Proceedings of

STATISTICS OF LANGUAGES THEORIES AND EXPERIMENTS

An Interdisciplinary Workshop

Warsaw, Poland, 19th–22nd July 2017

Sponsored by

SAMSUNG

Held at

Institute of Computer Science
Polish Academy of Sciences

Proceedings of

STATISTICS OF LANGUAGES THEORIES AND EXPERIMENTS

An Interdisciplinary Workshop

Warsaw, Poland, 19th–22nd July 2017

Edited by

Łukasz Dębowski
Ramon Ferrer-i-Cancho
Kumiko Tanaka-Ishii
Paweł Teisseyre

Workshop website

`http://statlang2017.ipipan.waw.pl/`

Program committee

- Łukasz Dębowski (Polish Academy of Sciences)
- Ramon Ferrer-i-Cancho (Universitat Politecnica de Catalunya)
- Kumiko Tanaka-Ishii (University of Tokyo)

Local organizing committee

- Łukasz Dębowski (Polish Academy of Sciences)
- Paweł Teisseyre (Polish Academy of Sciences)
- Jan Ziółkowski (Polish Academy of Sciences)

Contents

Preface

Statistical linguistics a.k.a. quantitative linguistics is a difficult domain of science. Ideally, researchers working in this area should have a solid background in both humanities and hard sciences. Whereas there are venues for linguists interested in statistical laws of language, we sought to establish a venue for a more interdisciplinary approach since, as we believe, fostering synergies between the methodologies of humanities and hard sciences is essential for the future development of this field. In recent years, there has been a growing pressure coming from computer science and artificial intelligence, including industrial applications, to think about some fundamental theoretical problems of language and probability together.

Probability distributions are a central concept of statistical linguistics, either in form of empirical counts or frequencies, or in form of theoretical probability measures and stochastic processes. We should be aware that discussing probability in language, we may look at very different things, such as:

- raw frequencies of single words or other units,
- empirical statistical laws, such as Zipf's law,
- more abstract principles of optimality (Zipf's forces),
- mathematical laws of probability,
- computer algorithms for text prediction or generation.

The level of abstraction we are likely to adopt is connected to our educational background, but we can learn from each other.

Probably, there are many intellectual lacunas on the way to understand probability in language and we need an organized effort of scientists of various disciplines. Statistical linguistics has been developing thanks to

contributions from different disciplines or communities: information theory, physics, computational linguistics, cognitive science, psychology and biology. Sometimes these communities provide with data, tools, or problems. Other times, they provide with solutions to statistical linguistics problems. Statistical linguistics welcomes people who are not linguists but have something creative to say about language. The challenge is that these people stay, become aware of each other, and build a solid community contributing continuously to the development of statistical linguistics.

As a result, we decided to organize an interdisciplinary workshop titled *Statistics of Languages: Theories and Experiments* in Warsaw, Poland, from 19th until 22nd July 2017. As announced, the aim of the workshop was to gather interdisciplinary researchers interested in both theoretical and experimental aspects of statistical laws of human language and similar digital systems (music, DNA, computer programs, animal communication). In particular, it was intended that:

- The workshop would foster stronger connections between empirical computational studies and theoretical mathematical models.
- The workshop would raise mutual awareness of investigations in these fields done by scholars of various sciences.

Whereas the attendance at the workshop was announced open, the talks at the workshop were drafted as short lectures by invitation. They were scheduled to last 45 minutes plus 15 minute discussion. Researchers having different backgrounds, such as linguists, mathematicians, computer scientists, physicists, and biologists were invited. To maintain a common perspective, it was requested that all talks have some computational or mathematical content. The following topics were proposed:

1. Empirical computational investigations of statistical laws of language (e.g., laws concerning the distribution of words, entropy, recurrence times).
2. Probabilistic and information-theoretic explanations of statistical laws of language, including mathematical connections among different laws.
3. Stochastic models of text generation that recover some statistical laws of language.
4. Meaningful comparison of statistical laws for various language-like systems such as human language, music, DNA, computer programs, and animal communication.

This booklet contains the abstracts of presentations submitted for the workshop. We collected 14 plenary talks from the invited speakers and 3 poster presentations from regular participants. A quick look at the table of contents confirms that the submitted titles largely conform with the call.

Organizing a new scientific event usually requires some funding. It was our great pleasure to learn that there is some interest in our initiative coming not only from academia but also from artificial intelligence industry supporting the development of related fundamental research. We wish to express our gratitude to Samsung Electronics Polska Sp. z o.o., who decided to be our sponsor and through its financial support, has helped to make this workshop possible.

Łukasz Dębowski
Ramon Ferrer-i-Cancho
Kumiko Tanaka-Ishii

Talks

A Statistical Interpretation of Linguistic Laws

Eduardo G. Altmann, Martin Gerlach

Keywords: word frequency, Zipf's law, Heaps' law

Regularities in the frequency of words in texts have long been summarized in form of linguistic laws, a cornerstone in the field of quantitative linguistics. The modern availability of large corpora renewed the interest on this subject, but it also brought new challenges due to the improved precision that is not only possible but also required in applications. In this talk I argue that linguistic laws, in their traditional formulations, are not falsifiable and have to be either re-interpreted or re-formulated. I then introduce a statistical framework to interpret, test, and apply linguistic laws.

References:

1. E. G. Altmann and M. Gerlach, *Statistical laws in linguistics*, chapter in the book Creativity and Universality in Language, M. Degli Esposti, E. G. Altmann, F. Pachet (Eds.), Lecture Notes in Morphogenesis, Springer, 2016.

2. M. Gerlach and E. G. Altmann, *Stochastic model for the vocabulary growth in natural languages*, Physical Review X 3, 021006, 2013.

3. M. Gerlach and E. G. Altmann, *Scaling laws and fluctuations in the statistics of word frequencies*, New Journal of Physics 15, 113010, 2014.

Word Entropy across more than 1000 Languages: The Linear Relationship between Unigram Entropy and Entropy Rate

Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, Ramon Ferrer-i-Cancho

Keywords: word entropy, natural language entropy, entropy rate

The entropy of words is a fundamental statistical property of natural languages. We here present results of word entropy estimations on an unprecedented sample of three massively parallel corpora—encompassing ca. 450 million words in 1916 texts and 1259 languages. We estimate both so-called unigram word entropies as well as word entropy rates. Unigram entropy can be conceptualized as the average information content of word types independent of the preceding co-text. The word entropy rate, on the other hand, is—under certain conditions to be discussed—the average information content of words taking into account the preceding co-text. Across languages of the world we find a strong linear relationship between unigram entropies and entropy rates. While unigram entropies are distributed around a mean of ca. 9 bits per word, entropy rates are distributed around a mean of 6 bits/word, while the difference is more narrowly distributed around a mean of 3 bits/word. In other words, the preceding co-text reduces the uncertainty of words by ca. 3 bits/word regardless of the language. This is unexpected, as from a linguistic point of view we would rather expect that co-occurrence patterns and syntactic dependencies vary considerably across languages, and hence also give rise to more variation in uncertainty reduction. The exact statistical and linguistic reasons for our findings are open for discussion and future research.

Language Adaptation and Change

Damian E. Blasi

Keywords: adaptation, language change, Neolithic, creole languages

The astonishing linguistic diversity displayed in the world's languages is largely driven by neutral processes that affect language transmission. In addition, it has been frequently argued that at least some of the observed diversity comes as a result of adaptive responses by languages to the varying pressures exerted on the human groups that speak or sign them. However, the question of how do languages adapt to these pressures through time is rarely discussed (if at all.) In this presentation, I will discuss two cases of different linguistic domains in which clear pressures for adaptation can be identified but where the historical outcomes tell different stories about the plasticity of language.

First, I discuss the development of a class of segments—labiodentals, in particular the fricatives "f" and "v"- that emerged from changes in diet and behaviour that started in the early Neolithic. A change in bite has lead to a reduced articulatory effort of those segments, thus paving the way to their widespread appearance across the languages of the world.

Second, I present the case of creole languages—languages that emerge from extreme contact situations, as for instance in multi-ethnic and multilinguistic slaveries. Creoles are considered to be excellent case studies where the pressure for communication "filters" unnecessary linguistic structure, thus producing particularly efficient (and novel) languages. Notwithstanding, a detailed evaluation of the available data suggests that creoles prefer to continue linguistic structure present in the languages spoken during their inception, against the expectation.

On the relationship between symmetry and structure in DNA sequences

Giampaolo Cristadoro, Mirko Degli Esposti, Eduardo G. Altmann

Keywords: DNA sequence, Chargaff symmetry, long-range correlations, transposable elements

We investigate a dynamics on symbolic sequences that mimics the action of some biological processes considered as one of the major mechanisms shaping DNA. We incorporate the relevant features emerging from such dynamics into a minimal model for genetic sequences. Our model predicts the existence of a nested hierarchy of symmetries at different structural scales. Numerical results on Homo Sapiens agree with our theoretical predictions.

From Boltzmann to Zipf through Jaynes

Alvaro Corral, Irina Espejo

Keywords: word-frequency distributions, Zipf's law, maximum-entropy principle

The word-frequency distribution provides the fundamental building blocks that generate language. It is well known, from empirical evidence, that the word-frequency distribution of any text is described by Zipf's law, approximately. Following Stephens and Bialek, we interpret the frequency of any word as arising from the interaction potential between its constituent letters. Indeed, Jaynes' maximum-entropy principle, with the constrains given by every empirical two-letter marginal, leads to a Boltzmann distribution of word frequencies, with an energy-like function given by the sum of all pairwise potentials. The iterative-scaling algorithm allows finding the potentials from the empirical two-letter marginals. Using this formalism, we explore the ability of several variations of these models to reproduce Zipf's law. In this way, a pure statistical-physics framework is used to describe word-frequency distributions.

References:

1. G. J. Stephens and W. Bialek, *Statistical mechanics of letters in words*, Physical Review E, 81, 066119, 2010.

Natural Language and Strong Nonergodicity

Łukasz Dębowski

Keywords: nonergodic processes, mutual information, PPM code

A stationary stochastic process is called ergodic when the relative frequencies of any events in the random text generated by the process converge to the probabilities of the events with probability one. In contrast, the hypothetical stochastic process responsible for generation of texts in natural language has been often supposed to be nonergodic, i.e., not ergodic. In the talk, we will argue that the process of natural language generation probably satisfies a stronger property, which we call strong nonergodicity. Intuitively, nonergodicity corresponds to existence of a random persistent topic in the random text, whereas strong nonergodicity occurs when we need an infinite sequence of random bits, called random facts, to describe this topic completely. We will show that strong nonergodicity of a stochastic process is a partly falsifiable property. Namely, we will exhibit a surprising assertion, which we call the theorem about facts and words. This proposition states that the number of random facts that can be inferred from a text generated by a stationary process must be roughly smaller than the number of word-like strings detected in this text by the standard PPM compression algorithm. Since the number of the word-like strings for texts in natural language follows an empirical stepwise power law, we may suppose that the number of inferrable facts also follows a power law. That is, natural language looks as if it were strongly nonergodic.

References:

1. Ł. Dębowski, *Is Natural Language Strongly Nonergodic? A Stronger Theorem about Facts and Words*, `http://arxiv.org/abs/1706.04432`, 2017.

Compression as a General Principle of Language

Ramon Ferrer-i-Cancho

Keywords: compression, linguistic laws, information theory, uniform information density, sociology of science

The principle of compression, namely, the minimization of the mean length of types under certain constraints (e.g., unique decipherability) is a highly predictive principle. First, it is able to shed light on the origins of various linguistic laws: Zipf's law of abbreviation, Menzerath's law and Zipf's law for word frequencies. Interestingly, compression is a side-effect of the general principle of distance minimization. Second, compression may also predict the phenomenon of reduction, namely, the tendency of types that appear in more predictive contexts to be shorter. Reduction is one of the main justifications for the uniform information density hypothesis and related approaches, that are regarded as the reference information theory of language specially by US researchers. Despite their many flaws, these approaches have enjoyed a tremendous success, that is easy to understand if sociological factors are taken into account.

References:

1. R. Ferrer-i-Cancho, C. Bentz and C. Seguin, *Compression and the origins of Zipf's law of abbreviation*, `http://arxiv.org/abs/1504.04884`, 2015.

2. M. L. Gustison, S. Semple, R. Ferrer-i-Cancho and T. J. Bergman, *Gelada vocal sequences follow Menzerath's linguistic law*, Proceedings of the National Academy of Sciences USA 113 (19), E2750-E2758, 2016.

3. R. Ferrer-i-Cancho, *Optimization models of natural communication*, Journal of Quantitative Linguistics, 2017.

4. R. Ferrer-i-Cancho, *The placement of the head that maximizes predictability. An information theoretic approach*, 2017.

Zipf's Law, Heaps' Law, and the Project Gutenberg Corpus

Francesc Font-Clos

Keywords: Zipf's law, Heaps' law, text length, fitting methods

I will give an overview of my work on statistics of languages: Zipf's law and Heaps' law, their relation, their dependence on text length and methods to rigorously test them. We will first see that, despite some claims, the functional form of the distribution of frequencies remains essentially unchanged when we take subsets of decreasing length of a given text (in physicists terms, it obeys a scaling law). This is true even if the distribution of frequencies is not a power law, but we will give some attention to the power-law case and show that Zipf's exponent cannot change with text length. Using proper fitting methods, we will prove our claim and demonstrate how biased fitting methods can create the illusion of length-dependent exponents. The relation between Zipf's and Heaps' exponents has been rediscovered so many times that it is usually taken for granted, but most rely on asymptotic limits or other approximations. As an alternative, I will present an exact derivation of Heaps' law where Zipf's law is assumed to be a power law. Surprisingly, the obtained form of Heaps' law is not a power law, but rather a convex curve in log-log space. We will see that the vocabulary growth of real books is well approximated by such form. Finally I will present some research on the Project Gutenberg corpus: how many books do really fulfill Zipf's law in its simplest form? Are shorter books more likely to pass fitting tests? I will try to give some answers

From Universality to Variability in the Statistics of Word Frequencies using Generalized Entropies

Martin Gerlach, Eduardo G. Altmann

Keywords: information theory, linguistic laws, universality, entropy, language change

The statistical analysis of texts has shown the existence of numerous linguistic laws, the most famous being Zipf's law. While the latter are usually formulated for the average behavior of an observable, topical variability or language change over time lead to much more intricate patterns when investigating the deviations from universality, e.g. much larger fluctuations around the average than expected. In this talk, I will address the question about the co-existence of universality and variability by proposing Information-theoretic measures based on generalized entropies to quantify the similarity between two instances of text. I will focus on two aspects involving the generic appearance of heavy-tailed distribution of word frequencies. First, the generalized entropies allow for a fine-tuning of which words in the frequency-spectrum contribute to the measured distance leading to an improved interpretation of the respective measures. Second, I will point out the caveats when estimating distances in (necessarily) finite samples of data caused by the large number of low-frequency symbols. I will show the practical relevance of these findings in: i) quantifying the evolution of the English language over the last two centuries in millions of digitized books and ii) investigating the organization and evolution of scientific fields in more than 21M scientific papers over the past three decades.

On Infinite Divisibility and the Large Number of Rare Events, in Texts and Elsewhere

Estate Khmaladze

Keywords: Levy measures, Zipf's Law, sparse data

Consider a sample of frequencies $\{\nu_{in}\}_{i=1}^{N}$ of N disjoint events, like occurrences of different words or sighting of different species in n independent observations. The specific property of the samples we have in mind is that both N and n are very large. In modern terminology, our observations are "sparse". Let us describe the family of these frequencies as a statistical ensemble, that is, study them through the measure

$$L_{Nn}(z) = \sum_{i=1}^{N} \mathbb{I}(\nu_{in} \geq z).$$

Without any claim that the probabilities $\{p_{in}\}_{i=1}^{N}$ of our events are random, we also consider them as a statistical ensemble and study the measure

$$K_{Nn}(z) = \sum_{i=1}^{N} \mathbb{I}(np_{in} \geq z).$$

What we want to understand is why the situation with very large N and n is so ubiquitous and why the measure L_{Nn} is so often stable and follows certain patterns, in spite of being based on very unstable and chaotic frequencies. Even more than this, we want to understand why K_{Nn} becomes stable and converges to a limit, and what the nature of this limit is.

Our main statement will be that the K_{Nn} converge to Lévy measures of infinitely divisible distributions. We will try to justify this statement, and if we accept it, this will broaden the theory of a large number of rare event much beyond Zipf's or Karlin-Rouault's, or other famous Laws, used to describe the asymptotic behaviour of L_{Nn}.

References:

1. E. Khmaladze, *Convergence properties in certain occupancy problems including the Karlin–Rouault Law*, Journal of Applied Probability, 48, 1095-1113, 2011.

Mommy Only Really Cares about My Grammar! Child-mother Dyads Integrate a Dynamical System

Fermin Moscoso del Prado

Keywords: causality, child language

Both the language used by young children (child language; CL) and the simplified language used by caretakers when talking to them (child-directed speech; CDS) become increasingly complex along development, eventually approaching regular adult language. Researchers disagree on whether children learn grammar from the input they receive (usage-based theories), or grammars are mostly innate, requiring only minimal input-based adjustments on the part of the children (nativist theories). A related question is whether parents adapt the complexity of CDS in specific response to their children's language abilities (fine-tuning), or only in response to their level of general cognitive development. Previous research suggests that parent-child interactions can be modelled by non-linear dynamical systems. Following this direction, I adapt a technique recently developed in Ecology—Convergent Cross-Mapping (CCM)—for assessing causal relations between the longitudinal co-development of aspects of CL and CDS. CCM enables reconstructing a network of causal relations involving aspects of CL and CDS. This network supports a mutual bootstrapping between lexical and grammatical aspects of CL. In addition, the network reveals explicit couplings between the language used by individual children and their mothers. This provides explicit evidence for fine-tuning: Mothers adapt CDS in response to the specific grammatical properties of CL (but apparently not its lexical properties). Our findings verify the strong causal predictions of usage-based theories, and are difficult for nativist theories to account for.

Do We Need Deep Learning for Recognizing Phrases in the Natural Language?

Urszula Oleszek

Keywords: deep learning, sequence labelling, conditional random fields, named entity recognition, sentiment recognition

During the lecture I would like to present various algorithms used in the sequence labelling task. They include classical approaches (such as Conditional Random Fields) and multiple types of recurrent deep neural networks. I will discuss their performance on natural language datasets and the problems of Named Entity and sentiment recognition.

Linguistic Laws in Non-human Primate Communication

Stuart Semple

Keywords: linguistic laws, primates, vocalisation, compression

My collaborators and I have conducted a number of studies testing whether the statistical laws of human language—known as linguistic laws—are also seen in the communication systems of non-human primates. We found that in Formosan macaques (a monkey native to Taiwan), calls in the vocal repertoire follow Zipf's law of abbreviation, which predicts a negative relationship between signal length and frequency of use. Subsequently, we demonstrated that in geladas (a monkey from the highlands of Ethiopia), vocal sequences follow Menzerath's law, according to which longer sequences are made up of shorter constituents. We mathematically linked both of these laws to compression, the information theoretic principle of minimising code length, and argued that this phenomenon underpins diverse biological information systems, reflecting a universal pressure for coding efficiency. In new work, we have conducted the first test of linguistic laws in non-human primate gestural communication. Analysing the play gestures of wild chimpanzees, we found a strong negative relationship between number of gestures in a sequence and mean duration of the constituent gestures (in line with Menzerath's law), but no relationship between gesture duration and frequency of use (contrary to Zipf's law of abbreviation). However, analysis of specific subsets of the overall gestural repertoire did reveal strong agreement with Zip's law of abbreviation, demonstrating that patterns consistent with this law were hidden when the entire repertoire was analysed. Overall, these studies highlight the value of exploring the full scope of linguistic laws outside the realm of human language.

The Zipf Law of Abbreviation from a Statistical Point of View

Isabel Serra, Christian Bentz, Alvaro Corral, Ramon Ferrer-i-Cancho

Keywords: power law, bivariate extreme value analysis

This work consists in the analysis in the frame of complex systems of the compression and the origins of Zipf's law of abbreviation. Languages across the world exhibit Zipf's law of abbreviation, namely more frequent words tend to be shorter. The current work analyze the apparent universality of this pattern in human language. We analyze the Bible (in 1400 different languages) and The Human Right Treaty (in 400 different languages). The main approach consists to consider the concept of mean code length as a mean energetic cost function over the probability and the magnitude of the types of the repertoire.

Generative Models Producing Power Laws of Language

Kumiko Tanaka-Ishii

Keywords: generative models, power law, long-range correlation

In this talk, I discuss a possible property of language through some generative models that produce power laws. First, I review power laws that hold for natural language, namely, Zipf's law, Heaps' law, and long-range correlation. For this last one, the long-range correlation, I introduce a state-of-the-art analysis method. Then, I reconsider how these power laws hold for various kinds of data, from literary texts, infants' utterances, and music. To investigate why such power laws should hold, I review some generative models highlighting a common property that human-generated data might possess.

Posters

A minimal model of text generation: random walk on structured scale free network

Dániel Czégel, Maxi San Miguel

Keywords: text generation, complex networks, stochastic processes

The large amount of digitized linguistic data opens up the unique possibility of using the methodology of complex systems to understand high-level human cognitive processes. Two such issues are i) the way we categorize the continuous space of real-world features into discrete concepts, and ii) the way we use language to copy a line a thought from one brain to another. In this work I address both questions by formulating a simple text generation model which reproduces the three major characteristic large-scale statistical laws of human language streams, namely Zipf's law, Heaps' law and Burstiness. Furthermore, the generation itself can be described as a random walk on a scale-free, highly clustered and low dimensional complex network, suggesting that this class of networks is appropriate as a minimal model of the semantic space. Entangling the global characteristics of the semantic space is an inevitable step towards analyzing texts as trajectories in such a space, with promising applications such as author or style identification, personal disorder diagnosis, or the evolution of cultural traits mirrored by text production characteristics.

What determines borrowability? A quantitative cross-linguistic study

Ezequiel Koile, Maya Inbar, Damian Blasi, Eitan Grossman

Keywords: loanwords, cross-linguistic borrowing, contact linguistics

Lexical borrowing plays an important role in the evolution of languages, since each contact situation among different speech communities can potentially lead to change in the languages involved. Our goal is to estimate the contribution of different factors which might affect the process of borrowing of lexical items.

The World Loanword Database (WOLD, Haspelmath and Tadmor 2009) is a cross-linguistically comparable set of 1460 meanings and their lexicalizations in 41 languages from all over the world. This project has generated a massive dataset particularly useful for a study aiming to tease apart the determining factors of lexical borrowing using quantitative methods. While the authors of WOLD have presented preliminary findings regarding universals of lexical borrowing (Tadmor 2009), any claim on the magnitude and the extent of the relevant factors are not feasible without proper statistical methods.

This work addresses this lacuna by conducting a statistical analysis of the WOLD database, with the aim of determining which factors define the lexemes to be borrowed. By implementing a mixed-effects model, we introduce the contribution of parameters such as the morphosyntactic properties the languages involved, the semantic field of each lexical item, and different sociolinguistic aspects in the borrowability of a given lexical item. On the basis of the model we draw a clearer picture of the dependency between factors, with the aim of accounting for the observed variation in borrowability and borrowing patterns across languages.

References:

1. M. Haspelmath and U. Tadmor (eds.). 2009. World Loanword Database. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at `http://wold.clld.org`, Accessed on 2016-02-15.)

2. M. Haspelmath and U. Tadmor. 2009. The loanword typology project and the world loanword database. In: M. Haspelmath and U. Tadmor (eds.), Loanwords in the world's languages: A comparative handbook, 1-34. Berlin: De Gruyter Mouton.

3. M. Haspelmath. 2009. Lexical borrowing: Concepts and issues. In: M. Haspelmath and U. Tadmor (eds.), Loanwords in the worlds languages: A comparative handbook, 35-54. Berlin: De Gruyter Mouton.

4. U. Tadmor. 2009. Loanwords in the worlds languages: Findings and results. In: M. Haspelmath and U. Tadmor (eds.), Loanwords in the worlds languages: A comparative handbook, 55-75. Berlin: De Gruyter Mouton.

Impact of lexical and sentiment factors on the popularity of scientific papers

Julian Sienkiewicz and Eduardo G. Altmann

Keywords: citation analysis, sentiment analysis, quantile regression

The number of citations an article receives can be considered a proxy for the attention or popularity the article achieved in the scientific community. Citations play a crucial role both in the evolution of science and in the bibliometric evaluation of scientists and institutions, in which case the number of citations is often tacitly taken as a measure of quality. We investigate how textual properties of the title and abstract of scientific papers relate to the number of citations they receive ten to twenty years after publication. Our main finding is that correlations are non-linear and the impact of factors on the most cited papers is typically very different from the impact on typical papers (in terms of number of citations). For instance, many previous works report that short titles correlate with citations. Here we find that this result holds only for the most cited papers, for a typical paper the correlation is reversed. We looked additionally on measures of the vocabulary complexity and on sentiment properties and found a strong relation between citations and the valence in the abstract.

References:

1. J. Sienkiewicz and E. G. Altmann, *Impact of lexical and sentiment factors on the popularity of scientific papers*, Royal Society Open Science, 3 (6), 160140, 2016, `http://rsos.royalsocietypublishing.org/content/3/6/160140`

List of participants

- Eduardo G. Altmann (University of Sydney)
  eduardo.altmann@sydney.edu.au
- Christian Bentz (University of Tübingen)
  chris@christianbentz.de
- Damian E. Blasi (University of Zurich/Max Planck Institute for the Science of Human History)
  damian.blasi@uzh.ch
- Giampaolo Cristadoro (University of Bologna)
  giampaolo.cristadoro@unibo.it
- Alvaro Corral (Centre de Recerca Matematica)
  acorral@crm.cat
- Dániel Czégel (MTA Centre for Ecological Research)
  czegel_d@yahoo.com
- Łukasz Dębowski (Polish Academy of Sciences)
  ldebowsk@ipipan.waw.pl
- Leandro Ezequiel Koile (Max Planck Institute for the Science of Human History)
  ezequielk@gmail.com
- Ramon Ferrer-i-Cancho (Universitat Politecnica de Catalunya)
  rferrericancho@cs.upc.edu
- Francesc Font-Clos (ISI Foundation)
  francesc.font@gmail.com

- Martin Gerlach (Northwestern University)
  gerlach.m@web.de
- Estate Khmaladze (Victoria University of Wellington)
  Estate.Khmaladze@vuw.ac.nz
- Fermin Moscoso del Prado (University of California Santa Barbara)
  fermosc@gmail.com
- Urszula Oleszek (Samsung Poland R&D Center)
  u.oleszek@samsung.com
- Stuart Semple (Univeristy of Roehampton)
  s.semple@roehampton.ac.uk
- Isabel Serra (Centre de Recerca Matemàtica)
  iserra@crm.cat
- Julian Sienkiewicz (Warsaw University of Technology)
  julasms@gmail.com
- Kumiko Tanaka-Ishii (University of Tokyo)
  kumiko@cl.rcast.u-tokyo.ac.jp
- Thomas Alexander Trost (Saarland University)
  thomas.trost@lsv.uni-saarland.de