Tiny Transactions on Computer Science (TinyToCS) is the premier venue for computer science research of 140 characters or less.

Volume 3 Index

Formal Verification and Programming Languages

- **I compute, therefore I am (buggy): methodic doubt meets multiprocessors**
  Inspired by Descartes' methodic doubt, we systematically test manycore chips to dispel and correct common false memory ordering assumptions.
  Jade Alglave (Microsoft Research Cambridge, University College London), Luc Maranget (INRIA), Daniel Poetzl (University of Oxford), Tyler Sorensen (University College London)

- **Embracing Overapproximation for Proving Nontermination**
  We embrace overapproximation to prove nontermination: a live abstraction's closed recurrence implies the input program's nontermination.
  Byron Cook, Carsten Fuhs, Kaustubh Nimkar, Peter O'Hearn (University College London)

- **How to Authenticate any Data Structure**
  In our new language, λ*, every data structure has an authenticated "merkleized" variant, safe to store on untrusted servers.
  Andrew Miller, Michael Hicks, Jonathan Katz, Elaine Shi (University of Maryland)

Machine Learning and Natural Language Processing

- **The Not-so-Magic Magic Barrier of Recommender Systems**
  We show that the magic barrier of recommender systems is not magic at all. In fact, it is an easily computed statistical rating property.
  Alejandro Bellogín (Universidad Autónoma de Madrid), Alan Said
Sarcasm Detection: Beyond Machine Learning Algorithms
Learning Algorithms just like humans need context to detect sarcasm.
Vineet Kumar (Capgemini)

Discriminating Similar Languages: Persian and Dari
Language Identification methods using surface features can distinguish close linguistic variants Persian and Dari with 96% accuracy.
Shervin Malmasi (Centre for Language Technology)

Non-transparent recommender system evaluation leads to misleading results
Evaluation of recommender systems is often misleading, transparent protocols should be used and reported by industry and academia.
Alan Said (Recorded Future), Alejandro Bellogín (Universidad Autónoma de Madrid)

Handling Information Overload: Automatic Generation of Wikipedia Articles
Semantic relationships among existing Wikipedia articles can be used to generate high-quality structured information on a given topic.
Vikrant Yadav, Faisal Khan (Indian Institute of Technology, Roorkee)

Pervasive Computing

Tiny Location Fingerprints Uniquely Identify Individuals
A sequence of 7 (± 2) significant places reconstructed from smartphone data, distinguishes an individual amongst a large human population.
George Roussos (Birkbeck College, University of London)

Security and Privacy

To Trust, or not to Trust, that is the Question: Structural Properties of X.509 Certificates
Even for a human observer with full knowledge it is impossible to decide if a new certificate is legitimate without out-of-band context.
Johanna Amann (International Computer Science Institute), Robin Sommer (International Computer Science Institute, Lawrence Berkeley National Laboratory), Matthias Vallentin (UC Berkeley), Seth Hall (International Computer Science Institute)

KISS On My List: Improving Platform Security and User Privacy by Simplifying the Android Permissions Model
Could simplifying Android's permission model improve platform security and privacy by easing development and promoting user understanding?
Nicole Borrelli
Preimage Attacks Against Spectral Hash and PTX Functions
We break the pre-image security of PTX and shash by reducing it to the trivially solvable problem of finding independent random vectors.
Ethan Heilman (Boston University)

Software Engineering

The More the Merrier: Leveraging on the Bug Inflow to Guide Software Maintenance
Humans obscured by bug overload, but machine learning benefits from plentiful training data. Practitioners confirm value of developed tools.
Markus Borg (Lund University), Leif Jonsson (Ericsson AB)

On Integration Repositories, Build Sheriffs, and Patch Backouts
Integration repositories and build sheriffs allow developers to test less comprehensively while keeping the main repository stable.
Rodrigo Souza, Christina Chavez (Federal University of Bahia), Roberto A. Bittencourt (State University of Feira de Santana)

Systems and Networks

Supporting Differentiated Services in Computers via Networking Technologies
A computer is inherently a network. It is promising to apply networking technologies like SDN to computer architecture.
Jiuyue Ma, Xiufeng Sui, Yupeng Li, Zihao Yu, Bowe Huang, Yungang Bao (State Key Laboratory of Computer Architecture, ICT, CAS)

Stream Processing Systems Have Arrived at the Big Data Party. But Where Are All the Benchmarks?
The ubiquity of stream processing systems and applications necessitates the development of a real-world TPC-like benchmark suite.
Zubair Nabi (IBM Research - Ireland)

The role of reconfigurable, policy-based middleware in supporting the Internet of Things
IoT requires middleware enabling the interactions of 'things' to be dynamically managed and regulated according to user/application policy.
Jatinder Singh (University of Cambridge)

Theory of Computation and Information Science

On Regular Expression Matching and Deterministic Finite Automata
Even ignoring construction time, deterministic finite automata do not solve regular expression matching in worst-case linear time.
Philip Bille (Technical University of Denmark, DTU Compute)
Cognitive computing leads to the next level of answering questions on the Web
From Search Engines to Question Answering Systems or
\[1]+[2]=[3].

Edy Portmann (*University of Bern*)

**Tiny ToCS Volume 3 Organizers**

Program Chair:

- **Heidy Khlaaf**, *University College London*

Program Committee:

- **Peter Bailis**, *University of California, Berkeley*
- **Yonatan Bisk**, *University of Illinois Urbana-Champaign*
- **Sarah Deisburg**, *University of Northern Iowa*
- **Mike Dodds**, *University of York*
- **Andrew Ferguson**, *Google*
- **Mital Kinderkhedia**, *University College London*
- **Zachary Kincaid**, *University of Toronto*
- **Lindsey Kuper**, *Intel*
- **Katie Kuksenok**, *University of Washington*
- **Jinna Lei**, *University of Washington*
- **Igor Mordatch**, *University of Washington*
- **Faisal Nawab**, *University of California Santa Barbara*
- **Sunil Pedapudi**, *Google*
- **Malte Schwarzkopf**, *University of Cambridge*
- **Divya Sharma**, *Carnegie Mellon University*
- **Stephen Strowes**, *Yahoo*
- **Yuhao Zhu**, *The University of Texas at Austin*

**Contact**: tiny-tocs-pc at googlegroups dot com

Include a link to Tiny ToCS on your website:
Discriminating Similar Languages: Persian and Dari

Shervin Malmasi
Centre for Language Technology
Macquarie University, Sydney, Australia
shervin.malmasi@mq.edu.au

ABSTRACT
Although widely-studied in recent years, Language Identification (LID) systems for determining the language of input texts often fail to discriminate between similar languages like Croatian-Serbian and Malay-Indonesian. This has brought attention to the task of discriminating similar languages, varieties and dialects – including a recent shared task [3].

Persian (also known as Farsi) and Dari (Eastern Persian, spoken predominantly in Afghanistan) are two close variants that have not hitherto been investigated in LID and we report the first results on this pair. Dari is a low-resourced but important language, particularly for the U.S. due to its ongoing involvement in Afghanistan, which has led to increasing research interest [1].

We developed a corpus of 28k sentences (14k per-language) and using character and word n-grams, we discriminated them with 96% accuracy. Out-of-domain cross-corpus evaluation was conducted to test the discriminative models' generalizability, achieving 87% accuracy in classifying 79k sentences from the Uppsala Persian Corpus. Feature analysis revealed lexical, morphological and orthographic inter-language differences.

Further to determining document languages, LID has applications in character encoding detection, statistical machine translation, inducing dialect-to-dialect lexicons and authorship profiling in the forensic linguistics domain. In Information Retrieval it can help filter documents (e.g. news articles or search results) by dialect.

LID can also be used in other Natural Language Processing tasks, including the adaptation of tools like part-of-speech taggers for low-resourced languages [2]. Since Dari is too different to directly apply Persian resources, the distinguishing features identified through LID can assist in adapting existing resources.

BODY
Language Identification methods using surface features can distinguish close linguistic variants Persian and Dari with 96% accuracy.

REFERENCES
Call for Papers

Tiny Transactions on Computer Science seeks papers describing significant research contributions to the field of computer science. Tiny ToCS is the premier venue for computer science research of 140 characters or less.

Why Tiny ToCS?

Tiny ToCS is a forum for high-quality scholarly discourse that challenges researchers to engage with their work in new capacities. Not only does the restricted format require a concise distillation of scholarly results, but the short format will be more accessible to the Computer Science community at large. Tiny ToCS favors accessible, carefully crafted, creative research, and invites submissions both about novel ideas as well as summaries or distillations of existing work.

We invite submissions on a wide range of computer science research, including, but not limited to:

- Artificial Intelligence
- Computer Architecture
- Computer Networking
- Computer Systems
- Computer Security
- Databases and Data
- Management
- Graphics
- Human-Computer Interaction
- Machine Learning
- Programming Languages
- Theory of Computation

Submissions

Submissions to Tiny ToCS can be up to 140 characters in length, with an abstract of no more than 250 words and a title of no more than 118 characters.

Alternate (video) format: Videos of up to 120 seconds may also be submitted. Authors should be careful to maintain anonymity in the video. The attached video description must fit within 118 characters so that it
may be tweeted alongside the video. Videos should be submitted directly to the PC via email: Heidy Khlaaf.

The primary content of submissions should fit into 140 characters. The abstract is not intended to expound upon your finding but instead to provide context for your work. What is the background for your research? Concisely, how does this work improve on related research? Submissions which violate this requirement will be rejected without consideration.

You may use three references in your submission. References will count as one word when used in the abstract section and three letters in the body section ("[1]”).

Your submission may not contain any non-text figures or special formatting. Plain-text only. ASCII art okay.

Reviewing will be double-blind and serve primarily to ensure basic quality of accepted papers. Submissions that are unverifiable or purely humorous will be rejected; submissions that highlight nuggets in or concisely summarize existing work are acceptable.

For inspiration, we have provided a sample submission, or refer to Tiny ToCS Volume 1.

### Submitting Papers and Logistics

Please validate your submission before submitting.

If possible, please format your submission(s) using the Tiny ToCS LaTeX class file. We’ve provided a sample skeleton file for you. At the least, please submit in PDF format (required by our conference software). And remember, leave names and affiliations off of your PDF.

Upload your submission here.

### Important Dates for Volume 3:

- Final submission: January 15, 2015, 11:59 PM PST
- Notification: February 15, 2015
- Camera Ready Due: March 1, 2015
- Publication: March 5, 2015

### Publication Details

The proceedings of Tiny ToCS Volume 3 will be posted to the arXiv under a Creative Commons Attribution-NonCommercial-ShareAlike license. Authors will retain all future publishing rights.

### Tiny ToCS Volume 3 Organizers

Program Chair:
Heidy Khlaaf, University College London

Program Committee:

- Peter Bailis, University of California, Berkeley
- Yonatan Bisk, University of Illinois Urbana-Champaign
- Sarah Deisburg, University of Northern Iowa
- Mike Dodds, University of York
- Andrew Ferguson, Google
- Mital Kinderkhedia, University College London
- Zachary Kincaid, University of Toronto
- Lindsey Kuper, Intel
- Katie Kuksenok, University of Washington
- Jinna Lei, University of Washington
- Igore Mordatch, University of Washington
- Faisal Nawab, University of California Santa Barbara
- Sunil Pedapudi, Google
- Malte Schwarzkopf, University of Cambridge
- Divya Sharma, Carnegie Mellon University
- Stephen Strowes, Yahoo
- Yuhao Zhu, The University of Texas at Austin

Contact: tiny-tocs-pc at googlegroups dot com

Include a link to Tiny ToCS on your website:
<a href="http://tinytocs.org"><img src="http://tinytocs.org/ttocs-thumb.png"></a>