CL4LC 2016

Computational Linguistics for Linguistic Complexity

Proceedings of the Workshop

December 11, 2016
Osaka, Japan

ISBN 978-4-87974-709-9

Preface

Welcome to the first edition of the "Computational Linguistics for Linguistic Complexity" workshop (CL4LC)! CL4LC aims at investigating "processing" aspects of linguistic complexity with the objective of promoting a common reflection on approaches for the detection, evaluation and modelling of linguistic complexity.

What has motivated such a focus on linguistic complexity? Although the topic of linguistic complexity has attracted researchers for quite some time, this concept is still poorly defined and often used with different meanings. Linguistic complexity indeed is inherently a multidimensional concept that must be approached from various perspectives, ranging from natural language processing (NLP), second language acquisition (SLA), psycholinguistics and cognitive science, as well as contrastive linguistics. In 2015, a one-day workshop dedicated to the question of *Measuring Linguistic Complexity* was organized at the catholic University of Louvain (UCL) with the aim of identifying convergent approaches in diverse fields addressing linguistic complexity from their specific viewpoint. Not only did the workshop turn out to be a great success, but it strikingly pointed out that more in–depth thought is required in order to investigate how research on linguistic complexity and its processing aspects could actually benefit from the sharing of definitions, methodologies and techniques developed from different perspectives.

CL4LC stems from these reflections and would like to go a step further towards a more multifaceted view of linguistic complexity. In particular, the workshop would like to investigate processing aspects of linguistic complexity both from a machine point of view and from the perspective of the human subject in order to pinpoint possible differences and commonalities.

We are glad to see that our expectations have been met since the workshop has generated great enthusiasm both within the Program Committee, whose members from various disciplines have wholeheartedly agreed to serve, and within authors, as we received 33 paper submissions in all, out of which eight were selected as oral presentations and seventeen as posters.

The multidisciplinary approach assumed by the workshop is reflected in the submissions that we received. We can classify them following one major "theoretical" distinction between absolute complexity (i.e. the formal properties of linguistic systems) and relative complexity (i.e. covering issues such as cognitive cost, difficulty, level of demand for a user/learner). Several papers that we received focused on language complexity per se, which is typically addressed comparing the structural complexity of different languages. Bentz et al. (a) thus compare typical measures of language complexity across 519 languages. In Bentz et al. (b), they also discuss language evolution, with a special focus on morphological complexity, in the light of learning pressures. Another approach is to assess document complexity and Chen and Meurers propose a web-based tool to this aim. Papers focusing deeper on a specific aspect of linguistic complexity were also proposed, such as Zaghouani et al., who report a method to detect lexical ambiguity, which is one of the major sources of language complexity, and on its impact on human annotation. Bjerva and Börstell investigate the impact of morphological complexity and animacy features on the order of verb and object in Swedish Sign Language. Takahira et al. adopt a broader approach and compare the entropy rates of six different languages by means of a state-of-the-art compression method. Finally, Shi et al. investigate the impact of polysemy on automatic word representation in order to improve the performance of word embeddings.

We also got very interesting papers on the relative complexity of language, i.e. the difficulty perceived by humans when processing linguistic input. Some of them are concerned with modeling human sentence processing through experimental and computational metrics to capture linguistic clues of sentence processing difficulty. Other papers address relative complexity from a more applicative point of view. The first is the case of van Schijndel and Schuler, who revisit the question of using eye-tracking data to predict the level of surprisal of sentences, showing that taking into consideration a word skipped during reading improves n-gram surprisal, but not surprisal measures based on PCFG. The work of Bloem also uses the construct of surprisal to investigate the relation between processing cost and the choice between near-synonymous verbal constructions in Dutch. The contribution of Shain et al. investigates the existence of a latency effect during sentence processing due to memory access. On their side, Li et al. regress various measures of textual complexity on fMRI timecourses while listening to a story to discuss the role of various regions of interest (ROI) in the humain brain. Chersoni et al. propose a very relevant contribution suggesting a Distributional Model for computing semantic complexity that is based on the MUC (Memory, Unification and Control) model for sentence comprehension. Heilmann and Neumann explore a completely different horizon and make use of keylogs to better model language complexity and the cognitive load it produces during the translation process. Finally, Becerra-Bonache and Jimenez-Lopez adopt a developmental approach of linguistic complexity that uses grammatical inference algorithms to simulate the acquisition of language by a native speaker.

The second more applicative point of view to relative complexity is addressed by Falkenjack and Jönsson, who are concerned with the scarcity of available texts for training readability models in languages other than English and suggest using a Bayesian Probit model coupled with a ranking classification strategy. Ströbel et al. suggest another approach to text readability based on a sliding-window that is used to create the distribution of linguistic complexity for a text. Wagner Filho et al. compare the efficiency of various machine learning algorithms and engineered features to automatically build a large corpus for readability assessment. Vajjala et al. provide an interesting example of a integrated view of text readability that correlates text characteristics with reader's language ability reflected in reading comprehension experiments. The paper by Deep Singh et al. also addresses readability prediction from a psycholinguistic point of view, using eye-tracking measures instead of grade level to train their model. Gonzalez-Dios et al. carry out an interesting in-depth analysis combining readability and text simplification, retaining the most predictive syntactic structures from a readability model and analysing how human writers simplify the syntactic structures concerned. Similarly, with the goal of connecting text simplification practises with real needs of readers, Gala et al. experiment with dyslexic children to verify the effects of lexical simplification on reading comprehension. Albertsson et al. on their part detect paraphrased segments between two corpora (one comprised of simple texts, while the other includes more advanced materials) for text simplification purposes. Finally, Pilan et al. use coursebook-based lists of vocabulary to improve the proficiency prediction of learner essays in Swedish.

A further perspective is assumed by those papers more focused on linguistic complexity from the automatic processing point of view, investigating differences and similarities with human sentence processing. This is the case of Delmonte's paper, which is concerned with syntactic complexity for a syntactic parser focusing in particular on those syntactic structures which are known to be difficult for human parsing. Mirzaei et al. use errors made by an automatic speech recognition system as indicators of second language learners's listening difficulties.

To conclude this nice programme, we wish to thank everyone who submitted a paper, all of the authors for their contributions, the members of the Program Committee for their thoughtful reviews, and everyone who attended this workshop for sharing time and thoughts on this increasingly important research topic.

Sincerely,

Dominique Brunato
Felice Dell'Orletta
Giulia Venturi
Thomas François
Philippe Blache

Organisers

Dominique Brunato, ItaliaNLP Lab, Istituto di Linguistica Computazionale "A. Zampolli", Pisa, Italy
Felice Dell'Orletta, ItaliaNLP Lab, Istituto di Linguistica Computazionale "A. Zampolli", Pisa, Italy
Giulia Venturi, ItaliaNLP Lab, Istituto di Linguistica Computazionale "A. Zampolli", Pisa, Italy
Thomas François, CENTAL, IL&C Université catholique de Louvain, Louvain-la-Neuve, Belgium
Philippe Blache, Laboratoire Parole et Langage, CNRS & Université de Provence, Aix-en-Provence, France

Programme Committee

Mohammed Attia (George Washington University, USA)
Delphine Bernhard (LilPa, Université de Strasbourg, France)
Joachim Bingel (University of Copenhagen, Denmark)
Nicoletta Calzolari (Istituto di Linguistica Computazionale "A. Zampolli", ILC-CNR, Italy - European Language Resources Association (ELRA), France)
Angelo Cangelosi, (Centre for Robotics and Neural Systems at the University of Plymouth, UK)
Benoît Crabbé (Université Paris 7, INRIA, France)
Scott Crossley (Georgia State University, USA)
Rodolfo Delmonte (Department of Computer Science, Università Ca' Foscari, Italy)
Piet Desmet (KULeuven, Belgium)
Arantza Díaz de Ilarraza (IXA NLP Group, University of the Basque Country)
Cédrick Fairon (Université catholique de Louvain, Belgium)
Marcello Ferro (Istituto di Linguistica Computazionale "A. Zampolli", ILC-CNR, Italy)
Nuria Gala (Aix-Marseille Université, France)
Ted Gibson (MIT, USA)
Itziar Gonzalez-Dios (IXA NLP Group, University of the Basque Country)
Alex Housen (Vrije Universiteit Brussel, Belgium)
Frank Keller (University of Edinburgh, UK)
Kristopher Kyle (Georgia State University, USA)
Alessandro Lenci (Università di Pisa, Italy)
Annie Louis (University of Essex, UK)
Xiaofei Lu (Pennsylvania State University, USA)
Shervin Malmasi (Harvard Medical School)
Ryan Mcdonald (Google Inc.)
Detmar Meurers (University of Tübingen, Germany)
Simonetta Montemagni (Istituto di Linguistica Computazionale "A. Zampolli", ILC-CNR, Italy)
Alexis Neme (Université Paris-Est, France)
Frederick J. Newmeyer (University of Washington, USA, University of British Columbia, Simon Fraser University, CA)
Joakim Nivre (Uppsala University, Sweden)
Gabriele Pallotti (Università di Modena e Reggio Emilia, Italy)
Magali Paquot (Université catholique de Louvain, Belgium)
Vito Pirrelli (Istituto di Linguistica Computazionale "A. Zampolli", ILC-CNR, Italy)

Barbara Plank (University of Groningen, Netherlands)
Massimo Poesio (University of Essex, UK)
Horacio Saggion (Universitat Pompeu Fabra, Spain)
Paul Smolensky (John Hopkins University, USA)
Kaori Sugiyama (Department of Literature, Seinan Gakuin University, Japan)
Benedikt Szmrecsanyi (KULeuven, Belgium)
Kumiko Tanaka-Ishii (University of Tokyo, Japan)
Joel Tetreault (Grammarly Inc.)
Sara Tonelli (FBK, Trento, Italy)
Sowmya Vajjala (Iowa State University, USA)
Aline Villavicencio (Institute of Informatics Federal University of Rio Grande do Sul, Brazil)
Elena Volodina (University of Gothenburg, Sweden)
Patrick Watrin (CENTAL, IL&C, Université catholique de Louvain, Louvain-la-Neuve, Belgium)
Daniel Wiechmann (University of Amsterdam, Netherlands)
Victoria Yaneva (University of Wolverhampton, UK)
Leonardo Zilio (CENTAL, IL&C, Université catholique de Louvain, Louvain-la-Neuve, Belgium)

Table of Contents

Conference Program

Sunday December 11, 2016 (Room 1002)

09:00–09:15 **Opening Remarks**

09:15–10:30 **Session 1 – Oral presentations**

09:15–09:40 *Could Machine Learning Shed Light on Natural Language Complexity?*
Maria Dolores Jimenez Lopez and Leonor Becerra-Bonache

09.40–10.05 *Towards a Distributional Model of Semantic Complexity*
Emmanuele Chersoni, Philippe Blache and Alessandro Lenci

10.05–10.30 *CoCoGen - Complexity Contour Generator: Automatic Assessment of Linguistic Complexity Using a Sliding-Window Technique*
Ströbel Marcus, Elma Kerz, Daniel Wiechmann and Stella Neumann

10:30–10:50 **Break**

10:50–12:05 **Session 2 – Oral presentations**

10:50–11:15 *Addressing surprisal deficiencies in reading time models*
Marten van Schijndel and William Schuler

11:15–11:40 *Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts*
Sowmya Vajjala, Detmar Meurers, Alexander Eitel and Katharina Scheiter

11:40–12:05 *Memory access during incremental sentence processing causes reading time latency*
Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson and William Schuler

12:05–14:00 **Lunch Break**

14:00–15:20 **Poster session**

Reducing lexical complexity as a tool to increase text accessibility for children with dyslexia
Nuria Gala and Johannes Ziegler

A Preliminary Study of Statistically Predictive Syntactic Complexity Features and Manual Simplifications in Basque
Itziar Gonzalez-Dios, María Jesús Aranzabe and Arantza Díaz de Ilarraza

Dynamic pause assessment of keystroke logged data for the detection of complexity in translation and monolingual text production
Arndt Heilmann and Stella Neumann

Sunday December 11, 2016 (continued)

Implicit readability ranking using the latent variable of a Bayesian Probit model
Johan Falkenjack and Arne Jonsson

CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis
Xiaobin Chen and Detmar Meurers

Coursebook Texts as a Helping Hand for Classifying Linguistic Complexity in Language Learners' Writings
Ildikó Pilán, David Alfter and Elena Volodina

Using Ambiguity Detection to Streamline Linguistic Annotation
Wajdi Zaghouani, Abdelati Hawwari, Sawsan Alqahtani, Houda Bouamor, Mahmoud Ghoneim, Mona Diab and Kemal Oflazer

Morphological Complexity Influences Verb-Object Order in Swedish Sign Language
Johannes Bjerva and Carl Börstell

A Comparison Between Morphological Complexity Measures: Typological Data vs. Language Corpora
Christian Bentz, Tatyana Ruzsics, Alexander Koplenig and Tanja Samardzic

Similarity-Based Alignment of Monolingual Corpora for Text Simplification Purposes
Sarah Albertsson, Evelina Rennes and Arne Jonsson

Automatic Construction of Large Readability Corpora
Jorge Alberto Wagner Filho, Rodrigo Wilkens and Aline Villavicencio

Syntactic and Lexical Complexity in Italian Noncanonical Structures
Rodolfo Delmonte

Testing the Processing Hypothesis of word order variation using a probabilistic language model
Jelke Bloem

Temporal Lobes as Combinatory Engines for both Form and Meaning
Jixing Li, Jonathan Brennan, Adam Mahar and John Hale

Automatic Speech Recognition Errors as a Predictor of L2 Listening Difficulties
Maryam Sadat Mirzaei, Kourosh Meshgi and Tatsuya Kawahara

Sunday December 11, 2016 (continued)

Quantifying sentence complexity based on eye-tracking measures
Abhinav Deep Singh, Poojan Mehta, Samar Husain and Rajkumar Rajakrishnan

Real Multi-Sense or Pseudo Multi-Sense: An Approach to Improve Word Representation
Haoyue Shi, Caihua Li and Junfeng Hu

15:20–15:40 **Break**

15:40–16:30 **Session 3 – Oral presentations**

15:40–16:05 *Upper Bound of Entropy Rate Revisited —A New Extrapolation of Compressed Large-Scale Corpora—*
Ryosuke Takahira, Kumiko Tanaka-Ishii and Łukasz Dębowski

16:05–16:30 *Learning pressures reduce morphological complexity: Linking corpus, computational and experimental evidence*
Christian Bentz and Aleksandrs Berdicevskis

16:30–17:00 **Round table and closing session**