Recreational Linguistics 1

Dragomir Radev *Editor*

Puzzles in Logic, Languages and Computation

The Red Book

Springer

Puzzles in Logic, Languages and Computation

Recreational Linguistics

Volume 1

For further volumes:
http://www.springer.com/series/11630

Dragomir Radev
Editor

Puzzles in Logic, Languages and Computation

The Red Book

Springer

Editor
Dragomir Radev
Department of Electrical Engineering and Computer Science
School of Information, Department of Linguistics
University of Michigan
Ann Arbor, MI, USA

Foreword by
James Pustejovsky
Brandeis University
Department of Computer Science
Volen Center for Complex Systems
Waltham, MA, USA

ISBN 978-3-642-34377-3 ISBN 978-3-642-34378-0 (eBook)
DOI 10.1007/978-3-642-34378-0
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013931838

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Axinia, Laura, and Victoria

Foreword

By
James Pustejovsky
TJX/Feldberg Professor of Computer Science
Brandeis University

This book brings together, for the first time in one collection, the best English-language problems created for students competing in the Computational Linguistics Olympiad. These problems are representative of the diverse areas presented in the competition and designed with three principles in mind:

- To challenge the student analytically, without requiring any explicit knowledge or experience in linguistics or computer science;

- To expose the student to the different kinds of reasoning required when encountering a new phenomenon in a language, both as a theoretical topic and as an applied problem.

- To foster the natural curiosity students have about the workings of their own language, as well as to introduce them to the beauty and structure of other languages.

The Linguistics Olympiad is designed to develop metalinguistic reasoning that is useful for any career involving human language and also to foster analytical problem-solving skills that are relevant for many technical and non-technical careers. The problems represented in this volume also emphasize an aptitude for computational thinking more than linguistics Olympiads in other countries. In addition to using logical and analytical skills, they explicitly focus on concepts and tools from computer science, such as finite state machines and graph search, while also introducing applications of computational linguistics, such as machine translation, information extraction, and automatic summarization.

Aside from being a fun intellectual challenge, the Olympiad mimics the skills used by researchers and scholars in the field of computational linguistics, which is increasingly important for the United States and other countries. Using computational linguistics, these experts can develop automated technologies such as translation software that cut down on the time and training needed to work with other languages, or software that automatically produces informative English summaries of documents in other languages or answers questions about information in these documents. In an increasingly global economy where businesses operate across borders and languages, having a strong pool of computational linguists is a competitive advantage, and an important component to both security and growth in the 21st century.

This collection of problems for the linguistics Olympiad is not only a valuable resource for high school students wishing to prepare themselves for the competition, but is a wonderful general introduction to the field of linguistics through the analytic problem solving technique.

Preface to Volume 1

This two-volume set includes more than 100 original problems (and their solutions) in (traditional) linguistics and computational linguistics. Many of the problems were used in the first five installments of NACLO[1] (North American Computational Linguistics Olympiad). NACLO, inaugurated in 2007, is an annual competition for high school students interested in human languages as well as the ways in which humans and computers deal with them using logic. NACLO is modeled after the IOL[2] (International Linguistics Olympiad) but, unlike IOL, includes a large percentage of problems in formal and computational linguistics. NACLO is a part of ELCLO (the consortium of English-language computational linguistics Olympiads, which includes Australia, Ireland, and Great Britain in addition to NACLO's members, USA and Canada).

This collection has been edited and augmented in order to make it appealing to a variety of audiences, from middle and high school students interested in languages, to teachers of languages, linguistics, and computer science, and to anyone fascinated by the phenomena of human language. All problems include detailed solutions that indicate how one can reach the answer even without any knowledge about the specific language or phenomenon on which the problems are based. The authors of the problems are linguistics and computer science professors and students and include several past contestants in the IOL, NACLO, and similar competitions.

The history of linguistics Olympiads started in 1965 in Russia when Andrey Zaliznyak organized the first local competition. Other contests followed in Eastern Europe. The first such contest in the USA was run in 1998-2000 by Thomas Payne for students in the Eugene, Oregon area. A hiatus followed and it was not until 2007 that another competition took place in North America. This time it was run nationwide in the USA and Canada. A number of students were able to participate remotely by having their high school teachers monitor the competition locally and send their papers to the NACLO committee for grading. The founders of NACLO include Tanya Korelsky (NSF) and also Lori Levin (CMU) and Thomas Payne (U. Oregon), as NACLO co-chairs, James Pustejovsky (Brandeis), as sponsorship chair, and Dragomir Radev (U. Michigan), as program chair and head coach of the US teams. A smaller and easier contest has also been held since 2007 by Aleka Blackwell in the Murphreesboro, TN area.

At the international level, a competition has taken place since 2003. It has been held ten times so far and has been hosted by Russia, Estonia, Poland, Bulgaria, the Netherlands, the USA, Slovenia, and Sweden. The eleventh IOL is scheduled for Manchester, UK in 2013. The international linguistics Olympiads have been modeled after similar competitions in Mathematics, Physics, Informatics, Biology, Chemistry, and other subjects. As of 2012, the strongest teams in Linguistics at the international level are the United States, Russia, Bulgaria, and Poland as well as the Netherlands, the United Kingdom, South Korea, and Latvia.

[1] http://www.naclo.cs.cmu.edu
[2] http://www.ioling.org

The goal of NACLO from its inception has been to popularize language and language technologies to high school students and to encourage them to pursue careers in these fields of study. Approximately 40% of all contestants in NACLO so far have been female. The very first NACLO winner was also female. The list of all NACLO winners so far includes: Rachel Zax (Ithaca, NY – 2007), Guy Tabachnick (New York, NY – 2008), Anand Natarajan (San Jose, CA – 2009), Ben Sklaroff (Palo Alto, CA – 2010), Daniel Mitropolsky (Oakville, Ontario, Canada – 2011), and Alex Wade (Reno, NV – 2012).

NACLO contestants can participate in two ways – at a university site nearby (if one exists) or at their own high school. The format is the same in both cases. The first round involves around 5-7 problems in a three hour time slot whereas the second (invitational) round includes 6-10 problems and takes 5 hours. Since 2010, first round problems have been graded by an auto-grading program while the second round problems are manually graded by the NACLO volunteer jury.

A large number of US students have performed extremely well at the ILO. The awards received form a really long list that includes over the past six years, the following: six out of 22 individual gold medals, four out of 7 team gold medals, and three out of five combined gold medals. The performance by year is shown here:

2007 Individual Gold (rank 1 of 3): Adam Hesterberg, Seattle, WA
    Team Gold (tied): Rebecca Jacobs, Joshua Falk, Michael Gottlieb, and Anna Tchetchetkine
2008 Individual Gold (rank 2 of 3): Hanzhi Zhu
    Team Gold (tied): Morris Alper, Rebecca Jacobs, Jae-kyu Lee, and Hanzhi Zhu
    Combined Gold: Josh Falk, Jeffrey Lim, Anand Natarajan, Guy Tabachnick
2009 Team Gold: Rebecca Jacobs, Anand Natarajan, Alan Huang, and Morris Alper
2010 Individual Gold (rank 3 of 3): Ben Sklaroff, Palo Alto, CA
    Combined Gold: Martin Camacho, Tian-Yi Damien Jiang, Alexander Iriza, and Alan Chang
2011 Individual Gold (rank 1 of 4): Morris Alper, Palo Alto, CA
    Team Gold: Morris Alper, Aaron Klein, Wesley Jones, and Duligur Ibeling
    Combined Gold: Morris Alper, Aaron Klein, Wesley Jones, and Duligur Ibeling
2012 Individual Gold (ranked 2 of 8): Alexander Wade, Reno, NV
    Individual Gold (ranked 4 of 8): Aderson Wang, PA
    Team Gold: Alexander Wade, Aidan Kaplan, Aaron Klein, and Erik Andersen

Other NACLO students who received medals at the international level:

Silver: Anand Natarajan (2008), Morris Alper (2008), Rebecca Jacobs (2009), Allen Yuan (2010), Martin Camacho (2010), Tian-Yi Damien Jiang (2010), Wesley Jones (2011), Allen Yuan (2011), Alexander Wade (2011), Duligur Ibeling (2011), Darryl Wu (2012), Aaron Klein (2012), Allan Sadun (2012).
Bronze: Jeffrey Lim (2008), Guy Tabachnick (2008), Rebecca Jacobs (2008), Alan Huang (2009), John Berman (2009), Sergei Bernstein (2009), Alexander Iriza (2010), Alan Chang (2010), Aaron Klein (2011), Daniel Mitropolsky (2011), Erik Andersen (2012).

More than 6,000 students have participated in NACLO so far. Each year, the top 100 (or so) get invited to the second "invitational" round which is used to determine the composition of the US and Canadian teams.

This book set includes two general classes of problems. The computational problems focus on formal and computational aspects of language understanding and automated language processing. The traditional problems pay more attention to linguistic phenomena such as morphology, phonology, phonetics, syntax, semantics, and pragmatics.

The traditional problems are on more than 50 languages and writing systems:
European/Middle Eastern: Ancient Greek, Swedish, Norwegian, Irish, Linear B, Bulgarian, Spanish, Armenian, Turkish, Romanian, Italian, Catalan, German, Etruscan, Hebrew, Welsh, Old Church Slavonic, Arabic, Russian, French, Greek, Albanian, English
Asian: Hmong, Huishu, Hindi, Japanese, Hmong, Tangkhul, Malayalam, Tamil, Vietnamese, Indonesian, Korean
Pacific: Ilocano, Manam Pile, Rotokas, Dyirbal, Abma, Minangkabau, Tanna, Arrernte, Warlpiri, Central Cagayan Agta, Hawaiian, Wembawemba, Pitjantjatjara, Nen, Enga
African: Swahili, Tadaksahak, Amharic, Maasai, Bambara
North and South American: Apinaye, Aymara, Tzolk'in, Quechua, Guaraní, Plains Cree, Tohono O'odham, Ulwa, Nahuatl, Ndyuka, Anishinaabemowin, Sahaptin, Zoque,

The computational problems cover: document retrieval, sentence similarity, optical character recognition, garden path sentences, semantics of noun-noun compounds, stemming, finite state automata, spectrograms, writing systems for the blind, spelling correction, text summarization, polarity induction, deixis, shift-reduce parsing, context-free grammars, named entity classification, text compression, machine translation, expansion of abbreviations, logical entailment, presupposition, word sense disambiguation, text processing, word reordering, syntactic ambiguity, handwriting recognition, word frequencies, syntactic transformations, recursion, modeling second language learning errors, sentence boundary identification, computational morphology, cognates, text classification, and Zipf's law.

In addition to the authors of the problems, I would like to thanks specifically the folks below for all their hard work over the years to make NACLO happen: Emily Bender, Mary Jo Bensasi, Marcus Berger, John Berman, Reed Blaylock, Eric Breck, Justin Brown, Rich Caneba, Hyunzoo Chai, Angie Chang, Ivan Derzhanski, Jason Eisner, Adam Emerson, Dominique Estival, Barbara di Eugenio, Jefferson Ezra, Eugene Fink, Anatole Gershman, Blumie Gourarie, Mercedes Harvey, Amy Hemmeter, Adam Hesterberg, Dick Hudson, Boris Iomdin, Alexander Iriza, Rebecca Jacobs, Ridley Jones, Wesley Jones, Tanya Korelsky, Nate LaFave, Andrew Lamont, Terry Langendoen, Rachael Leduc, Lillian Lee, Will Lewis, Pat Littell, Wanchen Lu, Rachel McEnroe, Ruslan Mitkov, Graham Morehead, David Mortensen, JP Obley, Martha Palmer, Tom Payne, Carrie Pichan, Ben Piche, Victor Pudeyev, James Pustejovsky, Vahed Qazvinian, Laura Radev, Adrienne Reed, Rahel Ringger, Meredith Rogan, David Ross, Andrea Sexton, Catherine Sheard, Ben Sklaroff, Catherine Arnott Smith, Noah Smith, Samuel Smolkin, Harold Somers, Richard Sproat, Kurnikova Stacy, Laine Stranahan, Rebecca Sundae, Jennifer Sussex, Roula Svorou, Aditya Tayade, Sally Thomason, Amy Troyani, Susanne Vejdemo, Zilin Wang, and Julia Workman.

The following organizations provided funding and other support for NACLO: the National Science Foundation, the Linguistics Society of America, the North American Chapter of the Association for Computational Linguistics, as well as Google, Yahoo!, Cambridge University Press, as well as many known and unknown supporters. They all deserve their own ticker tape parade for their contributions to NACLO.

How to use this book

The first volume contains 56 problems and the second one includes an additional 50. Based on the performance of the students in NACLO, each of the problems in the book set has been assigned a difficulty score from Beginner (one star) to Expert (five stars). Any person who has made it to high school has a shot at most of these problems. In fact, even the five-star problems were solved successfully during NACLO by dozens of high school students working under severe time pressure.

Problems of a more computational nature are marked with a small computer icon next to their names. The icon doesn't imply that a computer is needed to solve these problems but rather that they are related to computational linguistics.

The NACLO web site links to dozens of additional problems as well as numerous presentations, tutorials, and other educational materials for students and teachers alike.

Final words

In January 2011, the Linguistics Society of America awarded NACLO its "Linguistics, Language, and the Public" award for increasing awareness of linguistics in the general public. This honor is a strong indication of the recognition of the role that NACLO plays.

Dragomir Radev, NACLO Program Chair and US team head coach

September 30, 2012

Ann Arbor and New York

Table of Contents