SOE 8: Focus Session: Complex Systems Approaches to Language and Communication

In this session language and communication are investigated using ideas from statistical physics, time series analysis, and complex systems. New opportunities for quantitative studies in this field come both from the availability of large databases of human communication and from the richness of theoretical models developed in Physics. (Focus Session organized by Eduardo Altmann, Dresden)

Time: Tuesday 10:15–13:00 Location: GÖR 226

SOE 8.1 Tue 10:15 GÖR 226

A Comparative Study of Language Complexity in Wikipedia — •János Kertész[1,2], Taha Yasseri[2,3], and András Kornai[2,4] — [1]Central European University — [2]Budapest University of Technology and Economics — [3]University of Oxford — [4]Computer and Automation Research Institute of the Hungarian Academy of Sciences.

We present statistical analysis of English texts from Wikipedia [1]. We try to address the issue of language complexity empirically by comparing the Simple English Wikipedia (Simple) to comparable samples of the main English Wikipedia (Main). Simple is supposed to use a more simplified language with a limited vocabulary, and editors are explicitly requested to follow this guideline, yet in practice the vocabulary richness of both samples are at the same level. Detailed analysis of longer units (n-grams of words and part of speech tags) shows that the language of Simple is less complex than that of Main primarily due to the use of shorter sentences, as opposed to drastically simplified syntax or vocabulary. Comparing the two language varieties by the Gunning readability index supports this conclusion. We also report on the topical dependence of language complexity, that is, that the language is more advanced in conceptual articles compared to person-based (biographical) and object-based articles. Finally, we investigate the relation between conflict and language complexity by analyzing the content of the talk pages associated to controversial and peacefully developing articles, concluding that controversy has the effect of reducing language complexity.

[1] Yasseri T, Kornai A, Kertész J (2012) PLoS ONE 7(11): e48386.

SOE 8.2 Tue 10:45 GÖR 226

Statistical Mechanics of Human Language — •Kosmas Kosmidis — Computational Systems Biology Group, Jacobs University,Bremen — Computational Physics Group, Aristotle University of Thessaloniki,Thessaloniki,Greece

We use the formulation of equilibrium statistical mechanics in order to study some important characteristics of language. Using a simple expression for the Hamiltonian of a language system, which is directly implied by the Zipf law, we are able to explain several characteristic features of human language that seem completely unrelated, such as the universality of the Zipf exponent, the vocabulary size of children, the reduced communication abilities of people suffering from schizophrenia, etc. While several explanations are necessarily only qualitative at this stage, we have, nevertheless, been able to derive a formula for the vocabulary size of children as a function of age, which agrees rather well with experimental data.

SOE 8.3 Tue 11:00 GÖR 226

Topic models and scaling laws — •Martin Gerlach and Eduardo G. Altmann — Max Planck Institute for the Physics of Complex Systems, Dresden, Germany

In this talk we combine statistical analysis of large text databases and simple stochastic models to explain the appearance of scaling laws in the statistics of word frequencies. We focus on the well studied case of the vocabulary growth with database size (Heaps' law) and on a novel scaling law we observe using fluctuation scaling analysis. In order to simultaneously explain both scaling laws we show that it is essential to account for the heterogeneity in the vocabulary of texts by considering topic models (e.g. Latent Dirichlet Allocation). Our models are tested against three different databases: Google n-gram database, Wikipedia, and all articles published by PLoS.

SOE 8.4 Tue 11:15 GÖR 226

Reading Stockholm Riots 2013 in social media by text-mining — •Andrzej Jarynowski[1,2] and Amir Rostami[1] — [1]Department of Sociology, Stockholm University, Sweden — [2]Smoluchowski Institute of Physics, Jagiellonian University, Cracow, Poland

The riots in Stockholm in May 2013 were an event that reverberated in the world media for its dimension of violence that had spread through the Swedish capital. In this study we have investigated the role of social media in creating media phenomena via text mining and natural language processing. We have focused on two channels of communication for our analysis: Twitter and Poloniainfo.se (Forum of Polish community in Sweden). Our preliminary results show some hot topics driving discussion related mostly to Swedish Police and Swedish Politics by counting word usage. Typical features for media intervention are presented. We have built networks of most popular phrases, clustered by categories (geography, media institution, etc.). Sentiment analysis shows negative connotation with Police. The aim of this preliminary exploratory quantitative study was to generate questions and hypotheses, which we could carefully follow by deeper more qualitative methods.

15 min. break

SOE 8.5 Tue 11:45 GÖR 226

Agent-based models of consensus in language dynamics — •Martina Pugliese[1], Christine Cuskley[2], Claudio Castellano[2], Francesca Colaiori[2], Vittorio Loreto[1,3], and Francesca Tria[3] — [1]Physics Department, Sapienza University, Rome, Italy — [2]Institute for Complex Systems (ISC-CNR), Rome, Italy — [3]Institute for Scientific Interchange (ISI), Turin, Italy

The emergence of consensus in language dynamics can be studied in the framework known as the Naming Game (NG), where agents engage in pairwise interactions about naming an object and achieve a stable vocabulary, with a convergence-time depending on the network used and on the population replacement rate.

We have implemented a NG-like model investigating morphology where the focus is given on a fixed set of topic words chosen with an *a priori* frequency distribution. We examine processes of regularisation from two different perspectives: memory and development. Memory limitations are dictated by an expanding time window within which a word must be encountered to be recalled by an agent, otherwise the agent “forgets” the inflection for that word and falls back on the regular rule. The second strategy examines population turnover as a probabilistic replacement: if chosen as a speaker, a new agent defaults to the regular form.

Both strategies result in a frequency-based transition of regularity: high frequency words stabilise indefinitely in the irregular state, while those at low frequency regularise. With a mean-field approach we show that the transition is discontinuous.

SOE 8.6 Tue 12:15 GÖR 226

Spatial language dynamics in Northern Italy before standardization — •Gero Vogl[1], Michael Leitner[2], and Paul Videsott[3] — [1]Fakultät für Physik, Universität Wien, Austria — [2]Heinz Maier-Leibnitz Zentrum (MLZ), Technische Universität München, Germany — [3]Freie Universität Bozen, Italy

In the year 1200 every city in Northern Italy wrote its derivative from Latin in its own way. We have followed the change from the local to standard language over time (1200 to 1525) and space. Documents from 36 cities with more than 500.000 words have been scanned for 300 specific traits. We constructed similarity matrices and found: (a) there is certain continuity in one and the same city, i.e. similarity of written language over the time is greater than similarity to neighbour cities and even more so than similarity with the average North Italian city. (b) in the period around 1500, in cities along the Via Emilia the trend to standard Italian is strong, whereas in the North of the region towards the Alps local traits persist.

SOE 8.7 Tue 12:30 GÖR 226

Endogenous and exogenous effects on the adoption curves of linguistic innovations — •Fakhteh Ghanbarnejad, Martin Gerlach, Jose M. Miotto, and Eduardo G. Altmann — Max Planck Institute for the Physics of Complex Systems, Dresden,Germany

It is well accepted that adoption of innovations are described by S-curves (slow start, accelerating period, and slow end). In this talk we search for a quantitative description of the curve of total number of adopters as a function of time and we analyze how much information on the dynamics of innovation spreading can be obtained from them. We are particularly interested in the case of linguistic innovations because detailed databases of written texts from the last 200 years allow for an unprecedent statistical precision. Combining data analysis with simulations of simple models (e.g., the Bass dynamics on networks) we identify signatures of endogenous and exogenous factors in the adoption curves and we propose a method to quantify the strength of these factors from data. We obtain that in cases in which the exogenous factors are dominant (e.g., in the 1901 and 1996 orthographic reforms of German) the adoption curve is better described by an exponential than by an S-curve.

SOE 8.8 Tue 12:45 GÖR 226

A Coarse Grained Approach for Distinguishing Whale "Dialects" — •Sarah Hallerberg[1], Heike Vester[2], Kurt Hammerschmidt[3], and Marc Timme[1,4,5] — [1]Network Dynamics, Max Planck Institute for Dynamics and Self-Organization (MPIDS) — [2]Ocean Sounds, Henningsvaer, Norway — [3]Research Group Cognitive Ethology Lab, German Primate Center, Göttingen — [4]2Faculty of Physics, University of Göttingen, Göttingen, Germany — [5]Bernstein Center for Computational Neuroscience, Göttingen, Germany

Complex vocal communication simultaneously requires high cognitive abilities, a large flexibility in sound production, and advanced social interactions. Social whales, such as killer whales and pilot whales, fulfill all of these requirements. How their acoustic signals are used and how the acoustic patterns are organized, however, is largely unknown. Up to date, mostly human observers classify acoustic patterns through hearing and visual comparison of spectrograms. We decided to use a data analysis approach and study distributions of acoustic features (in particular, cepstral coefficients) generated from ensembles of pilot whale vocalizations. Comparing these distributions by computing Kullback-Leibler-divergences we find substantially different distributions for sounds produced by different groups of pilot whales.