$, with $P_{win} = 1/(1 + \exp(-a * \Delta r))$, $a = 0.032$ and 0.021, respectively.

2.5.2 Simulation results and discussions

The most important parameter in our model is a, the so-called competitiveness parameter. The number of players N_p and the number of tournaments N_t only have finite-size effects. It is natural to check the dependence of the simulation results on these parameters, which can reflect the resilience of our model.

First of all, we need to test whether the model can reproduce the power-laws of the cumulative distribution of scores. In Fig. 2.5, N_p equals 2048, and N_t is 128, while win probability, $P_{win} = 1/(1 + \exp(-a * \Delta r))$, with $a = 0.021$ and 0.032, as given by the empirical data of ATP and WTA, respectively. Here, we also use the same goodness-

of-fit test, and p-value equals 0.85 and 0.91 for the two distributions, respectively. Therefore, the cumulative distributions of scores given by the simulations indeed follow the power-law distributions with exponential decay, $P(S) \propto S^{-\tau}\exp(-S/S_c)$, with $\tau = 0.2, 0.22$, $S_c = 0.23, 0.19$, respectively for these two samples. Here, we notice that the values of the parameters are very close to what are obtained from the empirical data.

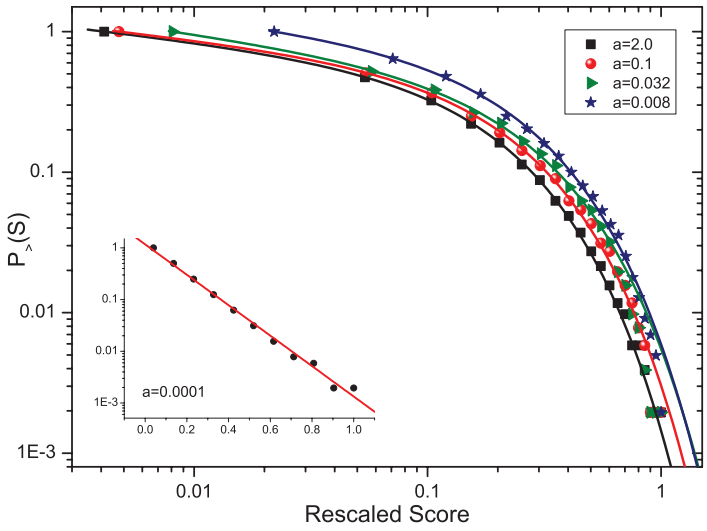

Fig. 2.6: (Color online) Influence of the critical parameter a on the final cumulative scores distributions, values of a ranging from 0.0001 to 2.0.

2.5.3 Robustness of the model

In the formula of win probability, smaller values of a correspond to more intensive competition. For instance, when $a = 0.0001$, $P_{win} \leqslant 0.525$ for $\Delta r \leqslant 1000$, which means the higher-ranked player only has slightly more chances than the lower-ranked player to win the match between them. While larger values of a suggest that the higher ranked players would win the match with a much larger probability. For example, when $a = 2.0$, $P_{win} \geqslant 0.88$ for $\Delta rank \geqslant 1$.

Thus here, to analyze the influence of win probability, we simulated our models with different values of a, $0.0001 \leqslant a \leqslant 2.0$. From Fig. 2.6, we can find that, as a gets smaller, the values of τ will become larger, while those of S_c will become smaller. When a is very small, such as $a = 0.0001$, the cumulative scores distributions change

from the power laws with exponential decay to exponential. Since in this case, all players nearly win the match randomly, thus the cumulative probabilities of the scores approximates 1, 1/2, $(1/2)^2$, ..., which results in the exponential distribution.

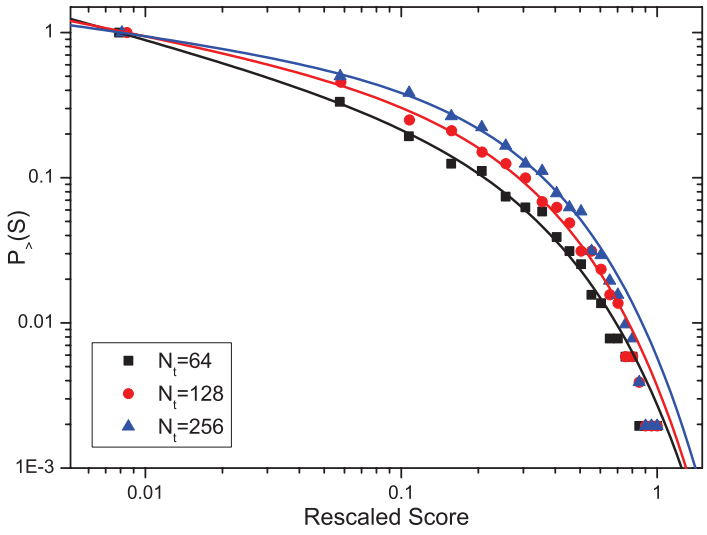

Fig. 2.7: (Color online) Simulation result of the cumulative scores distributions for different number of tournaments, with N_t=64, 128, and 256.

For different number of tournaments, $N_t = 64, 128$ and 256, the cumulative distributions of scores are shown in Fig. 2.7. As seen, all the cumulative distributions of scores are power-laws with exponential decay, values of the exponents τ and S_c being also very close to those of the empirical results.

In statistical physics, in order to determine the validity of the statistical approach, we often take the thermodynamic limit, in which the number of components N tends to infinity [108]. However, in real-world networks, the number of vertices or agents can never be that large and therefore we need to study the finite-size effect. For example, even the largest artificial net, the World Wide Web, whose size will soon approach 10^{11}, also shows qualitatively strong finite-size effect [109].

Therefore, here, in order to test the influence of the finite-size effect on the final cumulative distribution of scores, we consider the transformed score distribution $P(S) * S^{\tau}$ versus S/S_c, where S_c is the characteristic turning point of the exponential decay. For four different system sizes, such relationships were shown in Fig. 2.8, which suggests that, the tails of the four curves almost collapse with each other.

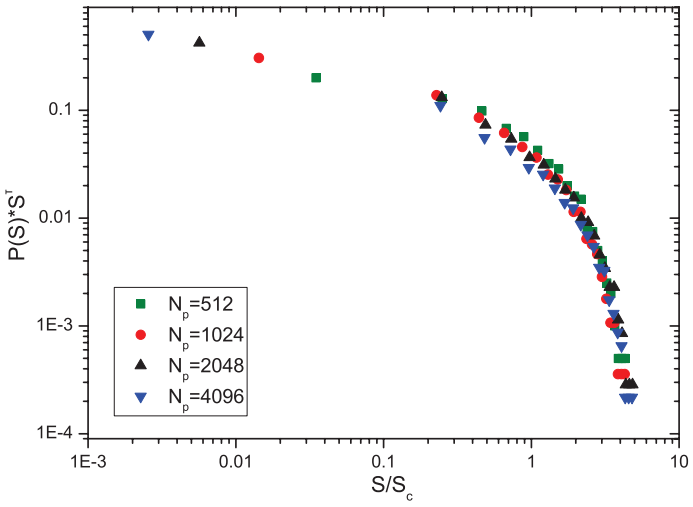

Fig. 2.8: (Color online) Finite-size effects analysis of the simulation results, with N_p=512, 1024, 2048 and 4096.

We also considered the influence of the players' initial score distributions on the final simulation results. For the same number of players, $N_p = 1024$, the same number of tournaments, $N_t = 128$, and the same win probability, $P_{win} = 1/(1 + \exp(-a * \Delta r))$ with $a = 0.015$, we conducted the simulations on several different initial score distributions, e.g., the uniform distributions, the Gaussian distributions with different standard deviations (*Mean value=50, Sigma=1, 3, 7, 11*). The simulation results are shown in Fig. 2.9, as one could observed, all the cumulative distributions of scores from the simulations are almost identical, and they could be characterized by $P(S) \propto S^{-\tau} \exp(-S/S_c)$, with $\tau = 0.23$, $S_c = 0.11$. Therefore, the initial score distributions have little influence on the final simulation results. The reasons behind should due to that, the awards to the winners accumulate in each round of tournaments, and this effect results that the final scores of players are much larger than their initial ones (*Mean value=50*).

As the major goal of our model is that it could reproduce the trend of empirical finding of cumulative score distributions. Therefore the predictive power of the model is rather modest. We don't think it could be a general framework for all kinds of sports systems. However, we are plotting of enriching the model by considering more ingredients so that the model could be more powerful. Of course in doing so we might have to consider the cost of introducing additional parameters.

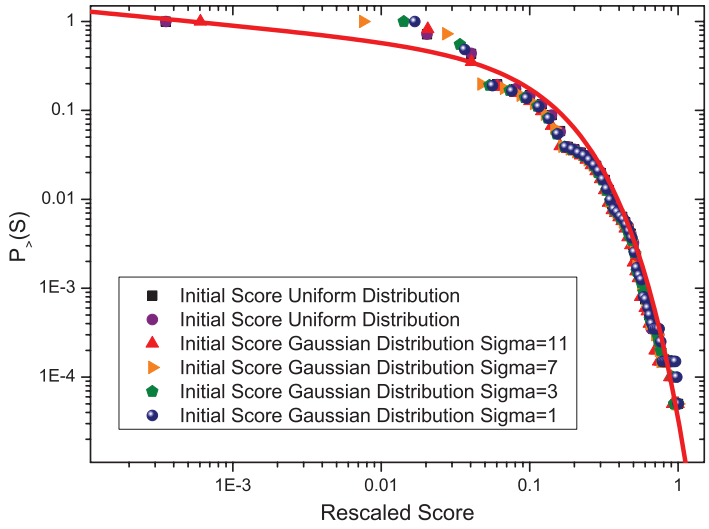

Fig. 2.9: (Color online) Different initial score distributions, e.g., the uniform distributions, the Gaussian distributions with different standard deviations, (*Mean value=50, Sigma=1, 3, 7, 11*), and their influences on the final simulation results.

2.6 Conclusion

Ranking is a direct measure of the individuals' performance in the whole system, in order to characterize the intrinsic common features and underlying dynamics of ranking systems, we chose to analyze the rankings in a specific kind of systems, i.e., sports systems, in which players or teams are ranked by their scores or prizes. Our concrete results concern: (i) Universal scaling is found in the distributions of scores and/or prize money, with values of the power exponents being close to each other for 40 samples of 12 sports ranking systems. (ii) Players' scores are found to obey the Pareto principle, which means, approximately 20% of players possess 80% of total scores of the whole system. (iii) Win probability is introduced to describe the chance that a higher-ranked player or a team will win when meeting a lower-ranked opponent. We relate the win probability solely to the rank difference Δr, and for tennis the win probability has been empirically verified to follow the sigmoid function, $P_{win} = 1/(1+\exp(-a*\Delta r))$. (iv) By employing the empirical features of win probability, we propose an agent-based model to simulate the process of the sports systems, and the universal scaling could be well reproduced by our model. And this result is quite robust when we change the values of parameters in the model.

Chapter 3

Explaining Zipf's Law via Mental Lexicon

3.1 Zipf's law for natural and artificial languages

For a given text in natural or artificial languages, if we order and normalize the frequency series of words in the text,

$$\{f_r\}_{r=1}^{n} \ , \ f_1 \geq f_2 \geq ... \geq f_n \ , \ \sum\nolimits_{r=1}^{n} f_r = 1 \ , \tag{3.1}$$

f_r is the normalized frequency of the word with rank r, n is the number of different words, the Zipf's law [110, 111] states that the normalized frequency f_r is inversely proportional to the rank r,

$$f_r = cr^{-\gamma}, \quad \gamma \approx 1.0, \tag{3.2}$$

c is the prefactor, γ is the exponent, and normally $\gamma \approx 1.0$. When the rank-frequency relation is plotted in the double-log scale (Fig. 3.1), we can find the observed linear relationship is strongest in the middle range, both very high and very low frequency items deviate from the log-log regression line (they are below the Zipf curve).

The Zipf's law is the major regularity of statistical linguistics, it applies to the texts written in many natural and artificial languages. For instance, in human language families, e.g., the Indo-European language family, such as English, French, German, Spanish, Russian, Italian, etc; the Sino-Tibetan language family, such as Chinese, etc. In artificial languages, e.g., the computer programming languages, such as the modern Java, C++, C language, etc.

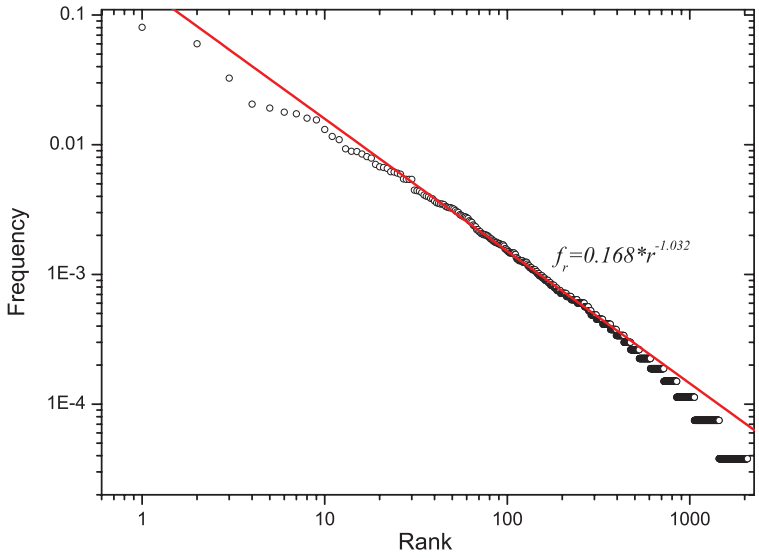

Fig. 3.1: (Color online) Log-log plot of the frequency vs. rank for one sample text. The Zipf's law holds in the middle range, both the very high and very low frequencies deviate from the log-log regression line (Zipf curve).

3.2 Origins of Zipf's law

The almost universal validity of the Zipf's law fascinated generations of scholars, however in spite of its venerable history (starting with Pareto 1897, Estoup 1916, Willis 1922, Yule, 1924) and considerable empirical support, Zipf's law remains one of the least understood phenomena in mathematical linguistics, i.e., its message is still not well understood: is it just a consequence of simple statistical regularities [112, 113], or it reflects a deeper structure of the text [114]? Many approaches were proposed for deriving the Zipf's law suggesting that it can have different origins, they can be divided into two groups.

3.2.1 Language as a media of communication

These theories deduce the Zipf's law from certain general premises of the language:

(1) The Zipf's idea that the language trades-off between maximizing the information transfer and minimizing the speaking-hearing effort [110]. Since this idea accounts for multi-functionality and short length of the most frequent words, it is so far still not conclusive [115, 116].

(2) The language employs words via the optimal setting of [algorithmic] information theory [112, 117, 118].

(3) Due to the competition of meanings, the derivation of the law is based on the idea that the words organize into hierarchical structure, where the most frequent words are the ones with wider meanings [119].

The general problem of derivations from this group is that explaining the Zipf's law for the language (and verifying it for a frequency dictionary) does not yet mean to explain the law for a concrete text, where the frequency of the same word varies widely from one text to another and is far from its value in a frequency dictionary [121].

3.2.2 Probabilistic model

The law can be derived from certain probabilistic models [113, 120, 122–125].

(1) Albeits some of these models assume relevance for realistic text-generating processes [123, 124], their a priori assumed probability structure is intricate, hence the question "why is there the Zipf's law?" translates into "why is there a specific probabilistic model?"

(2) There also belong derivations from various generalizations of the maximum entropy method [120, 121]. Here, however, the choice of the function to be maximized (and of the relevant constraints) is not clear, in contrast to the original method.

(3) Yet by far most known probabilistic model is a random text, where words are generated through random combinations of letters and the space symbol seemingly reproducing the $f_r \propto r^{-1}$ shape of the law [112, 113]. But the reproduction is elusive, since the model leads to a huge redundancy—many words have the same frequency and length—features absent in normal texts. A recent study outlines in detail the statistical differences between random and usual texts and review the previous literatures [126].

3.3 Motivations and Methods

Our approach for deriving the Zipf's law also uses a probability model, but it differs from previous models in several respects. First, it explains the law for a single text together with its limits of validity, i.e. together with the range of ranks where it holds. It also explains the rank-frequency relation for very rare words (hapax legomena) and

Table 3.1: Parameters of the texts: the total number of words N, the number of different words n, the lower $r_{\min}$ and the upper $r_{\max}$ ranks of the Zipfian domain, the fitted values of c and γ, and the difference d between the total frequency of the Zipfian domain got empirically and its value according to the Zipf's law: $d = \sum_{k=r_{\min}}^{r_{\max}}(ck^{-\gamma} - f_k)$. TF & TM means joining the texts TF and TM.

Texts	N	n	$r_{\min}$	$r_{\max}$	c	γ	$\lvert d\rvert$
TF	26624	2067	36	371	0.168	1.032	0.00333
TM	31567	2612	42	332	0.166	1.041	0.01004
AR	22641	1706	32	339	0.178	1.038	0.00048
DL	24990	1748	34	230	0.192	1.039	0.02145
TF & TM	54191	3408	30	602	0.139	1.013	0.02091
TF & AR	45265	2656	33	628	0.138	0.998	0.00239
TF & DL	47614	2877	28	527	0.162	1.014	0.01490
TM & AR	54208	3184	43	592	0.157	1.021	0.00491
TM & DL	56557	3154	45	493	0.161	1.023	0.01211
AR & DL	47631	2550	38	496	0.165	1.012	0.00947
Four texts	101822	4047	39	927	0.158	1.015	0.00187

relates it to the Zipf's law. Second, the a priori structure of our model get explained via Bayesian statistics, a branch of probability theory that is involved with explaining and interpreting the meaning of prior probability. For our situation, the a priori structure of our model relates to the general features of mental lexicon [127] of the author who produced the text. Third, the model is not *ad hoc*: though it is new in its entirety, its elements were already used successfully for text modelling, i.e., it is based on the latent semantic analysis.

3.4 Validity range of the Zipf's law

. Before introducing the model, we need to clarify the applicability of the Zipf's law. As is well known, Zipf's law applies mainly for the middle range of ranks, below we present the empirical results that clarify the valid range of the Zipf's law, confirm some known results, but also make several new points that motivate the theoretical model worked out in the sequel.

3.4.1 Database of English texts

We studied, in particular, four English texts written in different genres and epochs and having few thousands different words [see Tab 3.1]. This size is large enough to make the frequencies of words stable, but it is short enough for the text to have a well-defined meaning. Such texts enforce the understanding of the law, when we contrast the separate texts to their mixtures.

-*The Age of Reason* (AR) by T. Paine, 1794 (the major source of British deism, $N = 22641$ words).

-*Thoughts on the Funding System and its Effects* (TF) by P. Ravenstone, 1824 (economics, $N = 26624$ words).

-*Time Machine* (TM) by H. G. Wells, 1895 (a science fiction classics, $N = 31567$ words).

-*Dream Lover* (DL) by J. MacIntyre, 1987 (a romance novella, $N = 24990$ words).

3.4.2 Summary of empiric findings

We employ the linear fitting method to clarify the valid range of the Zipf's law ($r \in [r_{\min}, r_{\max}]$), to get the values of r_{min}, r_{max}, and the corresponding values of c and γ. For detail explanations, please refer to Appendix A. Followings are a list of the empiric results after using the linear fitting method:

1. For each text there is a specific (Zipfian) range of ranks $r \in [r_{\min}, r_{\max}]$, where the Zipf's law holds with $\gamma \approx 1$ and $c < 0.2$ [110, 111]; see Tab 3.1. Both for $r < r_{\min}$ and $r > r_{\max}$ the law is invalid, since the frequencies *are below* the Zipf curve (apart of very small exclusions, see Fig. 3.2) [110, 111].

2. Even if the same word enters into different texts it typically has quite different frequencies there [121], e.g. among 83 common words in the Zipfian ranges of AR and DL, only 12 words have approximately equal ranks and frequencies (most of them are function words).

3. The pre-Zipfian $1 \leq r < r_{\min}$ range contains mainly function words. They serve for establishing grammatical constructions (e.g., *the, a, such, this, that, where, were*). But the majority of words in the Zipfian range do have a narrow meaning (content words). A subset of those content words has a meaning that is specific for the

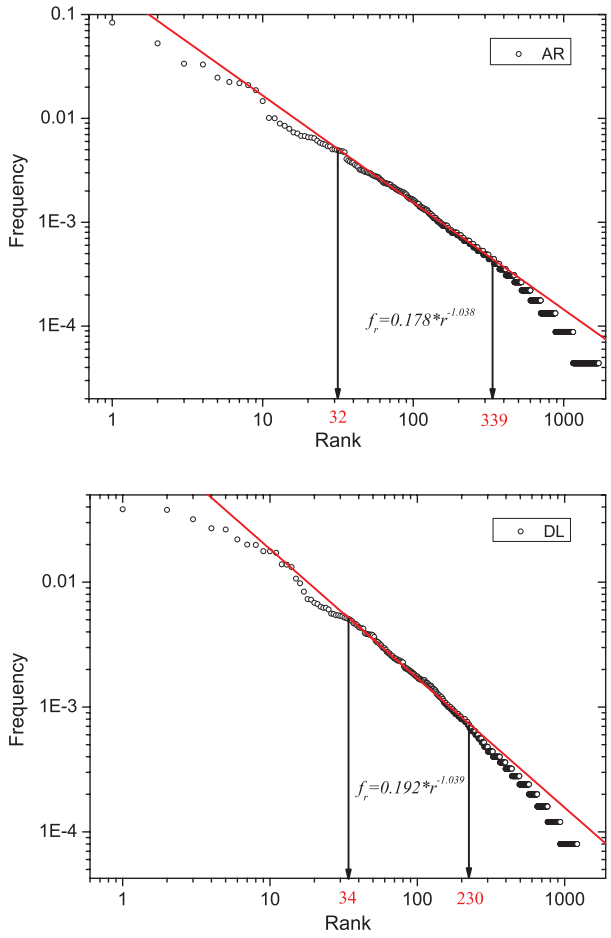

Fig. 3.2: (Color online) Frequency vs. rank for the texts *Age of Reason* (AR), *Dream Lover* (DL). Arrows and red numbers indicate on the range, where the Zipf law holds. The red line is the Zipf's law curve obtained from the fittings, $f_r = 0.178r^{-1.038}$ and $f_r = 0.192r^{-1.039}$.

text and can serve as its keywords. This is known and is routinely used in document processing [152]. (We will explain why the majority of key-words appear in the Zipfian domain in the solutions of our model.)

Few keywords appear also in the pre-Zipfian range, e.g. *love* and *miss* for DL and *god* and *man* for AR. Some keywords are also located in the post-Zipfian area, e.g. *eloi* for TM, but the majority of them are in the Zipfian range.

To confirm that the words from the Zipfian range relate to the meaning of the text, we excluded from our texts all the words from the third (post-Zipfian) range, and saw that not only the rough meaning of the text stayed intact, but also its basic conceptions and its deeper, intrinsic message (AR and TM do have such a message).

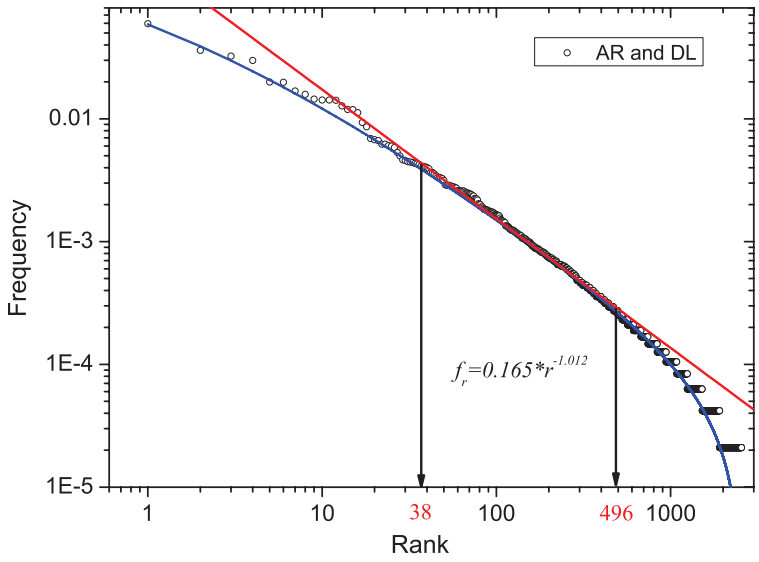

Fig. 3.3: (Color online) Frequency vs. rank for the joint text *Dream Lover* & *Age of Reason*. Red line: the Zipf curve $f_r = 0.165r^{-1.012}$. Arrows and red numbers indicate on the validity range of the Zipf's law. Blue line: the numerical solution of (3.11, 3.12) for $c = 0.165$. It coincides with the generalized Zipf law (3.17) for $r > r_{\min} = 38$. The step-wise behavior of f_r for $r > r_{\max}$ refers to hapax legomena.

4. The absolute majority of different words with ranks in $[r_{\min}, r_{\max}]$ have different frequencies. Only for $r \simeq r_{\max}$ the number of different words having the same frequency is $\simeq 10$. For $r > r_{\max}$ we meet the hapax legomena effect: words occurring only few times in the text ($f_r N$ is a small integer), and many words having the same frequency f_r [111]. The effect is not described by a smooth rank-frequency relation, including the Zipf's law. Generalizations of the law that do account for hapax legomena are reviewed in [111].

5. The minimal frequency of the Zipfian domain holds

$$f_{r_{\max}} > c/n. \tag{3.3}$$

We checked that (3.3) is valid not only for separate texts but also for the frequency dictionaries of English and Irish. For our texts a stronger relation holds $f_{r_{\max}} \gtrsim \frac{1}{n}$.

6. When joining (mixing) two texts (A and B), the word frequencies get mixed: $f_k(\text{A\&B}) = \frac{N_\text{A}}{N_\text{A}+N_\text{B}} f_k(\text{A}) + \frac{N_\text{B}}{N_\text{A}+N_\text{B}} f_k(\text{B})$, where N_A and $f_k(\text{A})$ are, respectively, the total number of words and the frequency of word k in the text A.

After joining two texts the range of the Zipf's law increases mainly via acquiring

more higher rank words, i.e. $r_{\min}$ stays approximately fixed, while $r_{\max}$ increases; see Tab 3.1. For instance, the Zipfian interval $\mathcal{Z}$[DL] ($\mathcal{Z}$[AR]) of DL (AR) contains 10 (30) words that do not enter to $\mathcal{Z}$[DL+AR] (the Zipfian range of the joint text), but instead $\mathcal{Z}$[DL+AR] has 93 new words that appear neither in $\mathcal{Z}$[DL] nor $\mathcal{Z}$[AR]. Hence the Zipfian range of the joint text increases.

The exponent γ gets closer to 1 and the prefactor c decreases. Also the overall precision of the Zipf's law increases; see Tab 3.1. The valid range of the law increases, since the Zipfian domains of different texts contain *mainly* different words [see **2**]. After joining, these domains combine. In particular, the keywords stay in the Zipfian range, e.g. after joining all four above texts, the keywords of each text are still in the Zipfian range (which now contains almost 900 words).

This feature of the law is consistent with statistical modelling. For this conclusion it is essential to account for its validity limits $[r_{\min}, r_{\max}]$, otherwise one can get the opposite (and incorrect) conclusion announced in [114].

3.5 Probabilistic Model

3.5.1 Three features of the model

According to the above empiric findings, a model for the Zipf's law is supposed to satisfy the following features:

(1) Apply to separate texts, i.e. explain how and why different texts can satisfy the same form of the rank-frequency relation despite the fact that the same words do *not* occur with same frequencies in the different texts.

(2) Derive the law in its totality, the prefactor c, the exponent γ, together with its extensions for all frequencies, limits of validity and hapax legomena effect.

(3) Relate the law to formation of a text.

3.5.2 Descriptions of the model

Two sources of the model are the latent semantic analysis [131], and the idea of applying ordered statistics for rank-frequency relations [117, 133, 134]. The descriptions ($\mathbf{A}-\mathbf{D}$) of the model are as follows:

A. The *bag-of-words picture* focusses on the frequency of the words that occur in a text and neglects their mutual disposition (i.e. syntactic structure, grammar structure, etc) [132]. Given n different words $\{w_k\}_{k=1}^n$, the joint probability for w_k to occur $\nu_k \geq 0$ times in a text T is multinomial

$$\pi[\boldsymbol{\nu}|\boldsymbol{\theta}] = \frac{N!\,\theta_1^{\nu_1}...\theta_n^{\nu_n}}{\nu_1!...\nu_n!}, \quad \boldsymbol{\nu} = \{\nu_k\}_{k=1}^n, \;\; \boldsymbol{\theta} = \{\theta_k\}_{k=1}^n, \tag{3.4}$$

where $N = \sum_{k=1}^n \nu_k$ is the length of the text, ν_k is the number of occurrences of w_k, and θ_k is the probability of w_k. The picture is well-known in computational linguistics and produces reasonable results for document classification [132]. But for our purposes this picture is incomplete, because it implies that each word has the same probability for different texts.

B. To improve this point we make $\boldsymbol{\theta}$ a random vector [132] with a text-dependent density $P(\boldsymbol{\theta}|T)$. The simplest assumption is that $(T, \boldsymbol{\theta}, \boldsymbol{\nu})$ form a Markov chain: the text T influences the observed $\boldsymbol{\nu}$ only via $\boldsymbol{\theta}$. Then the probability $p(\boldsymbol{\nu}|T)$ of $\boldsymbol{\nu}$ in a given text T reads

$$p(\boldsymbol{\nu}|T) = \int \mathrm{d}\boldsymbol{\theta}\, \pi[\boldsymbol{\nu}|\boldsymbol{\theta}]\, P(\boldsymbol{\theta}|T). \tag{3.5}$$

This form of $p(\boldsymbol{\nu}|T)$ is basic for probabilistic latent semantic analysis [131], a successful method of computational linguistics. There the density $P(\boldsymbol{\theta}|T)$ of latent variables $\boldsymbol{\theta}$ is determined from the data fitting. But we shall deduce $P(\boldsymbol{\theta}|T)$ theoretically.

C. $P(\boldsymbol{\theta}|T)$ is generated from a density $P(\boldsymbol{\theta})$ via conditioning on the ordering of $\mathbf{w} = \{w_k\}_{k=1}^n$ in T:

$$P(\boldsymbol{\theta}|T) = P(\boldsymbol{\theta})\, \chi_T(\boldsymbol{\theta}, \mathbf{w}) \Big/ \int \mathrm{d}\boldsymbol{\theta}'\, P(\boldsymbol{\theta}')\, \chi_T(\boldsymbol{\theta}', \mathbf{w})\,. \tag{3.6}$$

If different words of T are ordered as $(w_1, ..., w_n)$ with respect to the decreasing frequency of their occurrence in T (i.e. w_1 is more frequent than w_2), then $\chi_T(\boldsymbol{\theta}, \mathbf{w}) = 1$ if $\theta_1 \geq ... \geq \theta_n$, and $\chi_T(\boldsymbol{\theta}, \mathbf{w}) = 0$ otherwise.

As substantiated below, $P(\boldsymbol{\theta})$ refers to the mental lexicon of the author prior to generating a concrete text.

D. For simplicity, we assume that the probabilities θ_k are distributed identically and the dependence among them is due to $\sum_{k=1}^n \theta_k = 1$ only:

$$P(\boldsymbol{\theta}) \propto u(\theta_1)\,...\,u(\theta_n)\,\delta(\sum\nolimits_{k=1}^{n}\theta_k - 1), \tag{3.7}$$

where $\delta(x)$ is the delta function and the normalization ensuring $\int_0^\infty \prod_{k=1}^n \mathrm{d}\theta_k\, P(\boldsymbol{\theta}) = 1$ is omitted.

3.5.3 Solutions and Discussions

· The conditional probability $p_r(\nu|T)$ for the r'th most frequent word w_r to occur ν times in the text T reads from (3.4, 3.5)

$$p_r(\nu|T) = \frac{N!}{\nu!(N-\nu)!}\int_0^1 \mathrm{d}\theta\,\theta^\nu(1-\theta)^{N-\nu}P_r(\theta|T), \tag{3.8}$$

$$P_r(t|T) = \int \mathrm{d}\boldsymbol{\theta}\,P(\boldsymbol{\theta}|T)\delta(t-\theta_r), \tag{3.9}$$

where $P_r(t|T)$ is the marginal density for the probability t of w_r. For $n \gg 1$, we deduce in Appendix B from (3.6, 3.7) that $P_r(t|T)$ follows the law of large numbers, it is Gaussian,

$$P_r(t|T) \propto \exp[-\frac{n^3}{2\sigma_r^2}(t-\phi_r)^2], \tag{3.10}$$

where $\sigma_r = (c+n\phi_r)\sqrt{n\phi_r}$, and the mean ϕ_r is found from two equations for two unknowns μ and ϕ_r:

$$r/n = \int_{\phi_r}^\infty \mathrm{d}\theta\,u(\theta)\,e^{-\mu\theta n}\Big/\int_0^\infty \mathrm{d}\theta\,u(\theta)\,e^{-\mu\theta n}\,, \tag{3.11}$$

$$\int_0^\infty \mathrm{d}\theta\,\theta\,u(\theta)\,e^{-\mu\theta n} = \frac{1}{n}\int_0^\infty \mathrm{d}\theta\,u(\theta)\,e^{-\mu\theta n}. \tag{3.12}$$

Eq. (3.10) holds for $P_r(t|T)$ whenever its standard deviation $\sigma_r n^{-3/2}$ is much smaller than the mean ϕ_r; as checked below, this happens already for $r > 10$.

The meaning of (3.11, 3.12) is explained via the marginal density $P(\theta_1,...,\theta_m) = \int_0^\infty \prod_{k=m+1}^n \mathrm{d}\theta_k\,P(\boldsymbol{\theta})$ found from (3.7). For $n \gg 1$ and $m \ll n$ it factorizes (see details in Appendix C):[1]

$$P(\theta_1,...,\theta_m) = \prod\nolimits_{l=1}^m P(\theta_l) \propto \prod\nolimits_{l=1}^m u(\theta_l)e^{-\mu\theta_l n}. \tag{3.13}$$

[1]Eq. (3.13) can be established via the saddle point method in Appendix C, or heuristically via the exact relation $\overline{[\sum_{k=1}^n \theta_k]^2} = 1$, where $\overline{f}$ means averaging over $P(\boldsymbol{\theta})$. This relation predicts, together with $\overline{\theta_k} = \frac{1}{n}$, that $\overline{\theta_i\theta_j} - \overline{\theta_i}\,\overline{\theta_j} = \mathcal{O}(n^{-3})$, hence approximate factorization.

Eq. (3.12) ensures that $\int_0^\infty \mathrm{d}\theta\, \theta\, P(\theta) = \frac{1}{n}$. This relation follows from $\sum_{k=1}^n \theta_k = 1$ and it determines μ, an analogue of the chemical potential in statistical physics. The interpretation of (3.11) is that it equates the relative rank r/n to the (unconditional) probability $\int_{\phi_r}^\infty \mathrm{d}\theta\, P(\theta)$ of $\theta \geq \phi_r$.

Let us study implications of (3.8–3.12) for the Zipf's law.

In (3.8), $P_r(\theta|T)$ is much more narrow peaked than $\theta^\nu(1-\theta)^{N-\nu}$, since $n^3 \gg N \gg 1$ [see Tab 3.1]. Hence in this limit we approximate $P_r(\theta|T) \simeq \delta(\theta - \phi_r)$ [see (3.10)]:

$$p_r(\nu|T) = \frac{N!}{\nu!(N-\nu)!}\phi_r^\nu(1-\phi_r)^{N-\nu}. \tag{3.14}$$

Eq. (3.14) is the main outcome of the model; it shows that the conditional probability $p_r(\nu|T)$ for the occurence number ν of the word w_r has the same form (3.14) for different text. In (3.14), ϕ_r is the [effective] probability of the word w_r. If $N\phi_r \gg 1$, $p_r(\nu|T)$ is peaked at $\nu = N\phi_r$: the frequency of a word that appears many times equals its probability. Each word of the Zipfian domain occurs at least $\nu \sim N/n \gg 1$ times. For such words we approximate $f_r \equiv \nu/N \simeq \phi_r$.

Now we postulate in (3.7)

$$u(f) = (n^{-1}c + f)^{-2}, \tag{3.15}$$

where c will be related below to the prefactor of the Zipf's law. (We will explain the meaning of Eq. (3.15) and relates it to the features of the author's mental lexicon in the following section.)

For $c \lesssim 0.2$, $c\mu$ determined from (3.12, 3.15) is small and is found from integration by parts:

$$\mu \simeq c^{-1}\, e^{-\gamma_\mathrm{E} - \frac{1+c}{c}}, \tag{3.16}$$

where $\gamma_\mathrm{E} = 0.55117$ is the Euler's constant. One solves (3.11) for $c\mu \to 0$: $r/n = ce^{-n\phi_r\mu}/(c + n\phi_r)$. For $r > r_{\min}$, $\phi_r n\mu = f_r n\mu < 0.04 \ll 1$. We get

$$f_r = c(r^{-1} - n^{-1}). \tag{3.17}$$

This is the Zipf's law generalized by the factor n^{-1} at high ranks r. This cut-off factor ensures faster [than r^{-1}] decay of f_r for large r. In literature a cut-off factor similar to $\frac{1}{n}$ is introduced due to additional mechanisms (hence new parameters) [123]. In our situation the power-law and cut-off come from the same mechanism.

Table 3.2: For the text TF, r_k are the ranks, where the number of occurences changes from k to $k+1$; $\hat{r}_k = (\frac{k}{cN} + \frac{1}{n})^{-1}$ is the theoretical predictions of r_k. The maximal relative error $\frac{\hat{r}_k - r_k}{r_k} = 0.0357$ is reached for $k = 6$.

r/k	1	2	3	4	5	6	7	8	9	10
r_k	1446	1061	848	722	611	529	474	437	398	370
$\hat{r}_k$	1414	1074	866	726	624	547	488	440	400	368

Fig. 3.3 shows that (3.17) reproduces well the empirical behavior of f_r for $r > r_{\min}$. Our derivation shows that c is the prefactor of the Zipf's law, and that our assumption on $c < 0.2$ above (3.16) agrees with observations (Tab. 3.1). For $c \gg 0.2$, (3.11, 3.12) do not predict the Zipf's law (3.17).

For given prefactor c and the number of different words n, (3.11–3.15) predict the Zipfian range $[r_{\min}, r_{\max}]$ in agreement with empirical results (Fig. 3.3).

For $r < r_{\min}$, it is not anymore true that $f_r n\mu \ll 1$. So the fuller expression (3.11) is to be used. It reproduces qualitatively the empiric behavior of f_r, see Fig. 3.3. We do not expect any better agreement theory and observations for $r < r_{\min}$: since the behavior of frequencies of the words in this range is irregular.

According to (3.14), the probability ϕ_r is small for $r \gg r_{\max}$ and hence the occurrence number $\nu \equiv f_r N$ of a words w_r is a small integer (e.g. 1 or 2) that cannot be approximated by a continuous function of r, see (3.15) and Fig. 3.3. To describe this hapax legomena range, define r_k as the rank, when $\nu \equiv f_r N$ jumps from integer k to $k+1$. Since ϕ_r reproduces well the trend of f_r even for $r > r_{\max}$, r_k can be theoretically predicted from (3.17) by equating its left-hand-side to k/N:

$$\hat{r}_k = [\frac{k}{Nc} + \frac{1}{n}]^{-1}, \qquad k = 0, 1, 2, ... \tag{3.18}$$

Eq. (3.18) is exact for $k = 0$, and agrees with r_k for $k \geq 1$ (Tab 3.2). Hence it describes the hapax legomena phenomenon (many words have the same small frequency). For $k \gg Nc/n$ we deduce from (3.18) $\hat{r}_k - \hat{r}_{k+1} \propto k^{-2}$ for the number of words having the frequency k/N. This relation, which is a crude particular case of (3.18), is sometimes called the second Zipf's law [111].

Preliminary summary:

Thus till this end, our model explained, though different texts can have different frequencies for same words, the frequencies of words in a given text follow the Zipf's law. Without additional fitting parameters and new mechanisms we recovered the generalized form of this law applicable for large and small frequencies. But why we would select (3.15), if we would not know that it reproduces the Zipf's law? We answer this question in the following section.

3.6 Mental lexicon and the apriori density

. In this section, I explain why we employ the apriori probability density (3.15), and how it relates to the stable and efficient organizations of the mental lexicon of the author who produced the text. But before doing that, I would like to introduce some preliminary knowledge about the mental lexicon.

3.6.1 Mental lexicon, theories and perspective

The mental lexicon [137,138] is a mental dictionary in our human brain which contains the information regarding a word's meaning, pronunciation, syntactic characteristics, and so on. It differs from the general static book dictionary in that it is not just a general collection of words, instead, it deals with how those words are stored, processed, activated and retrieved. Therefore, normally, there are two basic questions related to mental lexicon:

1. The organization of the mental lexicon, i.e., how words are stored in long-term memory?

2. Lexical access, how words are retrieved from the mental lexicon?

Researches are conducted in various ways to identify the exact mode that words are linked and accessed. A common method to analyze these connections is through the *lexical decision task*, in which the participants are required to respond as quickly and accurately as possible to a string of letters presented on a screen to decide if the string is a non-word or a real word.

One important theory in the mental lexicon, namely the *semantic network theory*, proposed the idea of spreading activation, i.e., the nodes in the semantic network are

activated in three ways, priming effects, neighborhood effects, and frequency effects.

- Priming effects: it accounts for the decreased reaction times of related words in a lexical decision task, for example, the word *bread* "primed" *butter* to be retrieved quicker.

- Neighborhood effects: it refers to the activation of all similar "neighbors" of a target word, while neighbors are defined as items that are highly confusable with the target word due to overlapping features of other words. For examples, the word "game" has the neighbors "came, dame, fame, lame, name, same, tame, gale, gape, gate, and gave". The neighborhood effect depicts that words with larger neighborhood sizes will have quicker reaction times in a lexical decision task.

- Frequency effects: experiments found that high frequency words were responded faster than the low frequency ones in a lexical decision task.

3.6.2 Characteristics of the apriori density

Now, I began to explain why we chose

$$P(\boldsymbol{\theta}) \propto u(\theta_1) \dots u(\theta_n)\, \delta(\sum\nolimits_{k=1}^{n} \theta_k - 1), \tag{3.19}$$

$$u(\theta) = (cn^{-1} + \theta)^{-2}, \tag{3.20}$$

as the apriori probability density for the probabilities $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ of different words $(w_1, ..., w_n)$. To avoid the awkward term "probability for probability" we shall call $P(\boldsymbol{\theta})$ likelihood. We focus on the marginal likelihood [see (3.13)]:

$$P(\theta) = (n^{-1}c + \theta)^{-2} e^{-\mu n \theta}, \tag{3.21}$$

since the marginal likelihood $P(\theta)$ determines the rank-frequency relation (3.11).

We assume that during the conceptual planning of the text, i.e. when deciding on its topic, style and potential audience, the author already chooses (at least approximately) two structural parameters: the potential number n of different words to appear there and the constant c. This is why the marginal likelihood (3.21) depends on the parameters c and n. Moreover, c (along with n) is a structural parameter of the text, since c/n separates the Zipfian (keywords dominated) range from the hapax legomena range (rare words), see point **5** of the empiric results.

Note that different words have the same marginal likelihood (3.21), or that is to say, the likelihood $P(\boldsymbol{\theta})$ is symmetric with respect to interchanging the words $w_1, ..., w_n$.

This feature relates to an experimental fact that words are stored in the mental lexicon in the same way [142]. The difference between them—e.g. whether the word is more familiar to the author, and/or used by him more frequently—can be relevant during the (later) phonologization stage of speech/text production [142]. Naturally, the above symmetry holds only for the apriori likelihood. The posterior likelihood $P(\boldsymbol{\theta}|T)$, the one that is conditioned over the written text, does not and should not have such a symmetry.

The basic reason for the words to have random (not fixed) probabilities is that the text-producing author should be able to compose different texts, where the same word can have different frequencies. Hence the likelihood $P(\boldsymbol{\theta})$ of random probabilities relates to the prior knowledge (or lexicon) of the text-generating author on the words. This concept of *mental lexicon*—the store of words in the long-time memory so that the words are employed on-line for expressing thoughts via phrases and sentences—is well-established in psycholinguistics [137]. Though there is no a unique theory of mental lexicon—there is only a diverse set of competing models [137]—some of its basic features are well-established experimentally and are employed below for explaining the choice (3.19, 3.20).

3.6.3 Relations of the apriori density and mental lexicon

Once each word w_k has to have a variable (random) probability θ_k, there should be a way for the author to change (increase or decrease) this probability, e.g. when the author decides that the word w_k is to become the keyword of the text. The ensuing relation between the probability vectors $\boldsymbol{\theta}'$ (new) and $\boldsymbol{\theta}$ (old) should be a group, since the author should be able to come back from $\boldsymbol{\theta}'$ to $\boldsymbol{\theta}$, e.g. when revising the text.

Under certain conditions, the only group that (for $n \geq 3$) is [139]:

$$\theta'_k = \frac{\tau_k \theta_k}{\sum_{l=1}^{n} \tau_l \theta_l}, \;\; \tau_k > 0, \;\; k = 1, ..., n, \tag{3.22}$$

where τ_k are the group parameters. If the author wants to increase two times the probability of the word w_1, then $\tau_1 = 2$ and $\tau_{k\geq 2} = 1$.

In interpreting those changes, we adapt (3.22) to the probability increase of a single word w_1, whose probability the author decides to increase by $\tau_1 > 1$ times. Thus, (3.22)

is applied for $\tau_2 = ...\tau_n = 1$:

$$\theta_1' = \frac{\tau_1\theta_1}{1+(\tau_1-1)\theta_1},\ \theta_l' = \frac{\theta_l}{1+(\tau_1-1)\theta_1},\ \text{for } l \geq 2. \tag{3.23}$$

The inverse of transformation (3.23) reads

$$\theta_1 = \frac{\tau_1^{-1}\theta_1'}{1+(\tau_1^{-1}-1)\theta_1'},\ \theta_l = \frac{\theta_l'}{1+(\tau_1^{-1}-1)\theta_1'}. \tag{3.24}$$

In the frequency range we are interested in, $(\tau_1^{-1}-1)\theta_1'$ can be neglected, hence (3.24) just reduces to the scaling transformation:

$$\theta_1 = \tau_1^{-1}\theta_1',\ \theta_l = \theta_l'. \tag{3.25}$$

The change of the marginal likelihood for θ_1 is deduced from (3.21, 3.25):

$$P'(\theta_1') = \frac{1}{\tau_1}P(\frac{\theta_1'}{\tau_1}) = \frac{1}{\tau_1}(\frac{c}{n}+\frac{\theta_1'}{\tau_1})^{-2}. \tag{3.26}$$

Thus, for the ratio of the new to the old likelihood of the probability θ_1' we get

$$P'(\theta_1')/P(\theta_1') = \tau_1 > 1 \text{ for } \theta_1' \gg c\tau_1/n, \tag{3.27}$$

$$= \tau_1^{-1} < 1 \text{ for } \theta_1' \ll c\tau_1/n. \tag{3.28}$$

The meaning of (3.27) is that once the author decides to increase the probability of the word w_1 by τ_1 times, this word will be τ_1 times more likely produced with the higher probabilities, and τ_1 times less likely with smaller probabilities; see (3.28). The feature is unique to the form (3.21) of the marginal likelohood, which by itself is due to the form (3.20) of $u(\theta)$. This is the mechanism that ensures the appearance of the keywords in the Zipfian range.

If $P(\theta)$ is assumed to reflect the organization of the mental lexicon, then according to (3.27, 3.28) this organization is efficient, because the decision on increasing the probability of w_1 translates to increasing the likelihood of larger values of the probability. The organization is also stable, because the likelihood at large probabilities does increase right at that amount the author planned (not more).

The message of (3.27, 3.28) closely relates (but is not completely identical) to the word-frequency effect well-known for the mental lexicon: more frequently used words are produced (recalled) more easily [137, 142, 143]. In the context of (3.27, 3.28) this implies that the words that are decided to appear with more probability (e.g. the keywords) will be more likely produced with higher probabilities.

3.7 Concluding remarks

The Zipf's law—together with the limits of its validity, its generalization to high and low frequencies and hapax legomena—relates to the stable and efficient organization of the mental lexicon of the text-producing author. Practically, we expect these schemes to be more efficient for real texts, if the prior structure of the model conforms the Zipf's law. Our derivation of the Zipf's law will motivate the usage of priors (3.13, 3.15) in the schemes of *latent semantic analysis*, to improve the performance of *probabilistic latent semantic analysis algorithms* making them more consistent with the Zipf's law. Also, the proposed methods can find applications for studying rank-frequency relations and power laws in other fields.

Chapter 4

Rank-frequency Relation for Chinese Characters

4.1 Research motivation and outline

One widely known aspect of the rank-frequency relation that holds for texts written in many alphabetical languages is the Zipf's law [110, 144, 145]. This regularity was first discovered by Estoup [146]:

$$f_r \propto r^{-\gamma} \text{ with } \gamma \approx 1. \tag{4.1}$$

The message of a power-law rank-frequency relation is that there is no a single group of dominating words in a text, they rather hold some type of hierarchic, scale-invariant organization.

However, the Zipf's law was so far found to be absent for the rank-frequency relation of Chinese characters [154–158,161], which play—sociologically, psychologically and (to some extent) linguistically—the same role for Chinese readers and writers as the words do in Indo-European languages [163–165].

Rank-frequency relations for Chinese characters were first studied by Zipf and coauthors who did not find the Zipf's law [153]. They claimed to find another power law with exponent $\gamma = 2$ [153], but this result was later on shown to be incorrect [155], since it was not based on any goodness of fit measure. It was also proposed that the data obtained by Zipf are reasonably fit with a logarithmic function $f_r = a + b\ln(c + r)$ with constant a, b and c [155]. The result on the absence of the Zipf's law was then

confirmed by other studies [156–158, 161]. All these authors agree that the Zipf's law is absent (more generally a power law is absent), but have different opinions on the (non-power-law) form of the rank-frequency relation for Chinese characters: logarithmic [155], exponential $f_r \propto e^{-dr}$ (where $d > 0$ is a constant) [156, 157, 161] or a power-law with exponential cutoff [154, 158].

The Zipf's law is regarded as a universal feature of human languages on the level of words [162] [1]. Hence the invalidity of the Zipf's law for Chinese characters has contributed to the ongoing debate on controversies (coming from linguistics and experimental psychology) on whether and to which extent the Chinese writing system is similar to phonological writing systems [167–169]; in particular, to which extent it is based on characters in contrast to words [2].

In this chapter, the rank-frequency relations for short, long, and mixtures of English and Chinese texts have been analyzed, the research outline amounts to the following items:

– The Zipf's law holds for sufficiently short (few thousand different characters) Chinese texts written in Classic or Modern Chinese [3]. Short texts are important, because they are building blocks for understanding of long texts. For the sake of rank-frequency relations, but also more generally, one can argue that long texts are just mixtures (joining) of smaller, thematically homogeneous pieces. This premise of our approach is fully confirmed by our results.

– The validity scenario of the Zipf's law for short Chinese texts is basically the same as for short English texts [4]: the rank-frequency relation separates into three ranges.

[1] Applications of the Zipf's law to automatic keyword recognition are based on this fact [152], because keywords are located mostly in the validity range of the Zipf's law. A related set of applications of this law refers to distinguishing between artificial and natural texts, fraud detection [159] *etc*; see [160] for a survey of applications in natural language processing.

[2] We stress already here that the Zipf's law holds for Chinese words [158]. This is expected and intuitively follows from the possibility of literal translation from Chinese to English, where (almost) each Chinese word is mapped to an English one (see our glossary at Appendix E for definition of various special terms). In this sense, the validity of the Zipf's law for Chinese words is consistent with the validity of this law for English texts.

[3] Reforms started in the mainland China since late 1940's simplified about 2235 characters. Traditional characters are still used officially in Hong-Kong and Taiwan.

[4] Here and below we refer to a typical Indo-European alphabetical based language as English, meaning that for the sake of the present discussion differences between various Indo-European and/or Uralic languages are not essential. Likewise, we expect that the basic features of the rank-frequency

(1) The range of small ranks (more frequent characters) that contains mostly function characters; we call it the pre-Zipfian range. *(2)* The (Zipfian) range of middle ranks (more probable words) that contains mostly content characters. *(3)* The range of rare characters, where many characters have the same small frequency (hapax legomena).

– The essential difference between Chinese characters and English words comes in for long texts, or upon mixing (joining) different short texts. When mixing different English texts, the range of ranks where the Zipf's law is valid quickly increases, roughly combining the validity ranges of separate texts. Hence for a long text the major part of the overall frequency is carried out by the Zipfian range. When mixing different Chinese texts, the validity range of the Zipf's law increases very slowly. Instead there emerges another, exponential regime in the rank-frequency relation that involves a much larger range of ranks. However, the Zipfian range of ranks is still (more) important, since it carries out some 40% of the overall frequency. This overall frequency of the Zipfian range is approximately constant for all (numerous and very different) Chinese texts we studied.

– The reason of why different authors get different results for the rank-frequency relation of Chinese characters in big mixtures has to do with the fact that this relation is necessarily not universal: it emerges out of mixing of shorter texts that hold the Zipf's law, but the result of mixing crucially depends on what is mixed and in which proportion this is done. The resulting rank-frequency relation thus looses universality for those ranks, where the Zipf's law does not hold.

4.2 Zipf's law for short texts

We studied several English and Chinese texts of different lengths and genres written in different epochs; see Tabs. 4.1, 4.2 and 4.3. Some Chinese texts were written using modern characters, others employ traditional Chinese characters. Chinese and English texts are described in Appendix H. The texts can be classified as short (total number of characters or words is $N = 1 - 3 \times 10^4$) and long ($N > 10^5$). They generally have different rank-frequency characteristics, so we discuss them separately.

analysis of Chinese characters will apply for those languages (e.g. Japanese), where the Chinese characters are used.

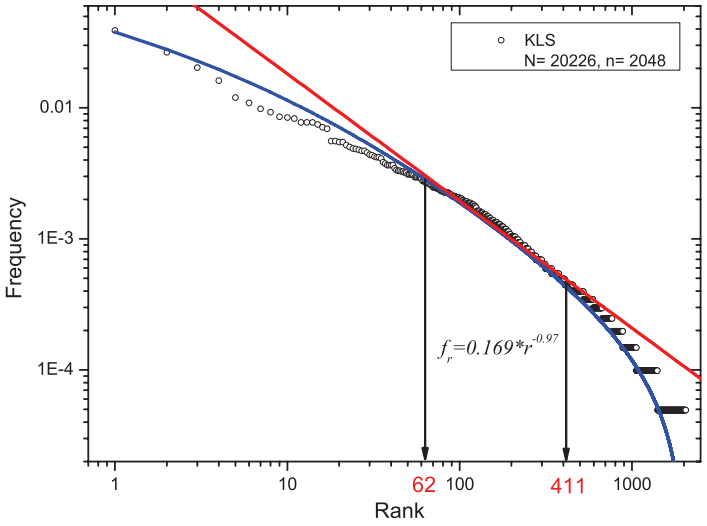

Fig. 4.1: (Color online) Frequency versus rank for the short modern Chinese text KLS; see Appendix H for its description. Red line: the Zipf curve $f_r = 0.169r^{-0.97}$ (Tab. 4.1). Arrows and red numbers indicate on the validity range of the Zipf's law. Blue line: the numerical solution of (4.3, 4.4) for $c = 0.169$. It coincides with the generalized Zipf law (4.7) for $r > r_{\min} = 62$. The step-wise behavior of f_r for $r > r_{\max}$ refers to hapax legomena.

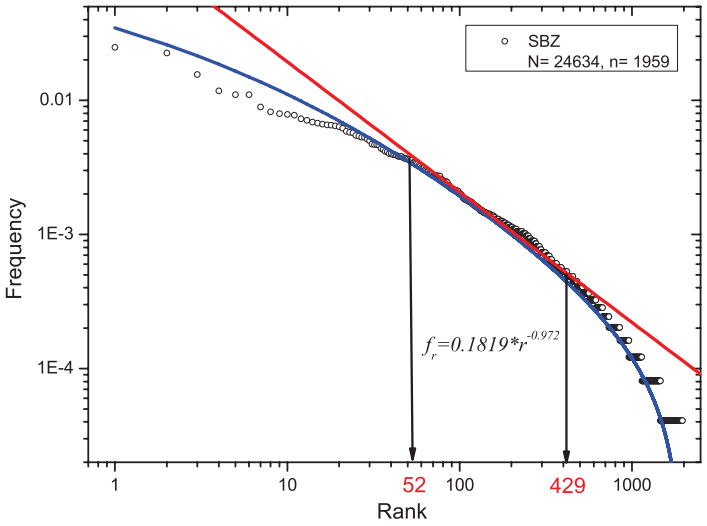

Fig. 4.2: (Color online) Frequency versus rank for the short classic Chinese text SBZ; see Appendix H for its description. Other notations have the same meaning as in Fig. 4.1.

Table 4.1: Parameters of the modern Chinese texts (see Appendix H for further details). N is the total number of characters in the text. The number of different characters is n. The Zipf's law $f_r = cr^{-\gamma}$ holds for the ranks $r_{\min} \leq r \leq r_{\max}$. Here $\sum_{k<r_{min}} f_k$ and $\sum_{k=r_{min}}^{r_{max}} f_k$ are the total frequencies carried out by the pre-Zipfian and Zipfian domain, respectively.
d is the difference between the total frequency of the Zipfian domain got empirically and its value according to the Zipf's law: $d = \sum_{k=r_{\min}}^{r_{\max}} (ck^{-\gamma} - f_k)$. Its absolute value d characterizes the global precision of the Zipf's law.
AQZ & KLS means joining the texts AQZ and KLS.
r_b is the conventional borderline rank between the exponential regime and the hapax legomena. Whenever we put "-" instead of it, we mean that either the exponential regime is absent or it is not distinguishable from the hapax legomena.

Texts	N	n	r_{min}	r_{max}	c	γ	$\sum_{k<r_{min}} f_k$	$\sum_{k=r_{min}}^{r_{max}} f_k$	$\lvert d \rvert$	r_b
AQZ	18153	1553	56	395	0.2239	1.03	0.42926	0.38424	0.00624	-
KLS	20226	2047	62	411	0.169	0.97	0.39971	0.379728	0.005728	-
AQZ & KLS	38379	2408	66	439	0.195	1.0	0.41684	0.369	0.0022	-
PFSJ	705130	3820	67	583	0.234	1.03	0.39544	0.425379	0.00842	1437
SHZ	704936	4376	78	590	0.225	1.02	0.39905	0.42	0.009561	1618

4.2.1 Empiric features of Zipf's law for short texts

We employ the linear fitting method to clarify the valid range of the Zipf's law, detailed explanations are presented in Appendix A. The followings are the results produced via the linear fitting method.

1. For each Chinese text there is a specific (Zipfian) range of ranks $r \in [r_{\min}, r_{\max}]$, where the Zipf's law $f_r = cr^{-\gamma}$ holds with $\gamma \approx 1$ and $c \lesssim 0.25$ [110, 111], (Tab. 4.1, Figs. 4.1 and 4.2). Both for $r < r_{\min}$ and $r > r_{\max}$ the frequencies *are below* the Zipf curve (Figs. 4.1 and 4.2).

Note that though the validity range $|r_{\max} - r_{\min}|$ is few times smaller than the maximal rank n, it is relevant, since it contains a sizable amount of the overall frequency: for Chinese texts (short or long) the Zipfian range carries 40 % of the overall frequency, i.e. $\sum_{k=r_{\min}}^{r_{\max}} f_k \simeq 0.4$.

2. In the pre-Zipfian range $1 \leq r < r_{\min}$ the overall number of function and empty characters is more than the number of content characters. Function and empty characters serve for establishing grammatical constructions (e.g. "的" *(de)*, "是" *(shì)*,

Table 4.2: Parameters of classic Chinese texts (see Appendix H for further details). Notations have the same meaning as in Tab. 4.1. Here 4 texts means joining of the texts CQF, SBZ, WJZ, and HLJ. Also, 7 (10,14) texts mean joining of the 4 with other 3 (6,11) classic texts, which we do not mention separately, because they give no new information.

Texts	N	n	r_{min}	r_{max}	c	γ	$\sum_{k<r_{min}} f_k$	$\sum_{k=r_{min}}^{r_{max}} f_k$	$\|d\|$	r_b
CQF	30017	1661	47	365	0.1778	0.985	0.43906	0.38997	0.00441	-
SBZ	24634	1959	52	357	0.1819	0.972	0.42828	0.408787	0.004353	-
WJZ	26330	1708	46	360	0.208	0.999	0.40434	0.418733	0.006923	-
HLJ	26559	1837	56	372	0.209	1.01	0.43674	0.379454	0.000832	-
CQF & SBZ	54651	2528	68	483	0.19498	0.989	0.42031	0.401661	0.00483	-
CQF & WJZ	56347	2302	66	439	0.20654	1.002	0.42815	0.383514	0.00564	-
CQF & HLJ	56576	2458	65	416	0.19498	0.998	0.43138	0.38654	0.00913	-
SBZ & WJZ	50964	2505	68	465	0.20512	0.992	0.40116	0.409017	0.00382	-
SBZ & HLJ	51193	2608	72	423	0.20893	1.000	0.41157	0.369598	0.00798	-
WJZ & HLJ	52889	2303	66	432	0.23988	1.035	0.43044	0.380801	0.002321	-
4 texts	107540	3186	75	528	0.22387	1.021	0.42526	0.391818	0.0007	681
7 texts	190803	4069	57	513	0.158	0.97	0.39381	0.4102	0.00331	789
10 texts	278557	4727	67	552	0.168	0.978	0.38058	0.4015	0.00217	1015
14 texts	348793	5018	78	625	0.176	0.98	0.39116	0.418983	0.00954	1223
SJ	572864	4932	76	535	0.236	1.025	0.40153	0.41253	0.007564	1336

"了" *(le)*, "不" *(bù)*, "在" *(zài)*). (We shall list them separately, though for our purposes they can be joined together; the main difference between them is that the empty characters are not used alone.)

But the majority of characters in the Zipfian range do have a specific meaning (content characters). A subset of those content characters has a meaning that is specific for the text and can serve as its key-characters.

Let us take for an example the modern Chinese text KLS (this text concerns military activities). The pre-Zipfian range of this text contains 61 characters. Among them there are, 24 function characters, 9 empty characters [5], 25 content characters, and finally there are 3 key-characters [6]: horn "号" *(hào)*, army "军" *(jūn)* and soldier "兵" *(bīn)*.

[5]We list empty characters separately, though for our purposes they can be joined with function characters.

[6]We present that meaning of the character which is most relevant in the context of the text.

Table 4.3: Parameters of four English texts and their mixtures: *The Age of Reason* (AR) by T. Paine, 1794 (the major source of British deism). *Time Machine* (TM) by H. G. Wells, 1895 (a science fiction classics). *Thoughts on the Funding System and its Effects* (TF) by P. Ravenstone, 1824 (economics). *Dream Lover* (DL) by J. MacIntyre, 1987 (a romance novella). TF & TM means joining the texts TF and TM.
The total number of words N, the number of different words n, the lower $r_{\min}$ and the upper $r_{\max}$ ranks of the Zipfian domain, the fitted values of c and γ, the overall frequencies of the pre-Zipfian and Zipfian range, and the difference d between the total frequency of the Zipfian domain got empirically and its value according to the Zipf's law: $d = \sum_{k=r_{\min}}^{r_{\max}}(ck^{-\gamma} - f_k)$.

Texts	N	n	$r_{\min}$	$r_{\max}$	c	γ	$\sum_{k<r_{min}} f_k$	$\sum_{k=r_{min}}^{r_{max}} f_k$	$\lvert d\rvert$
TF	26624	2067	36	371	0.168	1.032	0.44439	0.35158	0.00333
TM	31567	2612	42	332	0.166	1.041	0.45311	0.33876	0.01004
AR	22641	1706	32	339	0.178	1.038	0.47254	0.33947	0.00048
DL	24990	1748	34	230	0.192	1.039	0.47955	0.33251	0.02145
TF & TM	54191	3408	30	602	0.139	1.013	0.43508	0.40876	0.02091
TF & AR	45265	2656	33	628	0.138	0.998	0.45468	0.41045	0.00239
TF & DL	47614	2877	28	527	0.162	1.014	0.42599	0.42261	0.01490
TM & AR	54208	3184	43	592	0.157	1.021	0.47582	0.39687	0.00491
TM & DL	56557	3154	45	493	0.161	1.023	0.46726	0.38456	0.01211
AR & DL	47631	2550	38	496	0.165	1.012	0.45375	0.39236	0.00947
Four texts	101822	4047	39	927	0.158	1.015	0.44245	0.44158	0.00187

The Zipfian range of the KLS contains 350 characters. Among them, 91 are function, 10 are empty, 230 are content and 19 are key-characters (Tab. 4.4).

3. The absolute majority of different characters with ranks in $[r_{\min}, r_{\max}]$ have different frequencies. Only for $r \simeq r_{\max}$ the number of different characters having the same frequency is $\simeq 10$. For $r > r_{\max}$ we meet the hapax legomena effect: characters occurring only few times in the text (i.e. $f_r N = 1, 2, 3...$ is a small integer), and many characters having the same frequency f_r [111]. The effect is not described by a smooth rank-frequency relation, including the Zipf's law. The theory review below allows to explain the hapax legomena range together with the Zipf's law. Note that the very existence of hapax legomena is a non-trivial effect, since one can easily imagine (artificial) texts, where (say) no character appear only once.

4. All the above results hold for relatively short English texts [149] (Tab. 4.3 and Fig. 4.4). In particular, the Zipfian range of English texts also contains mainly

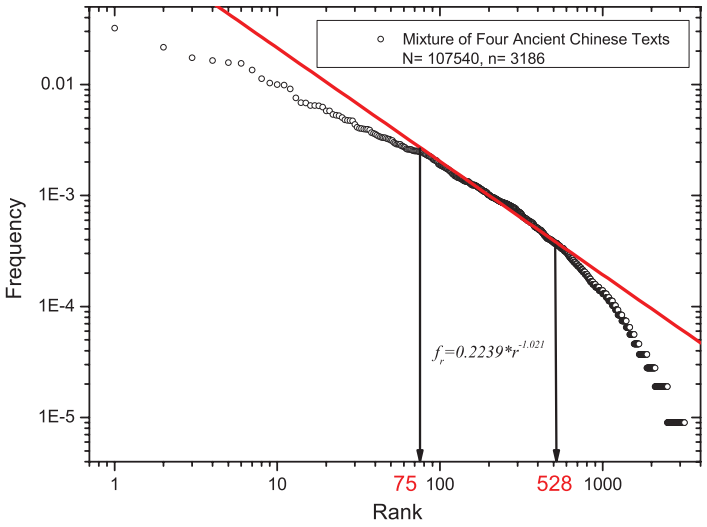

Fig. 4.3: (Color online) Frequency versus rank for the mixture of four short classic Chinese texts: CQF, SBZ, WJZ, HLJ (see also Appendix H). Other notations have the same meaning as in Fig. 4.1.

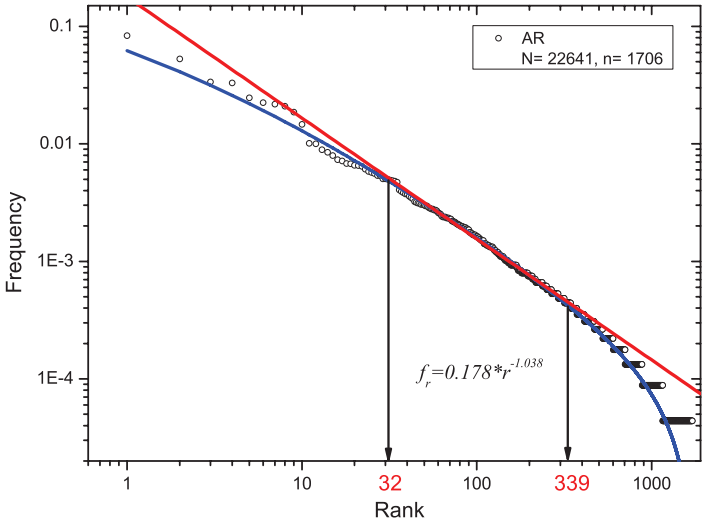

Fig. 4.4: (Color online) Frequency vs. rank for the English text AR (Tab. 4.3). Red line: the Zipf curve $f_r = 0.178r^{-1.038}$.

Table 4.4: List of the Key-Characters in the Pre-Zipfian and Zipfian range of the modern Chinese text, 昆仑觞, *Kūn Lún Shāng* (KLS) written by Shu-Ming BI in 1987. The text is about the arduous military training in the troops of Kun Lun mountain.

No.	Rank	Character	Pinyin	English	Frequency
1	14	号	*hào*	horn	157
2	32	军	*jūn*	army	86
3	44	兵	*bīn*	soldier	67
4	113	队	*duì*	troop	38
5	118	令	*lìng*	command	37
6	123	部	*bù*	troop	36
7	152	战	*zhàn*	fight/war	28
8	156	命	*mìng*	command	28
9	180	防	*fáng*	protect	24
10	213	血	*xuè*	blood	20
11	216	立	*lì*	stand straight	20
12	224	功	*gōng*	honor	19
13	225	枪	*qiāng*	gun	19
14	252	官	*guān*	officer	16
15	295	锅	*guō*	pan	14
16	299	保	*bǎo*	protect	14
17	300	卫	*wèi*	protect	13
18	352	营	*yíng*	camp	11
19	355	谋	*móu*	strategy	11
20	360	烧	*shāo*	burn	11
21	394	烈	*liè*	martyr	10
22	407	团	*tuán*	regiment	10

content words including the keywords. This is known and is routinely used in document processing [152].

We thus conclude that as far as short texts are concerned, the Zipf's law holds for Chinese characters in the same way as it does for English words.

5. To check our results on fitting the empiric data for word frequencies to the Zipf's law we carried out three alternative tests.

5.1 First we applied the Kolmogorov-Smirnov (KS) test to decide on the fitting quality of the data with the Zipf's law (in the range $[r_{\min}, r_{\max}]$). The test was carried out both with and without transforming to the logarithmic coordinates, it fully confirmed our result; see Table 4.5. For a detailed presentation of the KS test results see

Table 4.5: Comparison between different methods of estimating the exponent γ of the Zipf's law: LLS (linear least-square), NLS (nonlinear least-square), MLE (maximum likelihood estimation). We also present the p-value of the KS test when comparing the empiric word frequencies in the range $[r_{\min},\ r_{\max}]$ with the Zipf's-law within the linear lest-square method (LLS); for a more detailed presentation of the KS results see Appendix G. Recall that the p-values have to be sufficiently larger than 0.1 for fitting to be reliable from the viewpoint of KS test. This holds for the presented data; see Appendix G for details.

Texts	γ, LLS	γ, NLS	γ, MLE	p-value
TF	1.032	1.033	1.035	0.865
TM	1.041	1.036	1.039	0.682
AR	1.038	1.042	1.044	0.624
DL	1.039	1.034	1.035	0.812
AQZ	1.03	1.028	1.027	0.587
KLS	0.97	0.975	0.973	0.578
CQF	0.985	0.983	0.981	0.962
SBZ	0.972	0.967	0.973	0.796
WJZ	0.999	0.993	0.995	0.852
HLJ	1.01	1.015	1.011	0.923

Appendix G.

5.2 It was recently shown that even when the applicability range $[r_{\min}, r_{\max}]$ of a power law is known, the linear least-square method (that we employed above) may not give accurate estimations for the exponent γ of the power law. It was then argued that the method of Maximum Likelihood Estimation (MLE) is more reliable in this context. Hence to show that our results are robust, we calculated γ using the MLE method. We got that the difference with the linear least square method is quite small (changes come only at the third decimal place); see Table 4.5.

5.3 We also checked whether our results on the power law exponent γ are stable with respect to non-linear fitting schemes. Again, we find that non-linear fitting schemes (that we carried out via routines of Mathematica 7) produce very similar results for γ; see Table 4.5.

One reason for such a good coincidence between our linear fitting results and alternative tests is that we use a rather strict criteria ($SS^*_{\rm err} < 0.05$ and $R^2 > 0.995$) for determining *first* the Zipfian range $[r_{\min},\ r_{\max}]$ and then the parameters of the Zipf's law. Another reason is that in the vicinity of $r_{\max}$, the number of different words having

the same frequency is not large (it is smaller than 10). Hence there are no problems with lack of data points or systematic biases that can plague the applicability of the least square method for determination of the exponent γ.

4.2.2 Theoretical description of Zipf's law

A theoretical description of the Zipf's law that is specifically applicable to short English texts was recently proposed in [149]. We shall briefly remind it to demonstrate that also describes the rank-frequency relation for short Chinese texts. The theory is based on the ideas of latent semantic analysis and the concept of mental lexicon [149]. Its final outcome is the (binomial) probability $p_r(\nu|T)$ of the character (or word) with the rank r to appear ν times in a text T (with N total characters and n different characters):

$$p_r(\nu|T) = \frac{N!}{\nu!(N-\nu)!}\phi_r^{\nu}(1-\phi_r)^{N-\nu}, \tag{4.2}$$

where the effective probability ϕ_r of the character is found from two equations for two unknowns μ and ϕ_r:

$$r/n = \int_{\phi_r}^{\infty} \mathrm{d}\theta\, \frac{e^{-\mu\theta}}{(c+\theta)^2} \Big/ \int_0^{\infty} \mathrm{d}\theta\, \frac{e^{-\mu\theta}}{(c+\theta)^2}\,, \tag{4.3}$$

$$\int_0^{\infty} \mathrm{d}\theta\, \frac{\theta\, e^{-\mu\theta}}{(c+\theta)^2} = \int_0^{\infty} \mathrm{d}\theta\, \frac{e^{-\mu\theta}}{(c+\theta)^2}, \tag{4.4}$$

where c is a constant that will later on shown to coincide with the prefactor of the Zipf's law.

For $c \lesssim 0.25$, $c\mu$ determined from (4.4) is small and is found from integration by parts:

$$\mu \simeq c^{-1}\, e^{-\gamma_{\mathrm{E}} - \frac{1+c}{c}}, \tag{4.5}$$

where $\gamma_{\mathrm{E}} = 0.55117$ is the Euler's constant. One solves (4.3) for $c\mu \to 0$:

$$\frac{r}{n} = ce^{-n\phi_r\mu}/(c+n\phi_r). \tag{4.6}$$

Recall that according to (4.2), ϕ_r is the probability for the character (or the word in the English situation) with rank r. If ϕ_r is sufficiently large, $\phi_r N \gg 1$, the character with rank r appears in the text many times and its frequency $\nu \equiv f_r N$ is close to its maximally probable value $\phi_r N$; see (4.2). Hence the frequency f_r can be obtained

via the probability ϕ_r. This is the case in the Zipfian domain, since according to our empirical results (both for Chinese and English) $\frac{1}{n} \lesssim f_r$ for $r \leq r_{\max}$, and—upon identifying $\phi_r = f_r$—the above condition $\phi_r N \gg 1$ is ensured by $N/n \gg 1$.

Let us return to (4.6). For $r > r_{\min}$, $\phi_r n\mu = f_r n\mu < 0.04 \ll 1$; see (4.5), Figs. 4.1, 4.2 and 4.4. We get from (4.6):

$$f_r = c(r^{-1} - n^{-1}). \tag{4.7}$$

This is the Zipf's law generalized by the factor n^{-1} at high ranks r. This cut-off factor ensures faster [than r^{-1}] decay of f_r for large r.

Figs. 4.1, 4.2 and 4.4 show that (4.7) reproduces well the empirical behavior of f_r for $r > r_{\min}$. Our derivation shows that c is the prefactor of the Zipf's law, and that our assumption on $c \lesssim 0.25$ above (4.5) agrees with observations (Tabs. 4.1, 4.2 and 4.3).

For given prefactor c and the number of different characters n, (4.3) predict the Zipfian range $[r_{\min}, r_{\max}]$ in agreement with empirical results (Figs. 4.1, 4.2 and 4.4).

For $r < r_{\min}$, it is not anymore true that $f_r n\mu \ll 1$ (though it is still true that $f_r N = \phi_r N \gg 1$). So the fuller expression (4.3) is to be used instead of (4.6). It reproduces qualitatively the empiric behavior of f_r also for $r < r_{\min}$ (Figs. 4.1, 4.2 and 4.4). We do not expect any better agreement theory and observations for $r < r_{\min}$, since the behavior of frequencies in this range is irregular and changes sizably from one text to another.

4.2.3 Hapax legomena

According to (4.2), the probability ϕ_r is small for $r \gg r_{\max}$ and hence the occurrence number $\nu \equiv f_r N$ of the character with the rank r is a small integer (e.g. 1 or 2) that cannot be approximated by a continuous function of r (Figs. 4.1, 4.2 and 4.4). In particular, the reasoning after (4.6) on the equality between frequency and probability does not apply, although we see in Figs. 4.1, 4.2 and 4.4 that (4.7) roughly reproduces the trend of f_r even for $r > r_{\max}$.

To describe this hapax legomena range, define r_k as the rank, when $\nu \equiv f_r N$ jumps from integer k to $k+1$. Since ϕ_r reproduces well the trend of f_r even for $r > r_{\max}$, r_k

Table 4.6: Frequency of Chinese characters in the hapax legomena domain; $\hat{r}_k$ is calculated from (4.8), while r_k is found from empirical data.

Texts	k	1	2	3	4	5	6	7	8	9	10
AQZ	r_k	1097	857	702	595	522	461	414	370	339	311
	$\hat{r}_k$	1116	869	711	601	520	458	409	369	336	308
	$\frac{\lvert\hat{r}_k - r_k\rvert}{r_k}$	0.017	0.014	0.013	0.010	0.0038	0.0065	0.012	0.0027	0.0088	0.0096
KLS	r_k	1405	1060	885	767	662	582	520	455	408	377
	$\hat{r}_k$	1428	1093	884	750	656	575	515	445	404	369
	$\frac{\lvert\hat{r}_k - r_k\rvert}{r_k}$	0.016	0.031	0.0011	0.022	0.0091	0.012	0.0096	0.022	0.0098	0.021
SBZ	r_k	1460	1141	959	850	735	676	618	563	517	481
	$\hat{r}_k$	1481	1168	980	848	740	656	599	553	497	488
	$\frac{\lvert\hat{r}_k - r_k\rvert}{r_k}$	0.014	0.024	0.022	0.0024	0.0068	0.029	0.031	0.018	0.039	0.015
HLJ	r_k	1302	1045	872	756	669	604	551	501	467	430
	$\hat{r}_k$	1327	1080	900	783	684	607	545	494	462	420
	$\frac{\lvert\hat{r}_k - r_k\rvert}{r_k}$	0.019	0.033	0.032	0.035	0.022	0.0049	0.011	0.014	0.011	0.023

can be theoretically predicted from (4.7) by equating its left-hand-side to k/N:

$$\hat{r}_k = [\frac{k}{Nc} + \frac{1}{n}]^{-1}, \qquad k = 0, 1, 2, ... \tag{4.8}$$

Eq. (4.8) is exact for $k = 0$, and agrees with r_k for $k \geq 1$, see Tab. 4.6. Hence it describes the hapax legomena phenomenon, where many characters have the same small frequency.

We thus saw that a single formalism adequately describes both the Zipf's law for short texts and the hapax legomena range.

4.2.4 Summary

It is to be concluded from this section that—as far as the applicability of the Zipf's law to short texts is concerned—the Chinese characters behave similarly to English words. In particular, both situations can be adequately described by the same theory.

We should like to stress again why the consideration of short texts is important. One can argue that—at least for the sake of rank-frequency relations—long texts are just mixtures (joinings) of shorter, thematically homogeneous pieces (this premise is fully confirmed below). Hence the task of studying rank-frequency relations separates into two parts: first understanding short texts, and then long ones.

4.3 Rank-frequency relation for long texts and mixtures of short texts

4.3.1 Mixing English texts

When mixing (joining)[7] different English texts the valid range of the Zipf's law increases due to acquiring more higher rank words, i.e. $r_{\min}$ stays approximately fixed, while $r_{\max}$ increases (Tab. 4.3). The overall precision of the Zipf's law also increases upon mixing, as Tab. 4.3 shows.

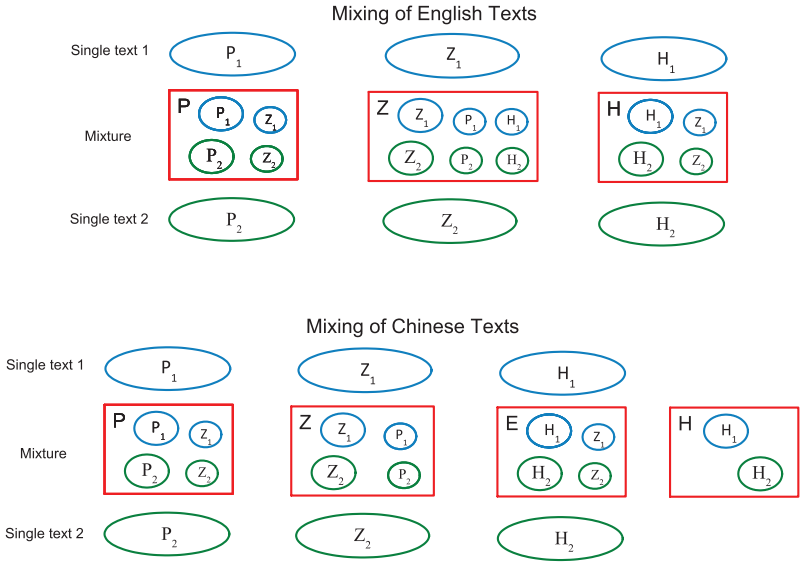

Fig. 4.5: (Color online) Schematic representation of various ranges under mixing (joining) two English (upper figure) and two Chinese (lower figure) texts. P_k, Z_k and H_k mean, respectively, the pre-Zipfian, Zipfian and hapax legomena ranges of the text k ($k = 1, 2$). P, Z and H mean the corresponding ranges for the mixture of texts 1 and 2. E means the exponential range that emerges upon mixing of two Chinese texts. For each range of the mixture we show schematically contributions from various ranges of the separate texts. The relative importance of each contribution is conventionally represented by different magnitudes of the circles.

The rough picture of the evolution of the rank-frequency relation under mixing two texts is summarized as follows, see Tab. 4.3 and Fig. 4.5 for a schematic illustration. The majority of the words in the Zipfian range of the mixture (e.g. AR & TM) come

[7]Upon joining two texts (A and B), the word frequencies get mixed: $f_k(\mathrm{A\&B}) = \frac{N_\mathrm{A}}{N_\mathrm{A}+N_\mathrm{B}} f_k(\mathrm{A}) + \frac{N_\mathrm{B}}{N_\mathrm{A}+N_\mathrm{B}} f_k(\mathrm{B})$, where N_A and $f_k(\mathrm{A})$ are, respectively, the total number of words and the frequency of word k in the text A.

from the Zipfian ranges of the separate texts. In particular, all the words that appear in the Zipfian ranges of the separate words do appear as well in the Zipfian range of the mixture (e.g. the Zipfian ranges of AR and TM have 130 common words). There are also relatively smaller contributions to the Zipfian range of the mixture from the pre-Zipfian and hapax legomena range of separate texts: note from Tab. 4.3 that the Zipfian range of the mixture AR & TM is 82 words larger than the sum of two separate Zipfian ranges, which is (307 + 290) minus 130 common words.

Some of the words that appear only in the Zipfian range of one of separate texts will appear in the hapax legomena range of the mixture; other words move from the pre-Zipfian range of separate texts to the Zipfian range of the mixture. But these are relatively minor effects: the rough effect of mixing is visualized by saying that the Zipfian ranges of both texts combine to become a larger Zipfian range of the mixture and acquire additional words from other ranges of the separate texts (Fig. 4.5). Note that the keywords of separate words stay in the Zipfian range of the mixture, e.g. after joining all four above texts, the keywords of each text are still in the Zipfian range, which now contains almost 900 words (Tab. 4.3).

The results on the behavior of the Zipf's law under mixing are new, but their overall message—the validity of the Zipf's law improves upon mixing) is expected—since it is known that the Zipf's law holds not only for short but also for long English texts and for frequency dictionaries (huge mixtures of various texts) [110].

4.3.2 Mixing Chinese texts

4.3.2.1 Stability of the Zipfian range

The situation for Chinese texts is different. Upon mixing two Chinese texts the validity range of the Zipf's law increases, but much slower as compared to English texts, see Tabs. 4.1 and 4.2. The valid ranges of the separate texts do not combine (in the above sense of English texts). Though the common words in the Zipfian ranges of separate texts do appear in the Zipfian range of the mixture, a sizable amount of those words that appeared in the Zipfian range of only one text do not show up in the Zipfian range of the mixture [8].

[8] As an example, let us consider in detail the mixing of two Chinese texts SBZ and CQF, see Tab. 4.2. The Zipfian ranges of CQF and SBZ contain, respectively, 306 and 319 characters. Among them

Importantly, the overall frequency of the Zipfian domain for very different Chinese texts (mixtures, long texts) is approximately the same and amounts to $\simeq 0.4$ (Tabs. 4.1 and 4.2). In contrast, for English texts this overall frequency grows with the number of different words in the text (Tab. 4.3). This is certainly consistent with the fact that for English texts the Zipfian range increases upon mixing.

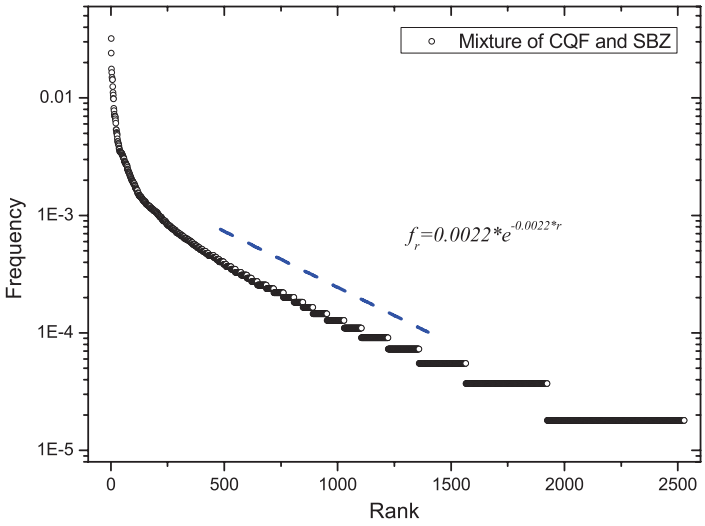

Fig. 4.6: (Color online) Rank frequency distribution for the mixture of CQF and SBZ. (Tab. 4.2) The scale of the frequency is chosen such that the exponential regime of the rank-frequency relation for $r > 500$ is made visible. For comparison, the dashed blue line shows a curve $f_r = 0.0022e^{-0.0022r}$. For the present example, the exponential regime is essentially mixed with hapax legomena, since for frequencies f_r with $r > r_{\max}$ the number of different words having this frequency is larger than 10. Recall that the Zipf's law holds for $r \in [r_{\min}, r_{\max}]$.

4.3.2.2 Emergence of the exponential regime

The majority of characters that appear in the Zipfian range of separate texts, but do not appear in the Zipfian range of the mixture, moves to the hapax legomena range of the mixture. Then, for larger mixtures and longer texts, a new, exponential regime of the rank-frequency relation emerges from within the hapax legomena range.

133 characters are common. The balance of the characters upon mixing is calculated as follows: 306 (from the Zipfian range of CQF) + 319 (from the Zipfian range of SBZ) - 133 (common characters) - 50 (characters from the Zipfian range of CQF that do not appear in the Zipfian range of CQF & SBZ) - 54 (characters from the Zipfian range of SBZ that do not appear in the Zipfian range of CQF+SBZ) +27 (characters that enter to the Zipfian range CQF & SBZ from the pre-Zipfian ranges of CQF or SBZ)= 415 (characters in the Zipfian range of CQF & SBZ).

To illustrate the emergence of the exponential regime, let us start with Fig. 4.6, here there are only two short texts mixed and hence the exponential regime cannot be reliably distinguished from the hapax legomena regime [9]: for all frequencies with the ranks $r > r_{\max}$ (i.e. for all frequencies beyond the Zipfian regime), the number of different characters having exactly the same frequency is larger than 10. (We conventionally take this number as a borderline of the hapax legomena.) However, the trace of the exponential regime is seen even within the hapax legomena, see Fig. 4.6.

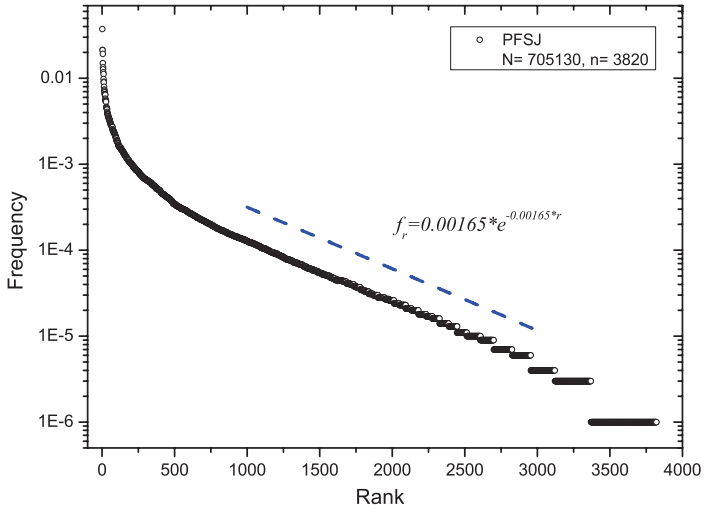

Fig. 4.7: (Color online) Rank frequency distribution of the long modern Chinese text PFSJ. The exponential behavior $f_r \propto e^{-0.00165r}$ of frequency f_r is visible for $r > 500$. For comparison, the dashed blue line shows a curve $f_r = 0.00165e^{-0.00165r}$. The boundary between the exponential regime and hapax legomena can be defined as the rank r_b, where the number of words having the same frequency f_{r_b} is equal to 10. For the present example $r_b = 1437$.

For bigger mixtures or longer texts, the exponential regime clearly differentiates from the hapax legomena. In this context, we define r_b as the borderline rank of the hapax legomena: for $r > r_b$, the number of characters having the frequency f_{r_b} is larger than 10. Then the exponential regime

$$f_r = ae^{-br} \text{ with } a < b, \tag{4.9}$$

exists for the ranks $r_{\max} < r \lesssim r_b$ (provided that $r_{\max}$ is sufficiently larger than r_b). Put differently, the exponential regime exists from ranks sufficiently larger than the

[9]Recall in this context that in the hapax legomena range many characters have the same frequency, hence no smooth rank-frequency relation is reliable.

Table 4.7: Parameters of the exponential range (lower and upper ranks and the overall frequency) for few long Chinese texts (Tabs. 4.1 and 4.2). Here n is the number of different characters. Recall that the lowest rank of the exponential range is $r_{\max}+1$, where $r_{\max}$ is the upper rank of the Zipfian range. The highest rank of the exponential range was denoted as r_b. (Tabs. 4.1 and 4.2)

Texts	n	Rank range	Overall frequency
PFSJ	3820	584–1437	0.12816
SHZ	4376	591–1618	0.14317
SJ	4932	536–1336	0.12887
14 texts	5018	626–1223	0.12291

upper rank $r_{\max}$ of the Zipfian range till the ranks, where the hapax legomenon starts. Tabs. 4.1, 4.2 and Fig. 4.7 show that the exponential regime is not only sizable by itself, but (for sufficiently long texts or sufficiently big mixtures) it is also bigger than the Zipfian range. This, of course, does not mean that the Zipfian range becomes less important, since, as we saw above, it carries out nearly 40 % of the overall frequency (Tabs. 4.1 and 4.2). The exponential range also carries out non-negligible frequency, though it is few times smaller than that of the Zipfian and pre-Zipfian ranges (Tabs. 4.1, 4.2 and 4.7).

Finally, we would like to stress that we considered various Chinese texts written with simplified or traditional characters, with Modern Chinese or different versions of Classic Chinese (Tabs. 4.1, 4.2 and Appendix H). As far as the rank-frequency relations are concerned, all these texts demonstrate the same features showing that the peculiarities of these relations are based on certain very basic features of Chinese characters. They do not depend (or depend much less) on specific details of texts.

4.4 Conclusions and discussions

4.4.1 Summary of results

1. As implied by the rank-frequency relation for characters, short Chinese texts demonstrate the same Zipf's law—together with its generalization to high and low frequencies (hapax legomena)—as short English texts. Assuming that authors write mainly relatively short texts (longer texts are obtained by mixing shorter ones), this similarity

implies that Chinese characters play the same role as English words.

2. As compared to English, there are two novelties of the rank-frequency relation of Chinese characters in long texts.

2.1 The overall frequency of the Zipfian range (the range of middle ranks, where the Zipf's law holds) stabilizes at $\simeq 0.4$. This holds for all texts we studied (written in different epochs, genres with different types of characters, see Tabs. 4.1, 4.2 and Appendix H). This effect of stabilization holds as well for the overall frequency of the pre-Zipfian range for both English and Chinese texts (Tabs. 4.1, 4.2 and 4.3).

2.2 There is a range with an exponential rank-frequency relation. It emerges for relatively longer texts from within the hapax legomena range. The range of ranks, where the exponential regime holds, is larger than that of the Zipf's law. But its overall frequency is few times smaller (Tabs. 4.1, 4.2 and 4.7).

Both these results are absent for English texts; there the overall frequency of the Zipfian range grows with the length of the text, while there is no exponential regime: the Zipfian range end with the hapax legomena (Tab. 4.3).

The results **2.1** and **2.2** imply that long Chinese texts do have a hierarchic structure: there is a group of characters that hold the Zipf's law with nearly universal overall frequency equal to $\simeq 0.4$, and yet another group of relatively less frequent characters that display the exponential range of the rank-frequency relation.

4.4.2 Interpretations and discussions

Chinese characters differ from English words, since only long Chinese texts have the above hierarchic structure. The underlying reason of the hierarchic structure is to be sought via the linguistic differences between Chinese characters and English words, as we outlined in Appendix D. In particular, the features **4, 6, 7** discussed in Appendix D can mean that certain homographic content characters play multiple role in different parts of a long Chinese text. They are hence distinguished and appear in the Zipfian range of the long text with (approximately) stable overall frequency $\simeq 0.4$. Since this frequency is sizable, and since the range of ranks carried out by the Zipf's law is relatively small, there is a relatively large range of ranks that has to have a relatively small overall frequency (Tabs. 4.1, 4.2 and 4.7). It is then natural that in this range there emerges an exponential regime that is related with a faster (compared to a power

law) decay of frequency versus rank.

Recall that the stabilization holds as well for the overall frequency of the pre-Zipfian domain both for English and Chinese texts. The explanation of this effect is similar to that given above (but to some extent is also more transparent): the pre-Zipfian range contains mostly function characters, which are not specific and used in different texts. Hence upon mixing the pre-Zipfian range has a stable overall frequency.

The above explanation for the coexistence of the Zipfian and exponential range suggests that there is a relation between the characters that appear in the Zipfian range of long texts and homography. As a preliminary support for this hypothesis, we considered the following construction. Assuming that a mixture is formed from separate texts $T_1, ..., T_k$, we looked at characters that appear in the Zipfian ranges of all the separate texts $T_1, ..., T_k$. This guarantees that these characters appear in the Zipfian range of the mixture. Then we estimated (via an explanatory dictionary of Chinese characters) the average number of different meanings for these characters. This average number appeared to be around 8, which is larger than the average number of meanings for an arbitrary Chinese character (i.e. when the averaging is taken over all characters in the dictionary) that is known to be not larger than 2 [175].

We should like to stress however that the above connection between the uncovered hierarchic structure and the number of meanings is preliminary, since we currently lack a reliable scheme of relating the rank-frequency relation of a given text to its semantic features.

The above discussion makes clear that a theory for studying the rank-frequency relation of a long text, as it emerges from mixing of different short texts, is currently lacking. Such a theory was not urgently needed for English texts, because there the (generalized) Zipf's law (4.7) describes well both long and short texts. But the example of Chinese characters clearly shows that the changes of the rank-frequency relation under mixing are essential. Hence the theory of the effect is needed.

Finally, one of main open questions is whether the uncovered hierarchical structure is really specific for Chinese characters, or it will show up as well for English texts, but on the level of the rank-frequency relation for morphemes and not the words. Factorizing English words into proper morphemes is not straightforward, but still possible.

Chapter 5

Conclusions and outlook

5.1 Conclusions

In this research, I am inspired by its interdisciplinary range of applications and its relevance to the fundamental issues, such as the human dynamics in sports systems, the schemes of human languages and scripts. Specifically, the main results and potential contributions are concluded as follows.

We found the distributions of scores/prize money for 40 data samples in 12 different sports fields, are governed by the same universal powers laws, which are also similar to the distributions of city size or human wealth. Moreover, the 40 data samples from all 12 sports fields seem to follow the Pareto principle, that is, 20% of the players or teams accumulate 80% of the scores/prize money. We also proposed an agent-based model which could simulate the competitions of players, when two players compete, we apply the empiric findings of win probability in tennis that, the probability the higher-ranked player will win is related to their rank difference. We expect these findings are relevant to reflect the features of human dynamics in competition-driven systems.

For the Zipf's law in human language texts, we showed that the Zipf's law could be analytically derived by the assumption that words are drawn into the text with random probabilities, while their apriori probability density relates to the stable and efficient organization of the mental lexicon of the text-producing author. Our approach could be applied to clarify the limits of its validity, and also its generalization to high and low frequencies including hapax legomena. We expect that our results will improve the performance of Probabilistic Latent Semantic Analysis (PLSA) algorithms making

them consistent with the Zipf's law. Also, the proposed methods can find applications for studying rank-frequency relations and power laws in other fields.

Concerning the rank-frequency relation for Chinese characters, we found that, for short Modern or Classic Chinese texts, they demonstrate the same Zipf's law–together with its generalization to high and low frequencies (hapax legomena)–as short English texts. While for long Modern or Classic Chinese texts, they appear a hierarchic structure: there is a group of characters that hold the Zipf's law with nearly universal overall frequency equal to $\simeq 0.4$, and another group of relatively less frequent characters that display the exponential range of the rank-frequency relation. We hope this research will contribute to document classification algorithms for Chinese characters, and may also provide a general method for distinguishing between logographic and phonetic scripts.

5.2 Future research plan

Following the above research works in Sports Systems and Human Languages, several aspects deserve the further investigations, so as to gain a deeper insight into the nature of these two fundamental issues.

Theoretical Explanation of the Universal Power Laws in Sports

Statistical analysis of score and/or prize money distributions of players, across 40 data samples in 12 sports fields: tennis, golf, table tennis, volleyball, football, snooker, badminton, basketball, baseball, hockey, handball and fencing, share similar universal power laws. The reasons why the sports systems have such common distributions are unknown, and they are the latest examples of the phenomena that abide by the mysterious power laws.

In the current work, we just proposed a sample toy model to simulate the real competition process of sports, simulations could yield results consistent with the empirical findings. However, it lacks the theoretical background, thus whether we can apply the theories in statistical physics, e.g. Markov models, or Self-organized Criticality, etc to study the common features in sports systems, this deserves the further investigation.

Structure of Tournaments and Rating Systems for Sports

Ranking is a direct measure of a player or a team's performance and come in

different forms. Some sports are ranked by using a points system, while others use the earnings. In practice, there can be a lot of factors which influence the rankings of players or teams, such as the structure of specific tournaments, the rating strategies of the tournaments, etc. Therefore, we shall try to provide some theoretical basis, so as to optimize the structure of the tournaments and make the rating strategies more efficient and fair.

Theoretical framework for Rank-Frequency Relations of Chinese Characters

The Zipf's law for short Chinese texts behaves much similar to those for short English texts, but for long texts, the rank-frequency relations of Chinese characters are quite different from those of the English words, e.g., for Chinese characters, there emerges a wide range of ranks where the rank-frequency relations is approximately exponential. So what does this imply, and what are the reasons behind these differences? Are they due to the different features of mental lexicons of Chinese and English writers, or the different mechanisms during the production of long texts out of smaller, thematically homogeneous pieces?

Rank-Frequency Relations for Phonetics of Human Languages

Phonetics is a branch of linguistics that comprises the study of the sounds of human speech, or in the case of the sign languages. The speech behavior and the voice of human beings could reflect the regional, social and personal identity. We shall aim to study the rank-frequency relations for the phonetics of human languages, to uncover some basic properties of the sounds of human speech, and provide some hints for the explanations.

Appendix A - Linear fitting method

This is the detailed explanations of the linear fitting method to clarify the valid range of the Zipf's law:

For each text we extract the ordered and normalized frequencies of different words [the number of different words is n; the overall number of words in a text is N]:

$$\{f_r\}_{r=1}^{n},\ \ f_1 \geq ... \geq f_n,\ \ \sum\nolimits_{r=1}^{n} f_r = 1. \tag{A.1}$$

We should now see whether the data $\{f_r\}_{r=1}^{n}$ fits to a power law: $\hat{f}_r = cr^{-\gamma}$. We represent the data as

$$\{y_r(x_r)\}_{r=1}^{n},\ \ y_r = \ln f_r,\ \ x_r = \ln r, \tag{A.2}$$

and fit it to the linear form $\{\hat{y}_r = \ln c - \gamma x_r\}_{r=1}^{n}$. Two unknowns $\ln c$ and γ are obtained from minimizing the sum of squared errors:

$$SS_{\rm err} = \sum\nolimits_{r=1}^{n} (y_r - \hat{y}_r)^2. \tag{A.3}$$

It is known since Gauss that this minimization produces

$$-\gamma^* = \frac{\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y})}{\sum_{k=1}^{n}(x_k - \overline{x})^2},\ \ \ln c^* = \overline{y} + \gamma^* \overline{x}, \tag{A.4}$$

where we defined

$$\overline{y} \equiv \frac{1}{n}\sum\nolimits_{k=1}^{n} y_k,\ \ \overline{x} \equiv \frac{1}{n}\sum\nolimits_{k=1}^{n} x_k. \tag{A.5}$$

As a measure of fitting quality one can take:

$$\min\nolimits_{c,\gamma}[SS_{\rm err}(c,\gamma)] = SS_{\rm err}(c^*,\gamma^*) \equiv SS^*_{\rm err}. \tag{A.6}$$

This is however not the only relevant quality measure. Another (more global) aspect of this quality is the coefficient of correlation between $\{y_r\}_{r=1}^{n}$ and $\{\hat{y}_r\}_{r=1}^{n}$ [136]:

$$R^2 = \frac{\left[\sum_{k=1}^{n}(y_k - \bar{y})(\hat{y}^*_k - \overline{\hat{y}^*})\right]^2}{\sum_{k=1}^{n}(y_k - \bar{y})^2 \sum_{k=1}^{n}(\hat{y}^*_k - \overline{\hat{y}^*})^2}, \tag{A.7}$$

where

$$\hat{y}^* = \{\hat{y}^*_r = \ln c^* - \gamma^* x_r\}_{r=1}^{n},\ \ \ \overline{\hat{y}^*} \equiv \frac{1}{n}\sum\nolimits_{k=1}^{n} \hat{y}^*_k. \tag{A.8}$$

For the linear fitting (A.4), the squared correlation coefficient is equal to the coefficient of determination,

$$R^2 = \sum_{k=1}^{n} (\hat{y}_k^* - \overline{y})^2 \Big/ \sum_{k=1}^{n} (y_k - \overline{y})^2, \tag{A.9}$$

the amount of variation in the data explained by the fitting [136]. Hence $SS_{\text{err}}^* \to 0$ and $R^2 \to 1$ mean good fitting. We minimize SS_{err} over c and γ for $r_{\min} \leq r \leq r_{\max}$ and find the maximal value of $r_{\max} - r_{\min}$ for which SS_{err}^* and $1 - R^2$ are smaller than, respectively, 0.05 and 0.005. This value of $r_{\max} - r_{\min}$ also determines the final fitted values c^* and γ^* of c and γ, respectively. Thus c^* and γ^* are found simultaneously with the validity range $[r_{\max}, r_{\max}]$ of the law. Whenever there is no risk of confusion, we for simplicity refer to c^* and γ^* as c and γ, respectively.

Appendix B - Derivation of Eqs. (3.10-3.12)

We defined $P_r(t|T)$ as the marginal density for the probability t of the word w_r,

$$P_r(t|T) \propto \int_0^\infty \mathrm{d}\theta_1 \int_0^{\theta_1} \mathrm{d}\theta_2 \int_0^{\theta_2} \mathrm{d}\theta_3 ... \int_0^{\theta_{n-1}} \mathrm{d}\theta_n \times P(\theta_1, ..., \theta_n)\, \delta(t - \theta_r), \quad \text{(A.10)}$$

where

$$P(\theta_1, ..., \theta_n) \propto u(\theta_1) \, ... \, u(\theta_n)\, \delta(\sum\nolimits_{k=1}^{n} \theta_k - 1). \quad \text{(A.11)}$$

In (A.11) we employ the Fourier representation of the delta-function,

$$\delta(\sum\nolimits_{k=1}^{n} \theta_k - 1) = \int_{-i\infty}^{i\infty} \frac{\mathrm{d}z}{2\pi i}\, e^{z - z\sum_{k=1}^{n}\theta_k}, \quad \text{(A.12)}$$

put (A.11) into (A.10) and then apply integration by parts. The result reads

$$P_r(t|T) \propto u(t) \int_{-i\infty}^{i\infty} \frac{\mathrm{d}z\, e^z}{2\pi i} \chi_0^{n-r}(t,z) \chi_1^{r-1}(t,z)\, e^{-tz}, \quad \text{(A.13)}$$

where

$$\chi_0(t,z) \equiv \int_0^t \mathrm{d}y e^{-zy} u(y),\ \chi_1(t,z) \equiv \int_t^\infty \mathrm{d}y e^{-zy} u(y).$$

The integral in (A.13) will be worked out via the saddle point method. But before that we need to fix the scales of the involved quantities. To this end, make the following changes of variables

$$\tilde{z} = z/n,\ \ \tilde{t} = tn,\ \ \tilde{y} = yn,\ \ \tilde{r} = r/n. \quad \text{(A.14)}$$

Then $P_r(t|T)$ reads from (A.13)

$$P_r(t|T) \ \propto \ u(t) \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\tilde{z}}{2\pi i}\, e^{n\varphi(\tilde{t},\tilde{z}) - \tilde{t}\tilde{z}}, \quad \text{(A.15)}$$

$$\begin{aligned} \varphi(\tilde{t},\tilde{z}) \ &= \ \tilde{z} + (1-\tilde{r}) \ln \int_0^{\tilde{t}} \frac{\mathrm{d}y\, e^{-\tilde{z}y}}{(c+y)^2} \\ &+ \ (\tilde{r} - \frac{1}{n}) \ln \int_{\tilde{t}}^\infty \frac{\mathrm{d}y\, e^{-\tilde{z}y}}{(c+y)^2}, \end{aligned} \quad \text{(A.16)}$$

where in (A.16) we already used $u(t) = (n^{-1}c + t)^{-2}$.

If $n \gg 1$ and $0 < \tilde{r} < 1$ is a finite number (neither close to one, nor to zero), the behavior of $\rho_r(t)$ in various averages, e.g. $\int \mathrm{d}t\, t\, \rho_r(t)$, is determined by the values of $\tilde{z} = \tilde{z}_s$ and $\tilde{t} = \tilde{t}_s$ that maximize $\phi(\tilde{t},\tilde{z})$. They are found from saddle-point equations

$$\partial_{\tilde{t}} \phi(\tilde{t}_s, \tilde{z}_s) = \partial_{\tilde{z}} \phi(\tilde{t}_s, \tilde{z}_s) = 0. \quad \text{(A.17)}$$

After reworking the two equations (A.17) we get Eqs. (3.11, 3.12) of the Chapter 3.

Due to (A.14), $\tilde{z}_s$ (that is real and positive) and $\tilde{t}_s$ stay finite for $n \gg 1$. Hence the integration line over $\tilde{z}$ in (A.15) is shifted to pass through $\tilde{z}_s$ (the saddle-point method). Now $\phi(\tilde{t}, \tilde{z})$ is expanded around $\tilde{z} = \tilde{z}_s$ and $\tilde{t} = \tilde{t}_s$ [first-order terms nullify due to (A.17)]:

$$\phi(\tilde{t}, \tilde{z}) = \phi(\tilde{t}_s, \tilde{z}_s) + \frac{1}{2}\partial_{\tilde{t}\tilde{t}}\phi(\tilde{t}_s, \tilde{z}_s)(\tilde{t} - \tilde{t}_s)^2 \tag{A.18}$$

$$+\frac{1}{2}\partial_{\tilde{z}\tilde{z}}\phi(\tilde{t}_s, \tilde{z}_s)(\tilde{z} - \tilde{z}_s)^2 \tag{A.19}$$

$$+\partial_{\tilde{t}\tilde{z}}\phi(\tilde{t}_s, \tilde{z}_s)(\tilde{t} - \tilde{t}_s)(\tilde{z} - \tilde{z}_s) + ... \tag{A.20}$$

Now only these terms can be retained in the integral over $\tilde{z}$. Since this integral goes over the imaginary axis, while $\tilde{z}_s$ is real, the integration contour is to be shifted to pass through $\tilde{z}_s$. For the convergence of the resulting Gaussian integral we need $\frac{1}{2}\partial_{\tilde{z}\tilde{z}}\phi(\tilde{t}_s, \tilde{z}_s) > 0$. Taking this Gaussian integral leads us to [up to factors that either constant or irrelevant for $n \gg 1$]

$$P_r(t|T) \propto e^{-\frac{n}{2\sigma^2}(\tilde{t}-\tilde{t}_s)^2} = e^{-\frac{n^3}{2\sigma^2}(t-\frac{\tilde{t}_s}{n})^2}, \tag{A.21}$$

$$\frac{1}{\sigma^2} = \frac{[\,\partial_{\tilde{t}\tilde{z}}\phi(\tilde{t}_s, \tilde{z}_s)\,]^2}{\partial_{\tilde{z}\tilde{z}}\phi(\tilde{t}_s, \tilde{z}_s)} - \partial_{\tilde{t}\tilde{t}}\phi(\tilde{t}_s, \tilde{z}_s). \tag{A.22}$$

Hence $P_r(t|T)$ is approximately Gaussian, with the standard deviation $\mathcal{O}(n^{-3/2})$ much smaller than the average for $\tilde{t}_s = \mathcal{O}(1)$.

In working out (A.22), we shall employ the fact that in (A.16) $\tilde{z}_s = \mu$ is a small parameter. This produces [up to smaller corrections]

$$\sigma = (c + \tilde{t}_s)\sqrt{\tilde{t}_s}. \tag{A.23}$$

Eq. (A.21) derives Eq. (3.10) of the Chapter 3, while (A.23) accounts for the estimate of σ that was presented after Eq. (3.10) of the Chapter 3.

Appendix C - Derivation of Eq. (3.21)

The marginal probability $P(t)$ is defined from (A.11) as

$$P(t) = \int \mathrm{d}\boldsymbol{\theta} P(\boldsymbol{\theta})\, \delta(t - \theta_r). \tag{A.24}$$

using (A.11, A.12) we obtain from (A.24)

$$P(t) \propto u(t) \int_{-i\infty}^{i\infty} \frac{\mathrm{d}\tilde{z}}{2\pi i}\, e^{n\phi(t,\tilde{z}) - \tilde{t}\tilde{z}}, \tag{A.25}$$

$$\phi(t,\tilde{z}) = (1-t)\tilde{z} + \ln \int_0^\infty \mathrm{d}y\, e^{-\tilde{z}y}\, (c+y)^{-2}. \tag{A.26}$$

We use the saddle-point method for (A.25), this produces the same saddle-point equation (A.17) for $\tilde{z}_s$,

$$1 = \frac{\int_0^\infty \mathrm{d}y\, e^{-\tilde{z}_s y}\, (c+y)^{-2}}{\int_0^\infty \mathrm{d}y\, y\, e^{-\tilde{z}_s y}\, (c+y)^{-2}}, \tag{A.27}$$

provided that we note the dominant range $t \propto 1/n \ll 1$ of t. Thus

$$P(\theta) \propto u(\theta) e^{-n\theta\tilde{z}_s}. \tag{A.28}$$

This validates Eq. (3.21) of the Chapter 3.

Likewise, one can show that the marginal density $P(\theta_1, ..., \theta_m)$ factorizes provided that $m \ll n$:

$$P(\theta_1, ..., \theta_m) \propto u(\theta_1) e^{-\mu\theta_1 n} \, ... \, u(\theta_m) e^{-\mu\theta_m n}. \tag{A.29}$$

Eq. (A.29) can be established more heuristically via the exact relation $\overline{[\sum_{k=1}^n \theta_k]^2} = 1$, where $\overline{f}$ means averaging over $P(\theta_1, ..., \theta_n)$. This relation predicts, together with $\overline{\theta_k} = \frac{1}{n}$, that $\overline{\theta_i \theta_j} - \overline{\theta_i}\,\overline{\theta_j} = \mathcal{O}(n^{-3})$, hence approximate factorization.

Using (A.28) with $u(\theta) = (\frac{c}{n} + \theta)^{-2}$ we note that the standard deviation $\langle (\theta - \langle\theta\rangle)^2 \rangle = \frac{1}{n}\sqrt{\frac{c}{\tilde{z}_s} - 1} \simeq \frac{1}{n}\sqrt{\frac{c}{\tilde{z}_s}}$ is larger than the average $\langle\theta\rangle = \int \mathrm{d}\theta\theta P(\theta) = \frac{1}{n}$, since $c/\tilde{z}_s \gg 1$.

Appendix D - Short introduction to Chinese characters

Here we shortly remind the Chinese writing systems, the main differences and similarities between Chinese characters and English words. This subject generated several controversies (myths as it was put in [168]), even among expert sinologists [164–169]. This appendix is necessary for a deeper understanding of our results and motivations.

The main qualitative conclusion of this appendix is that in contrast to English words, generally Chinese characters have more different meanings, they are more flexible, they could combine with other characters to convey different specific meanings. So there are characters, which appear many times in the text, but their concrete meanings are different in different places of the text.

I. Two features of Chinese writing systems

1. The basic unit of Chinese writing system is the character: a spatially marked pattern of strokes phonologically realized as a single syllable (please consult Appendix E for a glossary of various linguistic terms used in the paper). Generally, each character denotes a morpheme or several different morphemes.

2. The Chinese writing system evolved by emphasizing the concept of the character-morpheme, to some extent blurring the concept of the multi-syllable word. In particular, spaces in the Chinese writing system are put in between of characters and not in between of words [1]. Thus a given sentence can have different meanings when being separated into different sequences of words [166], and parsing a string of Chinese characters into words became a non-trivial computational problem; see [170] for a recent review.

[1] An immediate question is whether Chinese readers will benefit from reading a character-written text, where the words boundaries are indicated explicitly. For normal sentences the readers will not benefit, i.e. it does not matter whether the word boundaries are indicated explicitly or not [171]. But for difficult sentences the benefit is there [172].

II. Comparisons of Chinese characters and English words

We list the main differences and similarities between Chinese characters and English words as follows:

1. Psycholinguistic research shows that the characters are important cognitive and perceptual units for Chinese writers and readers [163–165]. We cite here one example. Chinese characters are more directly related to their meanings than English words to their meanings [165] [2]; see Appendix F for additional details. The explanation of this effect would be that characters (compared to English words) are perceived holistically as a meaning-carrying objects, while English words are yet to be reconstructed from a sequence of their constituents (letters) [3].

The word inferiority effect (see its description in Appendix F II) demonstrates that the perception of Chinese characters is not similar to that of English letters [164], and the perception of Chinese words is not similar to the perception of English words [164].

2. One-character words dominate in the following specific sense. Some 54% of modern Chinese word tokens are single-character, two-character word tokens amount to 42%; the remaining words have three or more characters [174]. For modern Chinese word types the situation is different: single character words amount to some 10% against 66% of two-character words [174]. Classic Chinese texts have more single-character words (tokens), the percentage varies between some 60% and 80% for texts written in different periods.

3. A minor part of multi-character words are multi-character morphemes, i.e. their separate characters do not normally appear alone (they are fully bound). Examples of this are the two-character Chinese words for *grape* "葡萄" *(pú táo)*, *dragonfly* "蜻蜓" *(qīng tíng)*, *olive* "橄榄" *(gǎn lǎn)*. Estimates show that some 10% of all characters are fully bound [168].

A related set of examples is provided by two-character words, where the separate

[2] To get a fuller picture of this effect let us denote $\tau_f(E)$ and $\tau_f(C)$ for English and Chinese phonology activation times, respectively, while $\tau_m(E)$ and $\tau_m(C)$ stand for respective meaning activation times. The phonology activation time is the time passed between seeing a word in English (or character in Chinese) and pronouncing it; likewise, for the meaning activation time. Now these quantities hold [165]: $\tau_f(E) < \tau_m(E) > \tau_m(C) \simeq \tau_f(C) > \tau_f(E)$.

[3] A simpler explanation would be that the characters are perceived as pictures (pictograms) directly pointing to their meaning. In its literal form this explanation is not correct, since characters-pictograms are not frequent in Chinese.

characters do have an independent meaning, but this meaning is not directly related to the meaning of the word, e.g. "东西" *(dōng xī)* means *thing*, but literally it amounts to *east-west*, or "手足" *(shǒu zú)* means *close partnership*, but literally *hand-foot*.)

4. The majority of the multi-character words are semantic compounds: their separate characters can stand alone and are related to the overall meaning of the word. Importantly, in most cases, the separate meanings of the component characters are wider than the (relatively unique) meaning of the compound two-character word. An example of this situation is the two-character Chinese word for *train* "火车" *(huǒ chē)*: its first character "火" *(huǒ)* has the meaning of *fire, heat, popular, anger, etc*, while the second character "车" *(chē)* has the meaning of *vehicle, machine, wheeled, lathe, castle, etc.*

Note that in Chinese there is a certain freedom in grouping morpheme into different combinations. Hence it is not easy to distinguish the semantic compounds from lexical phrases.

5. At this point we shall argue that in general Chinese characters have a larger number of different meanings than English words. This statement will certainly appear controversial, if it is taken without proper caution, and is explained without proper usage of linguistic terms (see our glossary at Appendix E).

First of all note the difference between polysemes and homographs: polysemes are two related meanings of the same character (word), homographs are two characters (words) that are written in the same way, but their meanings are far from each other [4]. Now many characters are simultaneously homographs and polysemes, e.g. character "明" *(míng)* means *brilliant, light, clear, next, etc.* Here the first three meanings are related and can be viewed as polysemes. The fourth meaning *next* is clearly different from the previous three. Hence this is a homograph. Another example is the character "发" *(fā or fà)* that can mean *hair, send out, fermentation, etc.* All these three meanings are clearly different; hence we have homographs. Note the following peculiarity of the above two examples: the first example is a non-heteronym (homophonic) character, i.e. it is read in the same way irrespectively whether it means *light* or *next*. The second example is a heteronym character: it written in the same way, but is read differently

[4]Note that polysemes are defined to be related meanings of the same word, while homographs are defined to be different words. This is natural, but also to some extent conventional, e.g. one can still define homographs as far away meanings of the same word.

depending on its meaning.

In most cases, heteronym characters—those which are written in the same way, but have different pronunciations—have at least two sufficiently different meanings. The disambiguation of their meaning is to be provided by the context of the sentence and/or the shared experience of the writer and reader [5].

Surely, also English words can be ambiguous in meaning (e.g. *get* means *obtain*, but also *understand* = *have knowledge*), but there is an essential difference. The major contribution of the meaning ambiguity in English is the polysemy: one word has somewhat different, but also closely related meanings. In contrast, many Chinese characters have widely different meanings, i.e. they are homographs rather than polysemes.

However, we are not aware of any quantitative comparison between homography of Chinese versus English. This may be related to the fact that it is sometimes not easy to distinguish between polysemy and homophony (see the glossary in Appendix E). Still the above statement on Chinese characters having a larger number of different meanings can be quantitatively illustrated via the relative prevalence of heteronyms in Chinese. The amount of heteronyms in English is negligible, e.g. in rather complete list of heteronyms presented in [176], we noted only 74 heteronyms [6], and only three of them had more than 2 meanings. This is a tiny amount of the overall number of English words ($> 5 \times 10^5$). To compare this with the Chinese situation, we note that at least some 14% of modern Chinese and 25% of traditional characters are heteronyms, which normally have at least two widely different meanings. Within the most frequent 5700 modern characters the number of heteronyms is even larger and amounts to almost 22% [174] [7].

[5]Note that homophony in Chinese is much larger than homography: in average a syllable has around 12–13 meanings [163]. Hence, in a sense, characters help to resolve the homophony of Chinese speech. This argument is frequently presented as an advantage of the character-based writing system, though it is not clear whether this system is here not solving the problem that was invited by its usage [173].

[6]Not counting those heteronyms that arise because an English word happens to coincide with a foreign special name, e.g. *Nancy* [English name] and *Nancy* city in France.

[7]One should not conclude that in average the Chinese character has more meanings than the English word, because there is a large number of characters—between 10 % and 14 % depending on the type of the dictionary employed [175]—that do not have lexical meaning, i.e. they are either function words (grammatical meaning mainly) or characters that cannot appear alone (bound characters). If now the number of meanings for each character is estimated via the number of entries in the explanatory dictionary—which is more or less traditional way of searching for the number of meanings, though

6. Chinese nouns are generally less abstract: whenever English creates a new word via conceptualizing the existing one, Chinese tends to explain the meaning via using certain basic characters (morphemes). Several basic examples of this scenario include: length=long+short "长短" *(cháng duǎn)*, landscape=mountains+water "山水" *(shān shuǐ)*, adult=big+person "大人" *(dà rén)*, population=person+ mouth "人口" *(rén kǒu)*, astronomy=heaven+script "天文" *(tiān wén)*, universe=great+emptiness "太空" *(tài kōng)*. English tools for making abstract words include prefixes, *poly-*, *super-*, *pro-*, *etc* and suffixes, *-tion*, *-ment*. These tools either do not have Chinese analogs, or their usage can generally be suppressed.

English words have inflections to indicate the tense of verbs, the number for nouns or the degree for adjectives. Chinese characters generally do not have such linguistic attributes [8], their role is carried out by the context of the sentence(s) [9].

The differences between Chinese and English writing systems can be viewed in the context of the two features: emphasizing the role of base (root) morphemes and delegating the meaning to the context of the sentence whenever this is possible [163].

The quantitative conclusion to be drawn from the above discussion is that Chinese characters have more different meanings, they are flexible, they could combine with other characters to convey different specific meanings. Anticipating our results in the sequel, we expect to see a group of characters, which appear many times in the text, but their concrete meanings are different in different places of the text.

---

it mixes up homography and polysemy—the average number of meanings per a Chinese character appears to be around 1.8–2 [175]. This is smaller than the average number of (necessarily polysemic) meanings for an English word that amounts to 2.3.

[8]Chinese expresses temporal ordering via context, e.g. adding words *tomorrow* or *yesterday*, or by aspects. The difference between tense and aspect is that the former implicitly assumes an external observer, whose reference time is compared with the time of the event described by the sentence. Aspects order events according to whether they are completed, or to which extent they are habitual. Indo-European languages tie up tense and aspect. The tie is weaker for Slavic Indo-European languages. Chinese has several tenses including perfective, imperfective and neutral.

[9]Chinese has certain affixes, but they can be and are suppressed whenever the issue is clear from the context.

Appendix E - Glossary

• Classic Chinese (*wén yán*) written language employed in China till the early XX (20th) century, and usually it is recognized that Classic Chinese evolved to Modern Chinese since the May Fourth Movement in 1919. Still the Modern Chinese keeps many elements of Classic Chinese. As compared to the Modern Chinese, the Classic Chinese has the following peculiarities (1) It is more lapidary: texts contain almost two times smaller amount of characters, since the Classic Chinese is dominated by one-character words. (2) It lacks punctuation signs and affixes. (3) It relies more on the context. (4) It frequently omits grammatical subjects.

• Content word (character): A word that has an independent meaning can be given by reference to a word outside any sentence in which the word may occur. Content words are said to have a lexical meaning, rather than indicating a syntactic (grammatical) function, as a function word does.

• Empty Chinese characters—e.g. "几" *(jǐ)* or "已" *(yǐ)* —serve for establishing numerals for nouns, aspects for verbs *etc*. In contrast, to function characters, they cannot be used alone, i.e. they are fully bound.

• Frequency dictionary collects words used in some activity (e.g. in exact science, or daily newspapers *etc*) and orders those words according to the frequency of usage. Frequency dictionaries can be viewed as big mixtures of different texts.

• Function word (character): is a word that has little lexical meaning or have ambiguous meaning, but instead serves to express grammatical relationships with other words within a sentence, or specify the attitude or mood of the speaker. Such words are said to have a grammatical meaning mainly.

• Hapax legomena: Set of words (characters) that appeared only once in a text. In a more general sense, set of words (characters) that appear in a text only few times. An important feature of a text written by a human subject is that the text contains a sizable amount of words (characters) that appear only one time; it is not difficult to imagine an artificial text (or purposefully modified natural text) that will not contain at all words that appear only once.

• Homophones: two different words that are pronounced in the same way, but may be written differently, e.g. *rain* and *reign*.

• Homographs: two different words (or characters) that are written in the same

way, but may be pronounced differently, e.g. *shower* [precipitation] and *shower* [the one who shows]. This example is a proper homograph, since the pronunciation is different. Another example (of both homography and homonymy) is *present* [gift] and *present* [the current moment of time]. Note that the distinction between homographs and polysemes is not sharp and sometimes difficult to make. There are various boundary situations, e.g. the verb *read* [present] and *read* [past] may qualify as homograph, but the meanings expressed are close to each other.

• Homonymes: two words (or characters) that are simultaneously homographs and homophones, e.g. *left* [past of *leave*] and *left* [opposite of *right*]. Some homonymes started out as polysemes, but then developed a substantial difference in meaning, e.g. *close* [near] and *close* [to shut (lips)].

• Heteronyms: two homographs that are not homophones, i.e. they are written in the same way, but are pronounced differently. Normally, heteronyms have at least two sufficiently different meanings, indicated by different pronunciations.

• Key-word (key-character): A content word (character) that characterizes a given text with its specific subject. The operational definition of a key-word (key-character) is that in a given text its frequency is much larger than in a frequency dictionary, which was obtained by mixing together a big mixture of different texts.

• Language family A set of related languages that are believed (or proved) to originate from a common ancestor language.

• Latent semantic analysis The analysis of word frequencies and word-word correlations (hence semantic relations) in a text that is based on the idea of hidden (latent) variables that control the usage of words; see [178] for reviews.

• Logographic writing system is based on the direct coding of morphemes.

• Mental lexicon: the store of words in the long-time memory. The words from the mental lexicon are employed on-line for expressing thoughts via phrases and sentences; see [127, 150] for detailed theories of the mental lexicon. Ref. [150] argues that in addition to mental lexicon humans contain a mental syllabary that is activated during the phonologization of a word that was already extracted from the mental lexicon.

• Morpheme: the "atom" of meaning: the smallest part of the speech or writing that has a separate (not necessarily unique) meaning, e.g. *cats* has two morphemes: *cat* and *-s*. The first morpheme can stand alone. The second one expresses the grammatical

meaning of plurality, but it is a bound morpheme, since it can appear only together with other morphemes.

• Polysemes are related meanings of the same word, e.g. the English word *get* means *obtain/have*, but also *understand* (= *have knowledge*). Another example is that many English nouns are simultaneously verbs (e.g. *advocate* [person] and *advocate* [to defend]).

• Syllable is the minimal phonetic unit characterized by acoustic integrity of its components (sounds), e.g. the word *body* is composed of two syllables: *bo-* and *-dy*, while *consider* consists of three syllables: *con- -si- -der*. In phonetic languages such as Russian the factorization of the word into syllables (syllabification) is straightforward, since the number of syllables directly relates to the number of vowels. In non-phonetic languages such as English, the correct syllabification can be complicated and not readily available to non-experts. Indo-European languages typically have many syllables, e.g. the total number of English syllables is more 10 000. However, 80 % of speech employs only 500-600 frequent syllables [150]. It was argued, based on psycholinguistic studies, that the frequent syllables are also stored in the long-term memory analogously to mental lexicon [150]. The total number of Chinese syllables is much less, around 500 (about 1200 together with tones) [150, 175]. Syllabification in Chinese is generally straightforward too, also because each character corresponds to a syllable.

• Writing system: the process or result of recording spoken language using a system of visual marks on a surface. There are two major types of writing systems: logographic (Sumerian cuneiforms, Egyptian hieroglyphs, Chinese characters) and phonographic. The latter includes syllabic writing (Japanese hiragana) and alphabetic writing (English, Russian, German).

Appendix F - Interference experiments distinguishing between the Chinese characters and the English words

The general scheme of interference experiments in psychology is described as follows [165, 166]. There are two tasks, the main one and the auxiliary one. Each task is defined via specific instructions. The subjects are asked to carry out the main task simultaneously trying to ignore the auxiliary task. The performance times for carrying out the main task in the presence of the auxiliary one are then compared with the performance times of the main task when the auxiliary task is absent, or at least it does not interfere with the main task.

I. The Stroop effect.

The main task is to call the color of words. The auxiliary task is not to pay attention at the meaning of those words. The experiment is designed such that there is an incongruency between the semantic meaning of the word and its color, e.g. the word *red* is written in black. As compared to the situation when the incongruency is absent, i.e. the word *red* is written in red, the reaction time of performing the main task is sizably larger. This is the essence of the Stroop effect: the semantic meaning interferes with the color perception.

It appears that the Stroop effect is larger for Chinese characters than for English words; see [165] for a review. This is one of the arguments that the getting to the meaning of a Chinese character is faster than to the meaning of an English word.

II. The word inferiority/superiority effect

If English-speaking subjects are asked to trace out (and count) a specific letter in a text, they make less errors, when the text is meaningless, i.e. it consists of meaningless strings of letters [164]. This is related to the fact that English words are recognized and stored as a whole. Hence the recognition of words interferes with the task of identifying the letter, and the English-speaking subjects make more errors when tracing out a letter in a meaningful text. In contrast to this, Chinese-speaking adults display the

word priority effect: they do less errors in tracing out a given character in a string of meaningful characters, as compared to tracing it out in a list of meaningless pseudo-characters [164]. The effect is reversed, if the Chinese subjects are asked to trace out a specific stroke within a character: in analogy to the English situation it is easier for Chinese speakers to trace out the stroke in a meaningless pseudo-character than in a meaningful character [164].

These results imply that the recognition of Chinese characters is more similar to the recognition of English words, while the recognition of Chinese words is less similar to the recognition of English words.

Appendix G: Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test (KS test) is used to determine if a data sample agrees with a reference probability distribution. The basic idea of the KS test is as follows.

We need to determine whether a given set X_1, X_2, ... , X_n is generated by i.i.d sampling a random variable with cumulative probability distribution $F(x)$ (null hypothesis). To this end we calculate the the empiric cumulative distribution function (CDF) $F_n(x)$ for X_1, X_2, ... , X_n:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x}, \tag{A.30}$$

where $I_{X_i \leq x}$ equals to 1 if $X_i \leq x$ and 0 otherwise. Next we define:

$$D_n = \sup_x |F_n(x) - F(x)|. \tag{A.31}$$

The advantage of using D_n (against other measures of distance between $F_n(x)$ and $F(x)$) is that if the null hypothesis is true, the probability distribution of D_n does not depend on $F(x)$. In that case it was shown that for $n \to \infty$, the cumulative probability distribution of $\sqrt{n}D_n$ is:

$$P(\sqrt{n}D_n \leq x) \equiv f(x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2}. \tag{A.32}$$

For not rejecting the null hypothesis we need that the observed value of $\sqrt{n}D_n^*$ is sufficiently small. To quantify that smallness we take a parameter (significance level) α ($0 < \alpha < 1$) and define κ_α as the unique solution of

$$f(\kappa_\alpha) = 1 - \alpha. \tag{A.33}$$

Now the null hypothesis is not rejected provided that

$$\sqrt{n}D_n^* < \kappa_\alpha, \tag{A.34}$$

where $\sqrt{n}D_n^*$ is the observed (calculated) value of D_n. Condition (A.34) ensures that if the null hypothesis is true, the probability to reject it is bounded from below by α. Hence in practice one takes, e.g. $\alpha = 0.05$ or $\alpha = 0.01$.

Note however that condition (A.34) will always hold provided that α is taken sufficiently small. Hence to quantify the goodness of the null hypothesis one should

Table 1: Kolmogorov-Smirnov test (KS test) for the fitting quality of our results. In the KS test, D and p denote the maximum difference (test statistics) and p-value respectively. D_1 and p_1 are calculated from the KS test between empiric data and numerical fitting, D_2 and p_2 are between empiric data and theoretical result, D_3 and p_3 are between numerical fitting and theoretical result. Note that for making the testing even more vigorous, the presented results for the KS characteristics are obtained in the original coordinates; similar results are obtained in logarithmical coordinates that are employed for the linear fitting.

Texts	D_1	p_1	D_2	p_2	D_3	p_3
TF	0.0418	0.865	0.0365	0.939	0.0381	0.912
TM	0.0529	0.682	0.0562	0.593	0.0581	0.568
AR	0.0564	0.624	0.0469	0.783	0.0443	0.825
DL	0.0451	0.812	0.0421	0.865	0.0472	0.761
AQZ	0.0586	0.587	0.0565	0.623	0.0601	0.564
KLS	0.0592	0.578	0.0641	0.496	0.0626	0.521
CQF	0.0341	0.962	0.0415	0.863	0.0421	0.857
SBZ	0.0461	0.796	0.0558	0.635	0.0616	0.538
WJZ	0.0427	0.852	0.0475	0.753	0.0524	0.691
HLJ	0.0375	0.923	0.0412	0.875	0.0425	0.862

calculate the p-value p : the maximal value of α, where (A.34) still holds. For the hypothesis to be reliable one needs that p is not very small. As an empiric creterion of reliability people frequently take $p > 0.1$.

We applied the KS test to our data on the character (word) frequencies. The empiric results on word frequencies f_r in the Zipfian range $[r_{\min}, r_{\max}]$ are fit to the power law, and then also to the theoretical prediction. With null hypothesis that empiric data follows the numerical fittings and/or theoretical results, we calculated the maximum differences (test statistics) D and the corresponding p-values in the KS tests. From the above table one could observe that all the test statistics D are quite small, while the p-values are *much larger* than 0.1. We conclude that from the viewpoint of the KS test the numerical fittings and theoretical results can be used to characterize the empiric data in the Zipfian range reasonably well.

Appendix H - A list of the studied texts

1) Two short modern Chinese texts:

- 昆仑觞, *Kūn Lún Shāng* (KLS) by Shu Ming Bi, 1987, (the total number of characters $N = 20226$, the number of different characters $n = 2047$). The text is about the arduous military training in the troops of Kun Lun mountain.

- 阿 Q 正传, *Ah Q Zhèng Zhuàn* (AQZ) by Xun Lu, 1922, ($N = 18153$, $n = 1553$). The story traces the "adventures" of a hypocrit and conformist called Ah Q, who is famous for what he presents as "spiritual victories".

2) Two long modern Chinese texts:

- 平凡的世界, *Píng Fán de Shì Jiè* (PFSJ) by Yao Lu, 1986, ($N = 705130$, $n = 3820$). The novel depicts many ordinary people's stories which include labor and love, setbacks and pursue, pain and joy, daily life and huge social conflict.

- 水浒传, *Shuǐ Hǔ Zhuàn* (SHZ) by Nai An Shi, 14th century, ($N = 704936$, $n = 4376$). The story tells how a group of 108 outlaws gathered at Mount Liang formed a sizable army before they were eventually granted amnesty by the government and sent on campaigns to resist foreign invaders and suppress rebel forces.

3) Four short classic Chinese texts:

- 春秋繁露, *Chūn Qiū Fán Lù* (CQF), by Zhong Shu Dong, 179-104 BC, (Vol.**1**-Vol.**8**, $N = 30017$, $n = 1661$). A commentary on the Confucian thought and teachings.

- 僧宝传, *Sēng Bǎo Zhuàn* (SBZ), by Hong Hui, 1124, (Vol.**1**-Vol.**7**, $N = 24634$, $n = 1959$). A commentary on the Taoist thought and teachings. Biographies of great Taoist masters.

- 武经总要, *Wǔ Jīng Zǒng Yào* (WJZ), by Gong Liang Zeng and Du Ding, 1040-1044, (Vol.**1**-Vol.**4**, $N = 26330$, $n = 1708$). A Chinese military compendium. The text covers a wide range of subjects, from naval warships to different types of catapults.

- 虎玲经, *Hǔ Líng Jīng* (HLJ), by Dong Xu, 1004, (Vol.**1**-Vol.**7**, $N = 26559$, $n = 1837$). Reviews various military strategies and relates them to factors of geography and climate.

4) A long classic Chinese text:

- 史记, *Shǐ Jì* (SJ), by Qian Sima, 109 to 91 BC, ($N = 572864$, $n = 4932$). Reviews imperial biographies, tables, treatises, biographies of feudal houses and eminent

persons.

5) Four short English texts:

-*The Age of Reason* (AR) by T. Paine, 1794 (the major source of British deism, $N = 22641$, $n = 1706$).

-*Thoughts on the Funding System and its Effects* (TF) by P. Ravenstone, 1824 (economics, $N = 26624$, $n = 2067$).

-*Time Machine* (TM) by H. G. Wells, 1895 (a science fiction classics, $N = 31567$, $n = 2612$).

-*Dream Lover* (DL) by J. MacIntyre, 1987 (a romance novella, $N = 24990$, $n = 1748$).

Bibliography

[1] L. M. A. Bettencourt, J. Lobo, D. Helbing, *et al.*, *PNAS*, **104**, 7301 (2007).

[2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D. U. Hwang, *Physics reports*, **424**, 175 (2006).

[3] C. S. Holling, *Ecosystems*, **4**, 390 (2001).

[4] D. Challet and Y. C. Zhang, *Physica A*, **246**, 407 (1997).

[5] T. Zhou, J. Ren, M. Medo and Y. C. Zhang, *Phys. Rev. E*, **76**, 046115 (2007).

[6] A. Schenkel, J. Zhang, Y. C. Zhang, *Fractals*, **1**, 47 (1993).

[7] M. Marsili and Y. C. Zhang, *Phys. Rev. Letts.*, **80**, 2741 (1998).

[8] R. Crane and D. Sornette, *PNAS*, **105**, 15649 (2008).

[9] B. Carlsson and R. Stankiewicz, *J. Evol. Econ.*, **1**, 93 (1991).

[10] C. Castellano, S. Fortunato, and V. Loreto, *Rev. Mod. Phys.*, **81**, 591 (2009).

[11] F. Leon, *Mathematical Problems in Engineering*, **2012**, 857512 (2012).

[12] M. Lagi, Yavni Bar-Yam, K.Z. Bertrand, Yaneer Bar-Yam, *arXiv:1203.1313*, March 6, 2012.

[13] M. Lim, R. Metzler, and Y. Bar-Yam, *Science*, **317**, 5844 (2007).

[14] D. Helbing, W. J. Yu, and H. Rauhut, *The Journal of Mathematical Sociology*, **35**, 177 (2011).

[15] R. Cont and J. P. Bouchaud, *Macroeconomic Dynamics*, **4**, 170 (2000).

[16] V. Robu, H. Halpin, and H. Shepherd, *ACM Transactions on the Web*, **3**, 14 (2009).

[17] D. Pumain, *SFI Working paper*, **2**, 2 (2004).

[18] R. Carvalho, S. Iida and A. Penn, *Proceedings of 4th International Space Syntax Symposium*, **34**, London (2003).

[19] T. Parsons, The social system, *Routledge Sociology Classics Series Major Languages*, Routledge, (1991).

[20] E. Ravenstein, *The Geographical Magazine III*, 173 - 177, 201 - 206, 229 - 233, 1876.

[21] G.K. Zipf, *American Sociological Review*, **11**, 677 (1946).

[22] J. Tinbergen, Shaping the World Economy, *Twentieth Century Fund*, 1962.

[23] V. Pareto, Translation of Manuale di economia politica ("Manual of political economy"), *A.M. Kelley*, 1971.

[24] I. Fisher, The Theory of Interest, *Augustus M. Kelley Publishers*, 1930.

[25] G. K. Zipf, The economy of geography, *Addison-Wesley Publishing Co. Inc*, pages 347 - 415, 1949.

[26] W. T. Li, *Glottometrics*, **5**, 14 (2002).

[27] X. Gabaix, *The Quarterly Journal of Economics*, **114**, 739 (1999).

[28] L. A. Adamic and B. A. Huberman, *Glottometrics*, **3**, 143 (2002).

[29] Z. K. Silagadze, *arXiv: 9901035v2*, 26 Jan 1999.

[30] B. B. Mandelbrot, *Journal of Business*, **36**, 394 (1963).

[31] R. N. Mantegna and E. H. Stanley, An Introduction to Econophysics: Correlations and Complexity in Finance, *Cambridge University Press*, November 1999.

[32] A. J. McNeil and R. Frey, *Journal of Empirical Finance*, **7**, 271 (2000).

[33] R. K. Merton, *Science*, **159**, 56 (1968).

[34] J. K. Galbraith, The rich got richer, *Salon Magazine*, June 8, 2004.

[35] T. W. Swan, *Economic Record*, **32**, 334 (1956).

[36] A. L. Barabási and R. Albert, *Science*, **286**, 509 (1999).

[37] C. A. E. Goodhart, Monetary relationships: a view from Threadneedle Street, *Monetary economics*, **1**, 1975.

[38] J. Danielsson, *Journal of Banking & Finance*, **26**, 1273 (2002).

[39] R. Dunbar, Grooming, gossip, and the evolution of language. *Harvard University Press*, 1998.

[40] M. E. J. Newman, *Contemporary Physics*, **46**, 323 (2005).

[41] A. Clauset, C. R. Shalizi, and M. E. J. Newman, *SIAM Review*, **51**, 661 (2009).

[42] D. R. Anderson, D. J. Sweeney, and T. A. Williams, Introduction to Statistics: Concepts and Applications, pp. 5 – 9. West Group. ISBN 978-0-314-03309-3, (1994).

[43] J. D. Hamilton, *Econometric Theory*, **11**, 625 (1995).

[44] A. O. Sykes, An Introduction to Regression Analysis, Chicago Working Paper in Law & Economics, 1992.

[45] R. Fisher, *Journal of the Royal Statistical Society, Series B (Methodological)*, **17**, 69 (1955).

[46] D. Madigan, Statistical Analysis and Data Mining, Wiley Public Online Library, online ISSN: 1932-1872.

[47] M. E. J. Newman, *SIAM Rev.*, **45**, 167 (2003).

[48] R. Albert, A. L. Barabási, *Rev. Mod. Phys.*, **74**, 47 (2002).

[49] S. H. Strogatz, *Nature*, **410**, 268 (2001).

[50] S. N. Dorogovtesev, J. F. F. Mendes, *Adv. Phys.*, **51**, 1079 (2002).

[51] D. J. Watts, Small Worlds: The Dynamics of Networks between Order and Randomness, *Princeton University Press*, Princeton, NJ, 1999.

[52] P. Mariolis, *Social Science Quarterly*, **56**, 425 (1975).

[53] V. Latora, and M. Marchiori, *Physica A*, **314**, 109 (2002).

[54] W. Li and X. Cai, *Phys. Rev. E*, **69** 046106 (2004).

[55] W. Li and X. Cai, *Physica A*, **382** 693 (2007).

[56] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles, *Nature*, **420**, 190 (2002).

[57] S. N. Dorogovtesev, J. F. F. Mendes, Evolution of Networks, *Oxford University Press*, Oxford, 2003.

[58] P. Holme, J. Saramakid, *Physics Reports*, **519**, 97 (2012).

[59] G. Rasch, Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests [J]. 1960.

[60] D. R. Rhodes, *et al*, *Nature Biotechnology*, **23**, 951 (2005).

[61] B. Merialdo, *Computational Linguistics archive*, **20**, 155 (1994).

[62] J. M. Ponte, W. B. Croft, *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 275, 1998.

[63] C. X. Zhai, J. Lafferty, *Proceedings of the tenth international conference on Information and knowledge management*, 403, 2001.

[64] B. Roark, *Computational Linguistics*, **27**, 249 (2006).

[65] Y. Bengio, R. Ducharme, P. Vincent, C. Janvin, *The Journal of Machine Learning Research archive*, **3**, 1137 (2003).

[66] D. Jurafsky, *Cognitive Science*, **20**, 137 (1996).

[67] E. Bonabeau, *PNAS*, **99**, 7280 (2002).

[68] D. Helbing and S. Balietti, *SFI Working paper*, **06**, 024 (2011).

[69] M. Niazi, A. Hussain, *Scientometrics*, **89**, 479 (2011).

[70] http://en.wikipedia.org/wiki/Agent-based_model

[71] J. H. Holland, J. H. Miller, *American Economic Review*, **81**, 365 (1991).

[72] A. Robert, The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration, *Princeton University Press*, Princeton: ISBN 978-0-691-01567-5 (1997).

[73] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, *J. Chem. Phys.*, **21**, 1087 (1953).

[74] E. J. Beltrami, Mathematics for dynamic modeling, NY: Academic Press, (1987).

[75] H. K. Khalil, Nonlinear Systems, Prentice Hall, ISBN 0-13-067389-7, (2001).

[76] S. H. Strogatz, Nonlinear dynamics and chaos, Reading, MA: Addison Wesley, (1994).

[77] D. Helbing, Quantitative sociodynamics (Springer, Hei- delberg, Germany), 1991.

[78] D. Helbing, *Physica A*, **196**, 546 (1993).

[79] W. Weidlich, *Phys. Rep.*, **204**, 1 (1991).

[80] W. Weidlich, Sociodynamics: A Systematic Approach to Mathematical Modelling in Social Sciences (Taylor and Francis, London, UK), 2002.

[81] H. Haken, Synergetics; An Introduction. Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry and Biology (Springer Verlag, Berlin-Heidelberg, Germany), 1978.

[82] D. Helbing, *Physica A*, **193**, 241 (1993).

[83] D. Helbing, *J. Math. Sociol.*, **19**, 189 (1994).

[84] http://en.wikipedia.org/wiki/Critical_phenomena

[85] M. E. Fisher, *Rev. Mod. Phys.*, **46**, 597 (1974).

[86] I. S. Aranson and L. S. Tsimring, *Rev. Mod. Phys.*, **78**, 641 (2006).

[87] R. J. Glauber, *J. Math. Phys.*, **4**, 294 (1963);

[88] I. D. Couzin and J. Krause, *Advances in the study of behavior*, **32**, 1 (2003).

[89] J. L. Deneubourg and S. Goss, *Ethology ecology & Evolution*, **1**, 295 (1989).

[90] F. Wu, *Rev. Mod. Phys.*, **54**, 235 (1982).

[91] R. Toral and C. J. Tessone, *Comm. Comput. Phys.*, **2**, 177 (2007).

[92] M. Marsili and Y. C. Zhang, *Phys. Rev. Lett.*, **80**, 2741 (1998).

[93] A. M. Petersen, F. Wang and H. E. Stanley, *Phys. Rev. E*, **81**, 036114 (2010).

[94] A. M. Petersen, W. S. Jung, J. S. Yang and H. E. Stanley, *Proc. Natl. Acad. Sci.*, **108**, 18 (2011).

[95] A. M. Petersen, H. E. Stanley and S. Succi, *Scientific Reports*, **1**, 181 (2011).

[96] F. Radicchi, *PLoS ONE*, **6**, e17249 (2011).

[97] E. Ben-Naim, S. Redner and F. Vazquez, *Europhys. Lett.*, **77**, 30005 (2007).

[98] E. Ben-Naim, F. Vazquez and S. Redner, *J. Korean Phys. Soc.*, **50**, 124 (2007).

[99] *www.atpworldtour.com, www.wtatennis.com, www.pga.com, www.lpga.com, www.ittf.com, www.fivb.org, www.fifa.com, www.worldsnooker.com, www.fie.ch, www.bwfbadminton.org, www.fiba.com, www.ibaf.org, www.fih.ch, www.ihf.info.*

[100] P. Minnhagen, *PHYSICS TODAY, POINTS OF VIEW* (November 3 2011).

[101] S. K. Baek, S. Bernhardsson and P. Minnhagen, *New J. Phys.*, **13**, 043004 (2011).

[102] S. K. Baek, P. Minnhagen and B. J. Kim, *New J. Phys.*, **13**, 073036 (2011).

[103] F. J. Massey, *J AM STAT ASSOC.*, **46**, 253 (1951).

[104] V. Pareto, A. N. Page, A. M. Kelley, Translation of Manuale di economia politica ("Manual of political economy") ISBN 9780678008812, 1971.

[105] R. A. Bradley and M. E. Terry, *Biometrica*, **39**, 324 (1952).

[106] P. Bak, C. Tang and K. Wiesenfeld, *Phys. Rev. Lett.*, **59**, 381 (1987).

[107] P. Bak, C. Tang and K. Wiesenfeld, *Phys. Rev. A*, **38**, 364 (1988).

[108] R. Toral and C. J. Tessone *Commun. Comput. Phys.*, **2**, 177 (2007).

[109] S. N. Dorogovtsev, A. V. Goltsev and J. F. F. Mendes, *Rev. Mod. Phys.*, **80**, 1275 (2008).

[110] *Zipf's law*, in http://en.wikipedia.org/wiki/Zipf_law.

[111] H. Baayen, *Word frequency distribution* (Kluwer Academic Publishers, 2001).

[112] B. Mandelbrot, *Fractal geometry of nature* (W. H. Freeman, New York, 1983).

[113] G.A. Miller, Am. J. Psyc. **70**, 311(1957). W.T. Li, IEEE Inform. Theory, **38**, 1842 (1992).

[114] Yu.A. Shrejder and A.A. Sharov, *Systems and Models* (Moscow, Radio i Svyaz, 1982) (In Russian).

[115] R. Ferrer-i-Cancho and R. Solé, PNAS, **100**, 788 (2003).

[116] M. Prokopenko *et al.*, JSTAT, P11025 (2010).

[117] V. Dunaev, Aut. Doc. Math. Linguistics, 14 (1984).

[118] B. Corominas-Murtra *et al.*, Phys. Rev. E **83**, 036115 (2011).

[119] D. Manin, Cognitive Science, **32**, 1075 (2008).

[120] Y. A. Shrejder, Prob. Inform. Trans. **3**, 57 (1967).

[121] M. V. Arapov and Y. A. Shrejder, in *Semiotics and Informatics*, v. 10, p. 74 (Moscow, VINITI, 1978).

[122] H. A. Simon, Biometrika **42**, 425 (1955).

[123] D. H. Zanette and M. A. Montemurro, J. Quant. Ling. **12**, 29 (2005).

[124] I. Kanter and D. A. Kessler, Phys. Rev. Lett. **74**, 4559 (1995).

[125] B. M. Hill, J. Am. Stat. Ass. **69**, 1017 (1974). G. Troll and P. beim Graben, Phys. Rev. E **57**, 1347 (1998). A. Czirok *et al.*, *ibid.* **53**, 6371 (1996).

[126] R. Ferrer-i-Cancho and B. Elveva, PLoS ONE, **5**, 9411 (2010).

[127] http://en.wikipedia.org/wiki/Mental_lexicon.

[128] W.J.M. Levelt *et al.*, Beh. Brain Sciences, **22**, 1 (1999).

[129] http://en.wikipedia.org/wiki/Coefficient_of_determination

[130] H. P. Luhn, IBM J. Res. Devel. **2**, 159 (1958).

[131] T. Hofmann, *Probabilistic Latent Semantic Analysis* , in Uncertainty in Artificial Intelligence, 1999.

[132] R. E. Madsen *et al.*, *Modeling word burstiness using the Dirichlet distribution*, in Proc. Intl. Conf. Machine Learning, 2005.

[133] S.M. Gusein-Zade, Prob. Inform. Trans. **24**, 338 (1988).

[134] L. Pietronero *et al.*, Physica A **293**, 297 (2001). L.A. Adamic and B.A. Huberman, Glottometrics, **3**, 143 (2002). R. Rousseau, *ibid.* **3**, 11 (2002).

[135] M. Jaeger, Int. J. Approx. Reas. **38**, 217 (2005).

[136] http://en.wikipedia.org/wiki/Coefficient_of_determination

[137] See http://en.wikipedia.org/wiki/Mental_ lexicon, and http://en.wikiversity.org/wiki/Psycholinguistics/The_ Mental_ lexicon,

[138] N. Jiang, *Applied Linguistics*, **21**, 47 (2000).

[139] M. Jaeger, Int. J. Approx. Reas. **38**, 217 (2005).

[140] E.T. Jaynes, IEEE Trans. Syst. Science & Cyb. **4**, 227 (1968).

[141] See http://en.wikipedia.org/wiki/Keywords

[142] W. J. M. Levelt and A.S. Meyer, European Journal of Cognitive Psychology, **12**, 433 (2000).

W. J. M. Levelt *et al.*, Behavioral and Brain Sciences, **22**, 1 (1999).

[143] C.M. McLeod and K. E. Kampe, J. Exp. Psychology: Learning, Memory, Cognition, **22**, 132 (1996).

[144] N. Hatzigeorgiu, G. Mikros, and G. Carayannis, Journal of Quantitative Linguistics, **8**, 175 (2001).

[145] B.D. Jayaram and M.N. Vidya, Journal of Quantitative Linguistics, **15**, 293 (2008).

[146] J.B. Estoup, *Gammes sténographique* (Institut Sténographique de France, Paris, 1916).

[147] G.A. Miller, Am. J. Psyc. **70**, 311(1957). W.T. Li, IEEE Inform. Theory, **38**, 1842 (1992).

[148] M.V. Arapov and Yu.A. Shrejder, in *Semiotics and Informatics*, v. 10, p. 74 (Moscow, VINITI, 1978). H.A. Simon, Biometrika **42**, 425 (1955). I. Kanter and D. A. Kessler, Phys. Rev. Lett. **74**, 4559 (1995). B.M. Hill, J. Am. Stat. Ass. **69**, 1017 (1974). G. Troll and P. beim Graben, Phys. Rev. E **57**, 1347 (1998). A. Czirok *et al.*, *ibid.* **53**, 6371 (1996). K. E. Kechedzhi *et al.*, *ibid.* **72** (2005).

[149] A.E. Allahverdyan, Weibing Deng and Q.A. Wang, *Explaining Zipf's law via mental lexicon*, arXiv:1302.4383. Available at http://xxx.lanl.gov/abs/1302.4383

[150] W.J.M. Levelt *et al.*, Beh. Brain Sciences, **22**, 1 (1999).

[151] http://en.wikipedia.org/wiki/Coefficient_of_determination

[152] H. P. Luhn, IBM J. Res. Devel. **2**, 159 (1958).

[153] G.K. Zipf, *Selected studies of the principle of relative frequency in language*, (Harvard University Press, Cambridge MA, 1932)

[154] K.-H. Zhao, Am. J. Phys. 58, 449 (1990).

[155] R. Rousseau and Q. Zhang, Scientometrics, **24**, 201 (1992).

[156] S. Shtrikman, Journal of Information Science, 142 **20** (1994).

[157] Wang Dahui *et al.*, Physica A **358**, 545 (2005).

[158] Q. Chen, J. Guo and Y. Liu, Journal of Quantitative Linguistics, **19**, 232 (2012).

[159] S.-M. Huang *et al.*, Decision Support Systems, **46**, 70 (2008).

[160] D.M.W. Powers, *Applications and explanations of Zipf's law*, in: D.M.W. Powers (ed.), New Methods in Language Processing and Computational Natural Language Learning (NEMLAP3/CONLL98), ACL, 1998, pp. 151-160.

[161] Le Quan Ha *et al.*, *Extension of Zipf's Law to Words and Phrases*, Proceedings of the 19th international conference on Computational linguistics, **1**, pp. 1-6, (2002).

[162] J. Elliott *et al.*, *Language identification in unknown signals*, Proceedings of the 18th conference on Computational linguistics, **2**, pp. 1021-1025, (2000).

[163] D. Aaronson and S. Ferres, J. Memory and Language, **25**, 136 (1986).

[164] H.-C. Chen, *Reading comprehension in Chinese*, in H.-C. Chen & O. J. L. Tzeng (Eds.), Language processing in Chinese (pp. 175- 205). Amsterdam, Elsevier, 1992.

[165] R. Hoosain, *Speed of getting at the phonology and meaning of Chinese words*, in *Cognitive neuroscience studies of Chinese language*, H.S.R. Kao, C.-K. Leong and D.-G. Gao (eds.) (Hong kong University Press, Hong kong, 2002).

[166] R. Hoosain, *Psychological reality of the word in Chinese*, in H.-C. Chen and J.-L. Tseng (eds.), Language processing in Chinese, pp. 111-130, (Amsterdam, Netherlands, 1992).

[167] G. Sampson, Linguistics, **32**, 117 (1994).

[168] J. DeFrancis, *Visible Speech: the Diverse Oneness of Writing Systems* (University of Hawaii Press, Honulu, 1989).

[169] J. L. Packard, *The Morphology of Chinese: A linguistic and cognitive approach* (Cambridge University Press, Cambridge, 2000).

[170] X. Luo, *A maximum entropy Chinese character-based parser*. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 2003.

[171] I.-M. Liu *et al.* Chinese Journal of Psychology, **16**, 25 (1974).

[172] S.-H. Hsu and K.-C. Huang, Perceptual and Motor Skills, **91**, 355 (2000); *ibid.* **90**, 81 (2000).

[173] Wm. C. Hannas, *Asia's Orthographic Dilemma* (University of Hawai'i Press, 1997).

[174] C.-Y. Chen *et al.*, *Some distributional properties of Madanrin Chinese*, Proceedings of the first Pasific Asia conference on formal and computational linguistics, p. 81 (Taipei, 1993).

[175] N.V. Obukhova, Quantitative linguistics and automatic text analysis (Proc. of Tartu university), **745**, 119 (1986).

[176] http://myweb.tiscali.co.uk/wordscape/wordlist/homogrph.html

[177] F.-L. Huang, Int. J. on Signal & Image Processing, **2**, 20 (2011).

[178] T. Hofmann, *Probabilistic Latent Semantic Analysis*, in Uncertainty in Artificial Intelligence, 1999.

Publications and pre-prints during PhD:

1. **W. B. Deng**, Armen E. Allahverdyan, and Q. A. Wang, Rank-frequency relation for Chinese characters, submitted to *PLOS ONE*.

2. Armen E. Allahverdyan, **W. B. Deng**, and Q. A. Wang, Explaining the Zipf's law via Mental Lexicon, submitted to *Physical Review E*.

3. **W. B. Deng**, W. Li, X. Cai, and Q. A. Wang, An investigation on the stock holding period: Empirical evidence and the agent-based model, submitted to *Physica A*.

4. **W. B. Deng**, W. Li, X. Cai, A. Bulou, and Q. A. Wang, Universal scaling in sports ranking, *New Journal of Physics* 14 (2012) 093038.

   This paper was highlighted by *Nature, Science Daily*, IOP and MIT *Technology Review*.

5. **W. B. Deng**, D. J. Wang, W. Li, and Q. A. Wang, English and Chinese language frequency time series analysis, *Chinese Science Bulletin* 56 (2011) 3717-3722.