Linguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
October 2015

Literature Lifts Up Computational Linguistics

David K. Elson
Anna Feldman
Anna Kazantseva
Stan Szpakowicz

Published by CSLI Publications

LiLT volume 12, issue 1 October 2015

Literature Lifts Up Computational Linguistics

David K. Elson, *Google, delson4@gmail.com*
Anna Feldman, *Montclair State University, feldmana@mail.montclair.edu*
Anna Kazantseva, *National Research Council Canada, anna@anna-kazantseva.com*
Stan Szpakowicz, *University of Ottawa, szpak@eecs.uottawa.ca*

Preface to the special issue

This collection of papers is a vignette of the first three of an ongoing series of workshops on Computational Linguistics for Literature (CLfL), collocated with conferences organized by the Association for Computational Linguistics.[1] The aim of the workshops is to create a forum for computational linguists who share a fascination with literature. The workshops have boasted papers on a wide variety of exciting topics such as computational treatment of poetry, automatic identification of quotable text, generation of music from literature, and computational models of narratives, to name but a few. This volume contains a small yet representative sample of research which our community carried out between 2012 and 2014.

The CLfL workshops cover a somewhat diffuse area. Some of the papers look at how computation can answer questions posed by the humanities. Other papers ask how the state of the art in Natural Lan-

[1]The fourth workshop (https://sites.google.com/site/clfl2015/) is already history.

LiLT Volume 12, Issue 1, October 2015.
Literature Lifts Up
Computational Linguistics.

guage Processing can help process literary data, both to make them more easily accessible and to suggest new areas of application. Perhaps the themes of the workshops are best defined by enumerating all papers accepted thus far; all papers are freely available in the ACL Anthology.[2]

What emerges from these papers is the thrill of exploring a new domain: computational treatment of literary text was rather rare as recently as the turn of the decade. As in any initial exploration of a problem space, these early attempts pursue a variety of subjects (in this case, literary modes and genres), evaluation methods, and even definitions of the problem. We might sum up that work as attempts to figure out how computational linguistics can work with the humanities to provide new insights into a body of world literature which has long since grown too massive for any scholar to read in a lifetime.

Literary prose and poetry are likely to be the genres most challenging for text understanding. This could be due to the predilection of many writers for treating ambiguity and indirectness as aesthetic methods—as opposed to, say, the relative clarity and conciseness that journalists favor. A literary author relies on a history of sensory, interpersonal and cultural experiences shared with a reader to relate an experience without explicating it, and just as a reader from a different culture may miss some of these references and nuances, a computer system misses most if not all of them. And yet, in exchange, the computer is an extremely attentive reader when it comes to surface-level considerations like style. It can pick up on even the most minute differences in syntax and word usage, at scale. To understand how a system reads literature, to extend its capabilities for doing so, and thereby to derive insights about where we have come as a species of fabulists, poets and stylists—this is the essence of the formidable challenge which our workshops have attempted to embrace.

Five of the papers from the CLfL anthology, substantially reworked, appear in this issue. Two of them tackle poetry, two take on prose, and one considers literary texts in general.

Julian Brooke, Adam Hammond and Graeme Hirst examine one very influential poem,[3] and they do it in quite some detail. "Distinguishing Voices in *The Waste Land* using Computational Stylistics" describes how stylistic analysis and text segmentation can automatically identify and cluster the voices of the multiple speakers in T. S. Eliot's mas-

[2] http://www.aclweb.org/anthology/W/W12/#2500
http://www.aclweb.org/anthology/W/W13/#1400
http://www.aclweb.org/anthology/W/W14/#0900
http://www.aclweb.org/anthology/W/W15/#0700

[3] http://www.bartleby.com/201/1.html

terpiece. The Authors also undertake quantitative comparison between the characters of the poem, and track how each individual voice changes from the beginning to the end. The paper contains a goodly share of literary analysis, which should please the less technically inclined readers.

Justine Kao and Dan Jurafsky attempt a computational take on a whole poetic movement: Imagism.[4] "A computational analysis of poetic style. Imagism and its influence on modern professional and amateur poetry" deploys quantitative methods to capture the distinct features of Imagist poetry and to compare those poems against a collection of more conventional nineteenth-century poems. The Authors also trace the influence of Imagist poets on the contemporary poetry scene, among both professional and amateur poets.

Mariona Coll Ardanuy and Caroline Sporleder propose a way to model novels using social networks of characters. "Clustering of Novels Represented as Social Networks" describes a system where each novel is represented by both a static graph of characters and a dynamic graph that captures the development of the plot. The Authors put to work several insightful metrics to describe the graphs and to cluster them by similarity. The paper shows the applicability of this method to authorship attribution and to clustering by genre.

Micha Elsner captures plot structure in a corpus of nineteenth-century English novels by relying mainly on lexical distributions. "Abstract Representations of Plot Structure" considers character similarity, sentiment-bearing words and generic content words in pursuit of robust, shallow metrics which can determine how similar two novels are to one another.

Finally, Yong Xu, Aurélien Max and François Yvon look at free e-books available in multiple languages as a potential source of high quality parallel corpora. "Sentence alignment for literary texts. The state-of-the-art and beyond" describes the difficulties in aligning translations of literary works, and describes a multi-pass system which addresses this problem. The technology which the paper advocates works for most text genres; it will serve literary data well.

Before we let you dip into this exciting volume, we wish to acknowledge the invaluable help of our wonderful reviewers:

Apoorv Agarwal
Cecilia Ovesdotter Alm
Chris Brew

[4] Imagism was *a 20th century movement in poetry advocating free verse and the expression of ideas and emotions through clear precise images* (http://www.merriam-webster.com/dictionary/imagism).

Mark Finlayson
Pablo Gervás
Graeme Hirst
Matthew Jockers
Mike Kestemont
Daniel Marcu
Rada Mihalcea
Nick Montfort
Vivi Nastase
Sebastian Padó
Caroline Sporleder

Thanks to you all!

David, Anna, Anna and Stan