

**INTERNATIONAL CONFERENCE  
RECENT ADVANCES  
IN NATURAL LANGUAGE PROCESSING**

**R A N L P 2 0 1 9**

**Natural Language Processing  
in a Deep Learning World**

**P R O C E E D I N G S**

Edited by Galia Angelova, Ruslan Mitkov, Ivelina Nikolova, Irina Temnikova

Varna, Bulgaria  
2–4 September, 2019

**INTERNATIONAL CONFERENCE  
RECENT ADVANCES IN  
NATURAL LANGUAGE PROCESSING 2019**

Natural Language Processing  
in a Deep Learning World

**PROCEEDINGS**

Varna, Bulgaria  
2–8 September 2019

Print ISBN 978-954-452-055-7  
Online ISBN 978-954-452-056-4  
Series Print ISSN 1313-8502  
Series Online ISSN 2603-2813

Designed and Printed by INCOMA Ltd.  
Shoumen, BULGARIA

## Preface

Welcome to the 12th International Conference on “Recent Advances in Natural Language Processing” (RANLP 2019) in Varna, Bulgaria, 2-4 September 2019. The main objective of the conference is to give researchers the opportunity to present new results in Natural Language Processing (NLP) based on modern theories and methodologies.

The Conference is preceded by the First Summer school on Deep Learning in NLP (29-30 August 2019) and two days of tutorials (31 August – 1 September 2019).

The Summer School lectures are given by Kyunghyun Cho (New York University), Marek Rei (University of Cambridge), Tim Rocktäschel (University College London) and Hinrich Schütze (Ludwig Maximilian University, Munich). Training in practical sessions is provided by Heike Adel (Stuttgart University), Alexander Popov (Institute of Information and Communication Technologies, Bulgarian Academy of Sciences), Omid Rohanian and Shiva Taslimipour (University of Wolverhampton).

Tutorials are given by the following lecturers: Antonio Miceli Barone (University of Edinburgh) and Sheila Castilho (Dublin City University), Valia Kordoni (Humboldt University, Berlin), Preslav Nakov (Qatar Computing Research Institute, HBKU), Vlad Niculae and Tsvetomila Mihaylova (Institute of Telecommunications, Lisbon).

The conference keynote speakers are:

- Kyunghyun Cho (New York University),
- Ken Church (Baidu),
- Preslav Nakov (Qatar Computing Research Institute, HBKU),
- Sebastian Padó (Stuttgart University),
- Hinrich Schütze (Ludwig Maximilian University, Munich).

This year 18 regular papers, 37 short papers, 95 posters, and 7 demos have been accepted for presentation at the conference. The selection rate of accepted papers is: regular papers 8,7%, short papers 26,7%, posters and demo papers – 72%.

The proceedings cover a wide variety of NLP topics, including but not limited to: deep learning; machine translation; opinion mining and sentiment analysis; semantics and discourse; named entity recognition; coreference resolution; corpus annotation; parsing and morphology; text summarisation and simplification; event extraction; fact checking and rumour analysis; NLP for healthcare; and NLP for social media.

In 2019 RANLP hosts four post-conference workshops on influential NLP topics: the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019), the

12th Workshop on Building and Using Comparable Corpora (BUCC), the Multiling 2019 Workshop: Summarization Across Languages, Genres and Sources as well as an Workshop on Language Technology for Digital Historical Archives with a Special Focus on Central-, (South-)Eastern Europe, Middle East and North Africa. The International Conference Biographical Data in a Digital World 2019 is another event held on 5-6 September 2019 in parallel with the RANLP post-conference Workshops.

We would like to thank all members of the Programme Committee and all additional reviewers. Together they have ensured that the best papers were included in the Proceedings and have provided invaluable comments for the authors.

Finally, special thanks go to the University of Wolverhampton, the Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences, the Bulgarian National Science Fund, Ontotext and IRIS.AI for their generous support of RANLP.

Welcome to Varna and we hope that you enjoy the conference!

The RANLP 2019 Organisers

**The International Conference RANLP–2019 is organised by:**

Research Group in Computational Linguistics,  
University of Wolverhampton, UK

Linguistic Modelling and Knowledge Processing Department,  
Institute of Information and Communication Technologies,  
Bulgarian Academy of Sciences, Bulgaria

**RANLP–2019 is partially supported by:**

National Science Fund, Ministry of Education and Science, Bulgaria

Ontotext AD

IRIS.AI

**Programme Committee Chair:**

Ruslan Mitkov, University of Wolverhampton, UK

**Organising Committee Chair:**

Galia Angelova, Bulgarian Academy of Sciences, Bulgaria

**Workshop Coordinator:**

Kiril Simov, Bulgarian Academy of Sciences, Bulgaria

**Tutorial Coordinator:**

Preslav Nakov, Qatar Computing Research Institute, HBKU, Qatar

**Proceedings Printing:**

Nikolai Nikolov, INCOMA Ltd., Shoumen, Bulgaria

**Programme Committee Coordinators:**

Ivelina Nikolova, Bulgarian Academy of Sciences  
Irina Temnikova, Bulgarian Academy of Sciences

## **Programme Committee:**

Ahmed Abdelali (Hamad Bin Khalifa University, Qatar)  
Cengiz Acarturk (Middle East Technical University, Turkey)  
Guadalupe Aguado-de-Cea (Polytechnic University of Madrid, Spain)  
Luis Alfonso Ureña López (University of Jaén, Spain)  
Hassina Aliane (Research Center on Scientific and Technical Information, Algeria)  
Pascal Amsili (University of Paris Diderot, France)  
Galia Angelova (Bulgarian Academy of Sciences, Bulgaria)  
Riza Batista-Navarro (University of Manchester, United Kingdom)  
Kalina Bontcheva (University of Sheffield, United Kingdom)  
Svetla Boytcheva (Bulgarian Academy of Sciences, Bulgaria)  
António Branco (University of Lisbon, Portugal)  
Chris Brew (Digital Operatives)  
Nicoletta Calzolari (Italian National Research Council, Italy)  
Sheila Castilho (Dublin City University, Ireland)  
Key-Sun Choi (Korea Advanced Institute of Science and Technology, South Korea)  
Kenneth Church (Baidu, United States of America)  
Kevin Cohen (University of Colorado School of Medicine, United States of America)  
Gloria Corpas Pastor (University of Málaga, Spain)  
Dan Cristea (University of Iași, Romania)  
Antonio Ferrández Rodríguez (University of Alicante, Spain)  
Fumiyo Fukumoto (University of Yamanashi, Japan)  
Prószéky Gábor (Pázmány Péter Catholic University & Bionics, Hungary)  
Alexander Gelbukh (National Polytechnic Institute, Mexico)  
Yota Georgakopoulou (Athena Consultancy, Greece)  
Ralph Grishman (New York University, United States of America)  
Veronique Hoste (Ghent University, Belgium)  
Diana Inkpen (University of Ottawa, Canada)  
Hitoshi Isahara (Toyohashi University of Technology, Japan)  
Miloš Jakubíček (Lexical Computing Ltd)  
Alma Kharrat (Microsoft)  
Udo Kruschwitz (University of Essex, United Kingdom)  
Sandra Kübler (Indiana University, United States of America)  
Katia Lida Kermanidis (Ionian University, Greece)  
Natalia Loukachevitch (Lomonosov Moscow State University, Russia)  
Eid Mohamed (Doha Institute for Graduate Studies, Qatar)  
Emad Mohamed (University of Wolverhampton, United Kingdom)  
Johanna Monti (University of Naples L'Orientale, Italy)  
Andrés Montoyo (University of Alicante, Spain)  
Alessandro Moschitti (Amazon)  
Rafael Muñoz Guillena (University of Alicante, Spain)  
Preslav Nakov (Qatar Computing Research Institute, Qatar)  
Roberto Navigli (Sapienza University of Rome, Italy)  
Raheel Nawaz (Manchester Metropolitan University, United Kingdom)  
Mark-Jan Nederhof (University of St Andrews, United Kingdom)  
Ivelina Nikolova (Bulgarian Academy of Sciences, Bulgaria)  
Kemal Oflazer (Carnegie Mellon University, Qatar)  
Maciej Ogrodniczuk (Polish Academy of Sciences, Poland)

Constantin Orasan (University of Wolverhampton, United Kingdom)  
 Petya Osenova (Sofia University and Bulgarian Academy of Sciences, Bulgaria)  
 Sebastian Padó (Stuttgart University, Germany)  
 Noa P. Cruz Diaz (Artificial Intelligence Excellence Center, Bankia, Spain)  
 Liviu P. Dinu (University of Bucharest, Romania)  
 Pavel Pecina (Charles University, Czech Republic)  
 Stelios Piperidis (Athena Research Center, Greece)  
 Massimo Poesio (University of Essex, United Kingdom)  
 Horacio Rodríguez (Polytechnic University of Catalonia, Spain)  
 Paolo Rosso (Polytechnic University of Valencia, Spain)  
 Vasile Rus (The University of Memphis, United States of America)  
 Frédérique Segond (Viseo)  
 Kiril Simov (Bulgarian Academy of Sciences, Bulgaria)  
 Vilemini Sosoni (Ionian University, Greece)  
 Keh-Yih Su (Institute of Information Science, Academia Sinica, Taiwan)  
 Stan Szpakowicz (University of Ottawa, Canada)  
 Hristo Tanev (European Commission, Belgium)  
 Shiva Taslimipoor (University of Wolverhampton, United Kingdom)  
 Irina Temnikova (Sofia University, Bulgaria)  
 Dan Tufiş (Romanian Academy of Sciences, Romania)  
 Aline Villavicencio (University of Essex, United Kingdom  
 and Federal University of Rio Grande do Sul, Brazil)  
 Yorick Wilks (Florida Institute for Human and Machine Cognition,  
 United States of America )  
 Mai Zaki (American University of Sharjah, United Arab Emirates)  
 Marcos Zampieri (University of Wolverhampton, United Kingdom)  
 Michael Zock (University of Aix-Marseille, France)

## **Reviewers:**

Ahmed AbuRa'ed (University Pompeu Fabra, Spain)  
 Mattia A. Di Gangi (University of Trento, Italy)  
 Itziar Aldabe (University of País Vasco, Spain)  
 Ahmed Ali (Hamad Bin Khalifa University, Qatar)  
 Ahmed Amine Aliane (Research Center on Scientific and Technical Information,  
 Algeria)  
 Le An Ha (University of Wolverhampton, United Kingdom)  
 Atefeh (Anna) Farzindar (University of Southern California, United States of America)  
 João António Rodrigues (University of Lisboa, Portugal)  
 Pepa Atanasova (University of Copenhagen, Denmark)  
 Mohammed Attia (George Washington University, United States of America)  
 Parnia Bahar (Aachen University, Germany)  
 Belahcene Bahloul (University of Khemis Miliana, Algeria)  
 Eduard Barbu (University of Tartu, Estonia)  
 Alberto Barrón-Cedeño (University of Bologna, Italy)  
 Leonor Becerra (Jean Monnet University, France)  
 Andrea Bellandi (National Research Council, Italy)  
 Fernando Benites (ZHAW School of Engineering, Switzerland)  
 Victoria Bobicev (Technical University of Moldova, Moldova)

Antonina Bondarenko (Lipetsk State Technical University, Russia)  
 Aurélien Bossard (University Paris 8, France)  
 Aljoscha Burchardt (German Research Centre for Artificial Intelligence, Germany)  
 Lindsay Bywood (University of Westminster, United Kingdom)  
 Ruket Cakici (Middle East Technical University, Turkey)  
 Iacer Calixto (University of Amsterdam, Netherlands and New York University, United States of America)  
 Pablo Calleja (Polytechnic University of Madrid, Spain)  
 Erik Cambria (Nanyang Technological University, Singapore)  
 Kai Cao (New York University, United States of America)  
 Thiago Castro Ferreira (Tilburg University, Netherlands)  
 Yue Chen (Queen Mary University of London, United Kingdom)  
 Mihaela Colhon (University of Craiova, Romania)  
 Daniel Dakota (Indiana University, United States of America)  
 Kareem Darwish (Hamad Bin Khalifa University, Qatar)  
 Orphee De Clercq (Ghent University, Belgium)  
 Kevin Deturck (Viseo)  
 Asma Djaidri (University of Science and Technology Houari Boumediene, Algeria)  
 Mazen Elagamy (Staffordshire University, United Kingdom)  
 Can Erten (University of York, United Kingdom)  
 Luis Espinosa Anke (Cardiff University, United Kingdom)  
 Kilian Evang (University of Düsseldorf, Germany)  
 Richard Evans (University of Wolverhampton, United Kingdom)  
 Stefan Evert (Friedrich–Alexander University, Germany)  
 Anna Feherova (University of Wolverhampton, United Kingdom)  
 Mariano Felice (University of Cambridge, United Kingdom)  
 Corina Forascu (The Alexandru Ioan Cuza University, Romania)  
 Vasiliki Foufi (University of Geneva, Switzerland)  
 Thomas Francois (Université catholique de Louvain, Belgium)  
 Adam Funk (University of Sheffield, United Kingdom)  
 Björn Gambäck (Norwegian University of Science and Technology, Norway)  
 Aina Garí Soler (The Computer Science Laboratory for Mechanics and Engineering Sciences, France)  
 Federico Gaspari (Dublin City University, Ireland)  
 José G. C. de Souza (eBay)  
 Goran Glavaš (University of Mannheim, Germany)  
 Darina Gold (University of Duisburg-Essen, Germany)  
 Reshmi Gopalakrishna Pillai (University of Wolverhampton, United Kingdom)  
 Rohit Gupta (University of Wolverhampton, United Kingdom)  
 Amir Hazem (Nantes University, France)  
 Tomáš Hercig (University of West Bohemia, Czech Republic)  
 Yasser Hifny (University of Helwan, Egypt)  
 Diliara Iakubova (Kazan Federal University, Russia)  
 Adrian Iftene (The Alexandru Ioan Cuza University, Romania)  
 Camelia Ignat (European Commission, Belgium)  
 Dmitry Ilvovsky (National Research University Higher School of Economics, Russia)  
 Miloš Jakubíček (Masaryk University, Czech Republic)  
 Arkadiusz Janz (Wroclaw University of Science and Technology, Poland)

Héctor Jiménez-Salazar (The Metropolitan Autonomous University, Mexico)  
 Olga Kanishcheva (National Technical University, Ukraine)  
 Georgi Karadzhov (Sofia University, Bulgaria)  
 David Kauchak (Pomona College, United States of America)  
 Yasen Kiproff (Sofia University, Bulgaria)  
 Jan Kocoń (Wroclaw University of Science and Technology, Poland)  
 Sarah Kohail (Hamburg University, Germany)  
 Yannis Korkontzelos (Edge Hill University, United Kingdom)  
 Venelin Kovatchev (University of Barcelona, Spain)  
 Peter Krejzl (University of West Bohemia, Czech Republic)  
 Sudip Kumar Naskar (Jadavpur University, India)  
 Maria Kunilovskaya (University of Tyumen, Russia)  
 Andrey Kutuzov (University of Oslo, Norway)  
 Sobha Lalitha Devi (Anna University, India)  
 Gabriella Lapesa (Universität Stuttgart, Institut für Maschinelle  
 Sprachverarbeitung, Germany)  
 Todor Lazarov (Bulgarian Academy of Sciences, Bulgaria)  
 Els Lefever (Ghent University, Belgium)  
 Ladislav Lenc (University of West Bohemia, Czech Republic)  
 Elena Lloret (University of Alicante)  
 Pintu Lohar (Dublin City University, Ireland)  
 Epida Loupaki (Aristotle University of Thessaloniki, Greece)  
 Lieve Macken (Ghent University, Belgium)  
 Mireille Makary (University of Wolverhampton, United Kingdom)  
 Michał Marcińczuk (Wroclaw University of Technology, Poland)  
 Angelo Mario Del Grosso (National Research Council of Italy, Italy)  
 Federico Martelli (Babelscape, Italy)  
 Patricia Martin Chozas (Polytechnic University of Madrid, Spain)  
 Eugenio Martínez-Cámara (University of Granada, Spain)  
 Irina Matveeva (NexLP, United States of America)  
 Flor Miriam Plaza del Arco (University of Jaén, Spain)  
 Arturo Montejo-Ráez (University of Jaén, Spain)  
 Paloma Moreda Pozo (University of Alicante, Spain)  
 Diego Moussallem (University of Paderborn, Germany)  
 Sara Moze (University of Wolverhampton, United Kingdom)  
 Nona Naderi (University of Toronto, Canada)  
 Marcin Oleksy (Wrocław University of Science and Technology, Poland)  
 Antoni Oliver (The Open University of Catalonia, Spain)  
 Mihaela Onofrei (University of Iasi, Romania)  
 Arzucan Özgür (Bogazici University, Turkey)  
 Santanu Pal (Saarland University, Germany)  
 Alexander Panchenko (University of Hamburg, Germany)  
 Sean Papay (University of Stuttgart, Germany)  
 Ljudmila Petković (University of Belgrade, Serbia)  
 Maciej Piasecki (Wroclaw University of Science and Technology, Poland)  
 Paul Piwek (The Open University, United Kingdom)  
 Alistair Plum (University of Wolverhampton, United Kingdom)  
 Alberto Poncelas (Dublin City University, Ireland)

Alexander Popov (Bulgarian Academy of Sciences, Bulgaria)  
 Maja Popović (Dublin City University, Ireland)  
 Dan Povey (Johns Hopkins University, United States of America)  
 Ondřej Pražák (University of West Bohemia, Czech Republic)  
 Prokopis Prokopidis (Research and Innovation Center in Information, Greece)  
 Tharindu Ranasinghe (University of Wolverhampton, United Kingdom)  
 Natalia Resende (Dublin City University, Ireland)  
 Pattabhi RK Rao (Anna University, India)  
 Omid Rohanian (University of Wolverhampton, United Kingdom)  
 Josef Ruppenhofer (Institute for the German Language, Germany)  
 Pavel Rychlý (Masaryk University, Czech Republic)  
 Magdaléna Rysová (Charles University, Czech Republic)  
 Branislava Šandrih (Belgrade University, Serbia)  
 Estela Saquete (University of Alicante, Spain)  
 Leah Schaefer (Indiana University, United States)  
 Ineke Schuurman (University of Leuven, Belgium)  
 Olga Seminck (Paris Diderot University, France)  
 Nasredine Semmar (Laboratory for Integration of Systems and Technology, France)  
 Matthew Shardlow (Manchester Metropolitan University, United Kingdom)  
 Artem Shelmanov (Russian Academy of Sciences, Russia)  
 Dimitar Shterionov (Dublin City University, Ireland)  
 Jennifer Sikos (University of Stuttgart, Germany)  
 João Silva (University of Lisboa, Portugal)  
 Vasiliki Simaki (Lancaster University, United Kingdom)  
 Sunayana Sitaram (Microsoft Research, India)  
 Mihailo Skoric (Researcher, Serbia)  
 Felix Stahlberg (University of Cambridge, Department of Engineering, United Kingdom)  
 Kenneth Steimel (Indiana University, United States)  
 Sebastian Stüker (Karlsruhe Institute of Technology, Germany)  
 Yoshimi Suzuki (Shizuoka University, Japan)  
 Liling Tan (Nanyang Technological University, Singapore)  
 Segun Taofeek Aroyehun (National Polytechnic Institute, Mexico)  
 Laura Tološi (Self employed data scientist)  
 Elena Tutubalina (Kazan Federal University, Russia)  
 Eleni Tziafa (National and Kapodistrian University of Athens, Greece)  
 Antonio Valerio Miceli Barone (University of Edinburgh, United Kingdom)  
 Mihaela Vela (Saarland University, Germany)  
 Cristina Vertan (University of Hamburg, Germany)  
 Manuel Vilares Ferro (University of Vigo, Spain)  
 Veronika Vincze (University of Szeged, Hungary)  
 Pidong Wang (National University of Singapore, Singapore)  
 Michael Wiegand (Heidelberg University, Germany)  
 Victoria Yaneva (University of Wolverhampton, United Kingdom)  
 Kristina Yordanova (University of Rostock, Germany)  
 Juntao Yu (Queen Mary University of London, United Kingdom)  
 Wajdi Zaghouani (Hamad Bin Khalifa University, Qatar)  
 Kalliopi Zervanou (Eindhoven University of Technology, Netherlands)  
 Inès Zribi (University of Sfax, Tunisia)

## Table of Contents

|                                                                                                                                       |     |
|---------------------------------------------------------------------------------------------------------------------------------------|-----|
| <i>Table Structure Recognition Based on Cell Relationship, a Bottom-Up Approach</i>                                                   |     |
| Darshan Adiga, Shabir Ahmad Bhat, Muzaffar Bashir Shah and Viveka Vyeth .....                                                         | 1   |
| <i>Identification of Good and Bad News on Twitter</i>                                                                                 |     |
| Piush Aggarwal and Ahmet Aker .....                                                                                                   | 9   |
| <i>Bilingual Low-Resource Neural Machine Translation with Round-Tripping: The Case of Persian-Spanish</i>                             |     |
| Benyamin Ahmadnia and Bonnie Dorr .....                                                                                               | 18  |
| <i>Enhancing Phrase-Based Statistical Machine Translation by Learning Phrase Representations Using Long Short-Term Memory Network</i> |     |
| Benyamin Ahmadnia and Bonnie Dorr .....                                                                                               | 25  |
| <i>Automatic Propbank Generation for Turkish</i>                                                                                      |     |
| Koray Ak and Olcay Taner Yıldız .....                                                                                                 | 33  |
| <i>Multilingual Sentence-Level Bias Detection in Wikipedia</i>                                                                        |     |
| Desislava Aleksandrova, François Lareau and Pierre André Ménard .....                                                                 | 42  |
| <i>Supervised Morphological Segmentation Using Rich Annotated Lexicon</i>                                                             |     |
| Ebrahim Ansari, Zdeněk Žabokrtský, Mohammad Mahmoudi, Hamid Haghdoust and<br>Jonáš Vidra .....                                        | 52  |
| <i>Combining Lexical Substitutes in Neural Word Sense Induction</i>                                                                   |     |
| Nikolay Arefyev, Boris Sheludko and Alexander Panchenko .....                                                                         | 62  |
| <i>Detecting Clitics Related Orthographic Errors in Turkish</i>                                                                       |     |
| Uğurcan Arıkan, Onur Güngör and Suzan Uskudarlı .....                                                                                 | 71  |
| <i>Benchmark Dataset for Propaganda Detection in Czech Newspaper Texts</i>                                                            |     |
| Vít Baisa, Ondřej Herman and Ales Horak .....                                                                                         | 77  |
| <i>Diachronic Analysis of Entities by Exploiting Wikipedia Page Revisions</i>                                                         |     |
| Pierpaolo Basile, Annalina Caputo, Seamus Lawless and Giovanni Semeraro .....                                                         | 84  |
| <i>Using a Lexical Semantic Network for the Ontology Building</i>                                                                     |     |
| Nadia Bebeshina-Clairret, Sylvie Despres and Mathieu Lafourcade .....                                                                 | 92  |
| <i>Naive Regularizers for Low-Resource Neural Machine Translation</i>                                                                 |     |
| Meriem Beloucif, Ana Valeria Gonzalez, Marcel Bollmann and Anders Søgaard .....                                                       | 102 |
| <i>Exploring Graph-Algebraic CCG Combinators for Syntactic-Semantic AMR Parsing</i>                                                   |     |
| Sebastian Beschke .....                                                                                                               | 112 |
| <i>Quasi Bidirectional Encoder Representations from Transformers for Word Sense Disambiguation</i>                                    |     |
| Michele Bevilacqua and Roberto Navigli .....                                                                                          | 122 |
| <i>Evaluating the Consistency of Word Embeddings from Small Data</i>                                                                  |     |
| Jelke Bloem, Antske Fokkens and Aurélie Herbelot .....                                                                                | 132 |
| <i>Cross-Domain Training for Goal-Oriented Conversational Agents</i>                                                                  |     |
| Alexandra Maria Bodîrlău, Stefania Budulan and Traian Rebedea .....                                                                   | 142 |

|                                                                                                                                                                                                      |     |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| <i>Learning Sentence Embeddings for Coherence Modelling and Beyond</i><br>Tanner Bohn, Yining Hu, Jinhang Zhang and Charles Ling .....                                                               | 151 |
| <i>Risk Factors Extraction from Clinical Texts Based on Linked Open Data</i><br>Svetla Boytcheva, Galia Angelova and Zhivko Angelov .....                                                            | 161 |
| <i>Parallel Sentence Retrieval From Comparable Corpora for Biomedical Text Simplification</i><br>Rémi Cardon and Natalia Grabar .....                                                                | 168 |
| <i>Classifying Author Intention for Writer Feedback in Related Work</i><br>Arlene Casey, Bonnie Webber and Dorota Glowacka .....                                                                     | 178 |
| <i>Sparse Victory – A Large Scale Systematic Comparison of Count-Based and Prediction-Based Vectorizers for Text Classification</i><br>Rupak Chakraborty, Ashima Elhence and Kapil Arora .....       | 188 |
| <i>A Fine-Grained Annotated Multi-Dialectal Arabic Corpus</i><br>Anis Charfi, Wajdi Zaghouani, Syed Hassan Mehdi and Esraa Mohamed .....                                                             | 198 |
| <i>Personality-Dependent Neural Text Summarization</i><br>Pablo Costa and Ivandré Paraboni .....                                                                                                     | 205 |
| <i>Self-Adaptation for Unsupervised Domain Adaptation</i><br>Xia Cui and Danushka Bollegala .....                                                                                                    | 213 |
| <i>Speculation and Negation Detection in French Biomedical Corpora</i><br>Clément Dalloux, Vincent Claveau and Natalia Grabar .....                                                                  | 223 |
| <i>Porting Multilingual Morphological Resources to OntoLex-Lemon</i><br>Thierry Declerck and Stefania Racioppa .....                                                                                 | 233 |
| <i>Dependency-Based Self-Attention for Transformer NMT</i><br>Hiroyuki Deguchi, Akihiro Tamura and Takashi Ninomiya .....                                                                            | 239 |
| <i>Detecting Toxicity in News Articles: Application to Bulgarian</i><br>Yoan Dinkov, Ivan Koychev and Preslav Nakov .....                                                                            | 247 |
| <i>De-Identification of Emails: Pseudonymizing Privacy-Sensitive Data in a German Email Corpus</i><br>Elisabeth Eder, Ulrike Krieg-Holz and Udo Hahn .....                                           | 259 |
| <i>Lexical Quantile-Based Text Complexity Measure</i><br>Maksim Ereemeev and Konstantin Vorontsov .....                                                                                              | 270 |
| <i>Demo Application for LETO: Learning Engine Through Ontologies</i><br>Suilan Estevez-Velarde, Andrés Montoyo, Yudivian Almeida-Cruz, Yoan Gutiérrez, Alejandro Piad-Morffis and Rafael Muñoz ..... | 276 |
| <i>Sentence Simplification for Semantic Role Labelling and Information Extraction</i><br>Richard Evans and Constantin Orăsan .....                                                                   | 285 |
| <i>OloBot - Towards A Text-Based Arabic Health Conversational Agent: Evaluation and Results</i><br>Ahmed Fadhil and Ahmed AbuRa'ed .....                                                             | 295 |
| <i>Developing the Old Tibetan Treebank</i><br>Christian Faggionato and Marieke Meelen .....                                                                                                          | 304 |

|                                                                                                                                                   |     |
|---------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| <i>Summarizing Legal Rulings: Comparative Experiments</i>                                                                                         |     |
| Diego Feijo and Viviane Moreira .....                                                                                                             | 313 |
| <i>Entropy as a Proxy for Gap Complexity in Open Cloze Tests</i>                                                                                  |     |
| Mariano Felice and Paula Buttery .....                                                                                                            | 323 |
| <i>Song Lyrics Summarization Inspired by Audio Thumbnailing</i>                                                                                   |     |
| Michael Fell, Elena Cabrio, Fabien Gandon and Alain Giboin .....                                                                                  | 328 |
| <i>Comparing Automated Methods to Detect Explicit Content in Song Lyrics</i>                                                                      |     |
| Michael Fell, Elena Cabrio, Michele Corazza and Fabien Gandon .....                                                                               | 338 |
| <i>Linguistic Classification: Dealing Jointly with Irrelevance and Inconsistency</i>                                                              |     |
| Laura Franzoi, Andrea Sgarro, Anca Dinu and Liviu P. Dinu .....                                                                                   | 345 |
| <i>Corpus Lexicography in a Wider Context</i>                                                                                                     |     |
| Chen Gafni .....                                                                                                                                  | 353 |
| <i>A Universal System for Automatic Text-to-Phonetics Conversion</i>                                                                              |     |
| Chen Gafni .....                                                                                                                                  | 360 |
| <i>Two Discourse Tree - Based Approaches to Indexing Answers</i>                                                                                  |     |
| Boris Galitsky and Dmitry Ilvovsky .....                                                                                                          | 367 |
| <i>Discourse-Based Approach to Involvement of Background Knowledge for Question Answering</i>                                                     |     |
| Boris Galitsky and Dmitry Ilvovsky .....                                                                                                          | 373 |
| <i>On a Chatbot Providing Virtual Dialogues</i>                                                                                                   |     |
| Boris Galitsky, Dmitry Ilvovsky and Elizaveta Goncharova .....                                                                                    | 382 |
| <i>Assessing Socioeconomic Status of Twitter Users: A Survey</i>                                                                                  |     |
| Dhouha Ghazouani, Luigi Lancieri, Habib Ounelli and Chaker Jebbari .....                                                                          | 388 |
| <i>Divide and Extract – Disentangling Clause Splitting and Proposition Extraction</i>                                                             |     |
| Darina Gold and Torsten Zesch .....                                                                                                               | 399 |
| <i>Sparse Coding in Authorship Attribution for Polish Tweets</i>                                                                                  |     |
| Piotr Grzybowski, Ewa Juralewicz and Maciej Piasecki .....                                                                                        | 409 |
| <i>Automatic Question Answering for Medical MCQs: Can It Go Further than Information Retrieval?</i>                                               |     |
| Le An Ha and Victoria Yaneva .....                                                                                                                | 418 |
| <i>Self-Knowledge Distillation in Natural Language Processing</i>                                                                                 |     |
| Sangchul Hahn and Heeyoul Choi .....                                                                                                              | 424 |
| <i>From the Paft to the Fiiture: A Fully Automatic NMT and Word Embeddings Method for OCR Post-Correction</i>                                     |     |
| Mika Härmäläinen and Simon Hengchen .....                                                                                                         | 432 |
| <i>Investigating Terminology Translation in Statistical and Neural Machine Translation: A Case Study on English-to-Hindi and Hindi-to-English</i> |     |
| Rejwanul Haque, Md Hasanuzzaman and Andy Way .....                                                                                                | 438 |
| <i>Beyond English-Only Reading Comprehension: Experiments in Zero-Shot Multilingual Transfer for Bulgarian</i>                                    |     |
| Momchil Hardalov, Ivan Koychev and Preslav Nakov .....                                                                                            | 448 |

|                                                                                                                          |     |
|--------------------------------------------------------------------------------------------------------------------------|-----|
| <i>Tweaks and Tricks for Word Embedding Disruptions</i>                                                                  |     |
| Amir Hazem and Nicolas Hernandez .....                                                                                   | 461 |
| <i>Meta-Embedding Sentence Representation for Textual Similarity</i>                                                     |     |
| Amir Hazem and Nicolas Hernandez .....                                                                                   | 466 |
| <i>Emoji Powered Capsule Network to Detect Type and Target of Offensive Posts in Social Media</i>                        |     |
| Hansi Hettiarachchi and Tharindu Ranasinghe .....                                                                        | 475 |
| <i>EoANN: Lexical Semantic Relation Classification Using an Ensemble of Artificial Neural Networks</i>                   |     |
| Rayehe Hosseini Pour and Mehrnoush Shamsfard .....                                                                       | 482 |
| <i>Opinions Summarization: Aspect Similarity Recognition Relaxes the Constraint of Predefined Aspects</i>                |     |
| Nguyen Huy Tien, Le Tung Thanh and Nguyen Minh Le .....                                                                  | 488 |
| <i>Discourse-Aware Hierarchical Attention Network for Extractive Single-Document Summarization</i>                       |     |
| Tatsuya Ishigaki, Hidetaka Kamigaito, Hiroya Takamura and Manabu Okumura .....                                           | 498 |
| <i>Semi-Supervised Induction of POS-Tag Lexicons with Tree Models</i>                                                    |     |
| Maciej Janicki .....                                                                                                     | 508 |
| <i>Word Sense Disambiguation Based on Constrained Random Walks in Linked Semantic Networks</i>                           |     |
| Arkadiusz Janz and Maciej Piasecki .....                                                                                 | 517 |
| <i>Classification of Micro-Texts Using Sub-Word Embeddings</i>                                                           |     |
| Mihir Joshi and Nur Zincir-Heywood .....                                                                                 | 527 |
| <i>Using Syntax to Resolve NPE in English</i>                                                                            |     |
| Payal Khullar, Allen Antony and Manish Shrivastava .....                                                                 | 535 |
| <i>Is Similarity Visually Grounded? Computational Model of Similarity for the Estonian Language</i>                      |     |
| Claudia Kittask and Eduard Barbu .....                                                                                   | 542 |
| <i>Language-Agnostic Twitter-Bot Detection</i>                                                                           |     |
| Jürgen Knauth .....                                                                                                      | 551 |
| <i>Multi-Level Analysis and Recognition of the Text Sentiment on the Example of Consumer Opinions</i>                    |     |
| Jan Kocoń, Monika Zaśko-Zielińska and Piotr Miłkowski .....                                                              | 560 |
| <i>A Qualitative Evaluation Framework for Paraphrase Identification</i>                                                  |     |
| Venelin Kovatchev, M. Antònia Martí, Maria Salamo and Javier Beltran .....                                               | 569 |
| <i>Study on Unsupervised Statistical Machine Translation for Backtranslation</i>                                         |     |
| Anush Kumar, Nihal V. Nayak, Aditya Chandra and Mydhili K. Nair .....                                                    | 579 |
| <i>Towards Functionally Similar Corpus Resources for Translation</i>                                                     |     |
| Maria Kunilovskaya and Serge Sharoff .....                                                                               | 584 |
| <i>Question Similarity in Community Question Answering: A Systematic Exploration of Preprocessing Methods and Models</i> |     |
| Florian Kunneman, Thiago Castro Ferreira, Emiel Krahmer and Antal van den Bosch .....                                    | 594 |
| <i>A Classification-Based Approach to Cognate Detection Combining Orthographic and Semantic Similarity Information</i>   |     |
| Sofie Labat and Els Lefever .....                                                                                        | 603 |

|                                                                                                                                                                                                                                                                  |     |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| <i>Resolving Pronouns for a Resource-Poor Language, Malayalam Using Resource-Rich Language, Tamil.</i><br>Sobha Lalitha Devi .....                                                                                                                               | 612 |
| <i>Semantic Role Labeling with Pretrained Language Models for Known and Unknown Predicates</i><br>Daniil Larionov, Artem Shelmanov, Elena Chistova and Ivan Smirnov .....                                                                                        | 620 |
| <i>A Structural Approach to Enhancing WordNet with Conceptual Frame Semantics</i><br>Svetlozara Leseva and Ivelina Stoyanova .....                                                                                                                               | 630 |
| <i>Compositional Hyponymy with Positive Operators</i><br>Martha Lewis .....                                                                                                                                                                                      | 639 |
| <i>The Impact of Semantic Linguistic Features in Relation Extraction: A Logical Relational Learning Approach</i><br>Rinaldo Lima, Bernard Espinasse and Frederico Freitas .....                                                                                  | 649 |
| <i>Detecting Anorexia in Spanish Tweets</i><br>Pilar López Úbeda, Flor Miriam Plaza del Arco, Manuel Carlos Díaz Galiano, L. Alfonso Urena Lopez and Maite Martin .....                                                                                          | 656 |
| <i>A Type-Theoretical Reduction of Morphological, Syntactic and Semantic Compositionality to a Single Level of Description</i><br>Erkki Luuk .....                                                                                                               | 665 |
| <i>v-trel: Vocabulary Trainer for Tracing Word Relations - An Implicit Crowdsourcing Approach</i><br>Verena Lyding, Christos Rodosthenous, Federico Sangati, Umair ul Hassan, Lionel Nicolas, Alexander König, Jolita Horbacauskienė and Anisia Katinskaia ..... | 675 |
| <i>Jointly Learning Author and Annotated Character N-gram Embeddings: A Case Study in Literary Text</i><br>Suraj Maharjan, Deepthi Mave, Prasha Shrestha, Manuel Montes, Fabio A. González and Thamar Solorio .....                                              | 685 |
| <i>Generating Challenge Datasets for Task-Oriented Conversational Agents through Self-Play</i><br>Sourabh Majumdar, Serra Sinem Tekiroglu and Marco Guerini .....                                                                                                | 694 |
| <i>Sentiment Polarity Detection in Azerbaijani Social News Articles</i><br>Sevda Mammadli, Shamsaddin Huseynov, Huseyn Alkaramov, Ulviyya Jafarli, Umid Suleymanov and Samir Rustamov .....                                                                      | 704 |
| <i>Inforex — a Collaborative System for Text Corpora Annotation and Analysis Goes Open</i><br>Michał Marcińczuk and Marcin Oleksy .....                                                                                                                          | 712 |
| <i>Semantic Language Model for Tunisian Dialect</i><br>Abir Masmoudi, Rim Laatar, Mariem Ellouze and Lamia Hadrich Belguith .....                                                                                                                                | 721 |
| <i>Automatic Diacritization of Tunisian Dialect Text Using Recurrent Neural Network</i><br>Abir Masmoudi, Mariem Ellouze and Lamia Hadrich Belguith .....                                                                                                        | 731 |
| <i>Comparing MT Approaches for Text Normalization</i><br>Claudia Matos Veliz, Orphee De Clercq and Veronique Hoste .....                                                                                                                                         | 741 |
| <i>Sentiment and Emotion Based Representations for Fake Reviews Detection</i><br>Alimuddin Melleng, Anna Jurek-Loughrey and Deepak P. ....                                                                                                                       | 751 |
| <i>Turning Silver into Gold: Error-Focused Corpus Reannotation with Active Learning</i><br>Pierre André Ménard and Antoine Mougeot .....                                                                                                                         | 759 |

|                                                                                                                                  |     |
|----------------------------------------------------------------------------------------------------------------------------------|-----|
| <i>NLP Community Perspectives on Replicability</i>                                                                               |     |
| Margot Mieskes, Karën Fort, Aurélie Névél, Cyril Grouin and Kevin Cohen . . . . .                                                | 769 |
| <i>Unsupervised Data Augmentation for Less-Resourced Languages with no Standardized Spelling</i>                                 |     |
| Alice Millour and Karën Fort . . . . .                                                                                           | 777 |
| <i>Neural Feature Extraction for Contextual Emotion Detection</i>                                                                |     |
| Elham Mohammadi, Hessam Amini and Leila Kosseim . . . . .                                                                        | 786 |
| <i>Empirical Study of Diachronic Word Embeddings for Scarce Data</i>                                                             |     |
| Syrielle Montariol and Alexandre Allauzen . . . . .                                                                              | 796 |
| <i>A Fast and Accurate Partially Deterministic Morphological Analysis</i>                                                        |     |
| Hajime Morita and Tomoya Iwakura . . . . .                                                                                       | 805 |
| <i>incom.py - A Toolbox for Calculating Linguistic Distances and Asymmetries between Related Languages</i>                       |     |
| Marius Mosbach, Irina Stenger, Tania Avgustinova and Dietrich Klakow . . . . .                                                   | 811 |
| <i>A Holistic Natural Language Generation Framework for the Semantic Web</i>                                                     |     |
| Axel-Cyrille Ngonga Ngomo, Diego Moussallem and Lorenz Bühmann . . . . .                                                         | 820 |
| <i>Building a Comprehensive Romanian Knowledge Base for Drug Administration</i>                                                  |     |
| Bogdan Nicula, Mihai Dascalu, Maria-Dorinela Sîrbu, Ștefan Trăușan-Matu and Alexandru Nuță . . . . .                             | 830 |
| <i>Summary Refinement through Denoising</i>                                                                                      |     |
| Nikola Nikolov, Alessandro Calmanovici and Richard Hahnloser . . . . .                                                           | 838 |
| <i>Large-Scale Hierarchical Alignment for Data-Driven Text Rewriting</i>                                                         |     |
| Nikola Nikolov and Richard Hahnloser . . . . .                                                                                   | 845 |
| <i>Dependency-Based Relative Positional Encoding for Transformer NMT</i>                                                         |     |
| Yutaro Omote, Akihiro Tamura and Takashi Ninomiya . . . . .                                                                      | 855 |
| <i>From Image to Text in Sentiment Analysis via Regression and Deep Learning</i>                                                 |     |
| Daniela Onita, Liviu P. Dinu and Adriana Birlutiu . . . . .                                                                      | 863 |
| <i>Building a Morphological Analyser for Laz</i>                                                                                 |     |
| Esra Önal and Francis Tyers . . . . .                                                                                            | 870 |
| <i>Term Based Semantic Clusters for Very Short Text Classification</i>                                                           |     |
| Jasper Paalman, Shantanu Mullick, Kalliopi Zervanou, Yingqian Zhang . . . . .                                                    | 879 |
| <i>Quotation Detection and Classification with a Corpus-Agnostic Model</i>                                                       |     |
| Sean Papay and Sebastian Padó . . . . .                                                                                          | 889 |
| <i>Validation of Facts Against Textual Sources</i>                                                                               |     |
| Vamsi Krishna Pendyala, Simran Sinha, Satya Prakash, Shriya Reddy and Anupam Jamatia . . .                                       | 896 |
| <i>A Neural Network Component for Knowledge-Based Semantic Representations of Text</i>                                           |     |
| Alejandro Piad-Morffis, Rafael Muñoz, Yudivian Almeida-Cruz, Yoan Gutiérrez, Suilan Estevez-Velarde and Andrés Montoyo . . . . . | 905 |
| <i>Toponym Detection in the Bio-Medical Domain: A Hybrid Approach with Deep Learning</i>                                         |     |
| Alistair Plum, Tharindu Ranasinghe and Constantin Orăsan . . . . .                                                               | 913 |

|                                                                                                                      |      |
|----------------------------------------------------------------------------------------------------------------------|------|
| <i>Combining SMT and NMT Back-Translated Data for Efficient NMT</i>                                                  |      |
| Alberto Poncelas, Maja Popović, Dimitar Shterionov, Gideon Maillette de Buy Wenniger and Andy Way .....              | 923  |
| <i>Unsupervised Dialogue Intent Detection via Hierarchical Topic Model</i>                                           |      |
| Artem Popov, Victor Bulatov, Darya Polyudova and Eugenia Veselova .....                                              | 933  |
| <i>Graph Embeddings for Frame Identification</i>                                                                     |      |
| Alexander Popov and Jennifer Sikos .....                                                                             | 940  |
| <i>Know Your Graph. State-of-the-Art Knowledge-Based WSD</i>                                                         |      |
| Alexander Popov, Kiril Simov and Petya Osenova .....                                                                 | 950  |
| <i>Are Ambiguous Conjunctions Problematic for Machine Translation?</i>                                               |      |
| Maja Popović and Sheila Castilho .....                                                                               | 960  |
| <i>ULSAna: Universal Language Semantic Analyzer</i>                                                                  |      |
| Ondřej Pražák and Miloslav Konopík .....                                                                             | 968  |
| <i>Machine Learning Approach to Fact-Checking in West Slavic Languages</i>                                           |      |
| Pavel Přibáň, Tomáš Hercig and Josef Steinberger .....                                                               | 974  |
| <i>NE-Table: A Neural Key-Value Table for Named Entities</i>                                                         |      |
| Janarthanan Rajendran, Jatin Ganhotra, Xiaoxiao Guo, Mo Yu, Satinder Singh and Lazaros Polymenakos .....             | 981  |
| <i>Enhancing Unsupervised Sentence Similarity Methods with Deep Contextualised Word Representations</i>              |      |
| Tharindu Ranasinghe, Constantin Orăsan and Ruslan Mitkov .....                                                       | 995  |
| <i>Semantic Textual Similarity with Siamese Neural Networks</i>                                                      |      |
| Tharindu Ranasinghe, Constantin Orăsan and Ruslan Mitkov .....                                                       | 1005 |
| <i>Analysing the Impact of Supervised Machine Learning on Automatic Term Extraction: HAMLET vs TermoStat</i>         |      |
| Ayla Rigouts Terryn, Patrick Drouin, Veronique Hoste and Els Lefever .....                                           | 1013 |
| <i>Distant Supervision for Sentiment Attitude Extraction</i>                                                         |      |
| Nicolay Rusnachenko, Natalia Loukachevitch and Elena Tutubalina .....                                                | 1023 |
| <i>Self-Attentional Models Application in Task-Oriented Dialogue Generation Systems</i>                              |      |
| Mansour Saffar Mehrjardi, Amine Trabelsi and Osmar R. Zaiane .....                                                   | 1032 |
| <i>Whom to Learn From? Graph- vs. Text-Based Word Embeddings</i>                                                     |      |
| Małgorzata Salawa, António Branco, Ruben Branco, João António Rodrigues and Chakaveh Saedi .....                     | 1042 |
| <i>Persistence Pays Off: Paying Attention to What the LSTM Gating Mechanism Persists</i>                             |      |
| Giancarlo Salton and John Kelleher .....                                                                             | 1053 |
| <i>Development and Evaluation of Three Named Entity Recognition Systems for Serbian - The Case of Personal Names</i> |      |
| Branislava Šandrih, Cvetana Krstev and Ranka Stankovic .....                                                         | 1061 |
| <i>Moral Stance Recognition and Polarity Classification from Twitter and Elicited Text</i>                           |      |
| Wesley Santos and Ivandré Paraboni .....                                                                             | 1070 |

|                                                                                                                                                                                                   |      |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| <i>The "Jump and Stay" Method to Discover Proper Verb Centered Constructions in Corpus Lattices</i><br>Bálint Sass .....                                                                          | 1077 |
| <i>Offence in Dialogues: A Corpus-Based Study</i><br>Johannes Schäfer and Ben Burtenshaw .....                                                                                                    | 1086 |
| <i>EmoTag – Towards an Emotion-Based Analysis of Emojis</i><br>Abu Awal Md Shoeb, Shahab Raji and Gerard de Melo .....                                                                            | 1095 |
| <i>A Morpho-Syntactically Informed LSTM-CRF Model for Named Entity Recognition</i><br>Lilia Simeonova, Kiril Simov, Petya Osenova and Preslav Nakov .....                                         | 1105 |
| <i>Named Entity Recognition in Information Security Domain for Russian</i><br>Anastasiia Sirotina and Natalia Loukachevitch .....                                                                 | 1115 |
| <i>Cross-Family Similarity Learning for Cognate Identification in Low-Resource Languages</i><br>Eliel Soisalon-Soininen and Mark Granroth-Wilding .....                                           | 1122 |
| <i>Automatic Detection of Translation Direction</i><br>Ilia Sominsky and Shuly Wintner .....                                                                                                      | 1132 |
| <i>Automated Text Simplification as a Preprocessing Step for Machine Translation into an Under-Resourced Language</i><br>Sanja Štajner and Maja Popović .....                                     | 1142 |
| <i>Investigating Multilingual Abusive Language Detection: A Cautionary Tale</i><br>Kenneth Steimel, Daniel Dakota, Yue Chen and Sandra Kübler .....                                               | 1152 |
| <i>Augmenting a BiLSTM Tagger with a Morphological Lexicon and a Lexical Category Identification Step</i><br>Steinþór Steingrímsson, Örvar Kárasen and Hrafn Loftsson .....                       | 1162 |
| <i>Comparison of Machine Learning Approaches for Industry Classification Based on Textual Descriptions of Companies</i><br>Andrey Tagarev, Nikola Tulechki and Svetla Boytcheva .....             | 1170 |
| <i>A Quantum-Like Approach to Word Sense Disambiguation</i><br>Fabio Tamburini .....                                                                                                              | 1177 |
| <i>Understanding Neural Machine Translation by Simplification: The Case of Encoder-Free Models</i><br>Gongbo Tang, Rico Sennrich and Joakim Nivre .....                                           | 1187 |
| <i>Text-Based Joint Prediction of Numeric and Categorical Attributes of Entities in Knowledge Bases</i><br>V Thejas, Abhijeet Gupta and Sebastian Padó .....                                      | 1195 |
| <i>SenZi: A Sentiment Analysis Lexicon for the Latinised Arabic (Arabizi)</i><br>Taha Tobaili, Miriam Fernandez, Harith Alani, Sanaa Sharafeddine, Hazem Hajj and Goran Glavaš .....              | 1204 |
| <i>Mining the UK Web Archive for Semantic Change Detection</i><br>Adam Tsakalidis, Marya Bazzi, Mihai Cucuringu, Pierpaolo Basile and Barbara McGillivray .....                                   | 1213 |
| <i>Cross-Lingual Word Embeddings for Morphologically Rich Languages</i><br>Ahmet Üstün, Gosse Bouma and Gertjan van Noord .....                                                                   | 1223 |
| <i>It Takes Nine to Smell a Rat: Neural Multi-Task Learning for Check-Worthiness Prediction</i><br>Slavena Vasileva, Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño and Preslav Nakov ..... | 1230 |

|                                                                                                                                                                                                                                                                                      |      |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| <i>Deep Learning Contextual Models for Prediction of Sport Events Outcome from Sportsmen Interviews</i><br>Boris Velichkov, Ivan Koychev and Svetla Boytcheva .....                                                                                                                  | 1241 |
| <i>Exploiting Frame-Semantics and Frame-Semantic Parsing for Automatic Extraction of Typological Information from Descriptive Grammars of Natural Languages</i><br>Shafqat Mumtaz Virk, Azam Sheikh Muhammad, Lars Borin, Muhammad Irfan Aslam, Saania Iqbal and Nazia Khurram ..... | 1248 |
| <i>Exploiting Open IE for Deriving Multiple Premises Entailment Corpus</i><br>Martin Vítá and Jakub Klímek .....                                                                                                                                                                     | 1258 |
| <i>Towards Adaptive Text Summarization: How Does Compression Rate Affect Summary Readability of L2 Texts?</i><br>Tatiana Vodolazova and Elena Lloret .....                                                                                                                           | 1266 |
| <i>The Impact of Rule-Based Text Generation on the Quality of Abstractive Summaries</i><br>Tatiana Vodolazova and Elena Lloret .....                                                                                                                                                 | 1276 |
| <i>ETNLP: A Visual-Aided Systematic Approach to Select Pre-Trained Embeddings for a Downstream Task</i><br>Son Vu Xuan, Thanh Vu, Son Tran and Lili Jiang .....                                                                                                                      | 1286 |
| <i>Tagger for Polish Computer Mediated Communication Texts</i><br>Wiktor Walentynowicz, Maciej Piasecki and Marcin Oleksy .....                                                                                                                                                      | 1296 |
| <i>Evaluation of Vector Embedding Models in Clustering of Text Documents</i><br>Tomasz Walkowiak and Mateusz Gniewkowski .....                                                                                                                                                       | 1305 |
| <i>Bigger versus Similar: Selecting a Background Corpus for First Story Detection Based on Distributional Similarity</i><br>Fei Wang, Robert J. Ross and John D. Kelleher .....                                                                                                      | 1313 |
| <i>Predicting Sentiment of Polish Language Short Texts</i><br>Aleksander Wawer and Julita Sobiczewska .....                                                                                                                                                                          | 1322 |
| <i>Improving Named Entity Linking Corpora Quality</i><br>Albert Weichselbraun, Adrian M.P. Brasoveanu, Philipp Kuntschik and Lyndon J.B. Nixon ..                                                                                                                                    | 1329 |
| <i>Sequential Graph Dependency Parser</i><br>Sean Welleck and Kyunghyun Cho .....                                                                                                                                                                                                    | 1339 |
| <i>Term-Based Extraction of Medical Information: Pre-Operative Patient Education Use Case</i><br>Martin Wolf, Volha Petukhova and Dietrich Klakow .....                                                                                                                              | 1347 |
| <i>A Survey of the Perceived Text Adaptation Needs of Adults with Autism</i><br>Victoria Yaneva, Constantin Orăsan, Le An Ha and Natalia Ponomareva .....                                                                                                                            | 1357 |
| <i>An Open, Extendible, and Fast Turkish Morphological Analyzer</i><br>Olçay Taner Yıldız, Begüm Avar and Gökhan Ercan .....                                                                                                                                                         | 1365 |
| <i>Self-Attention Networks for Intent Detection</i><br>Sevinj Yolchuyeva, Géza Németh and Bálint Gyires-Tóth .....                                                                                                                                                                   | 1374 |
| <i>Turkish Tweet Classification with Transformer Encoder</i><br>Atif Emre Yüksel, Yaşar Alim Türkmen, Arzucan Özgür and Berna Altınel .....                                                                                                                                          | 1381 |

|                                                                                              |      |
|----------------------------------------------------------------------------------------------|------|
| <i>Multilingual Dynamic Topic Model</i>                                                      |      |
| Elaine Zosa and Mark Granroth-Wilding .....                                                  | 1389 |
| <i>A Wide-Coverage Context-Free Grammar for Icelandic and an Accompanying Parsing System</i> |      |
| Vilhjálmur Þorsteinsson, Hulda Óladóttir and Hrafn Loftsson .....                            | 1398 |

# Linguistic Classification: Dealing Jointly with Irrelevance and Inconsistency

**Laura Franzoi**  
Faculty of Mathematics  
and Computer Science  
University of Bucharest  
laura.franzoi@gmail.com

**Andrea Sgarro**  
DMG  
University of Trieste  
sgarro@units.it

**Anca Dinu**  
Faculty of  
Foreign Languages  
and Literatures  
University of Bucharest  
ancaddinu@gmail.com

**Liviu P. Dinu**  
Faculty of Mathematics  
and Computer Science  
University of Bucharest  
liviu.p.dinu@gmail.com

## Abstract

In this paper we present new methods for language classification which put to good use both syntax and fuzzy tools, and are capable of dealing with irrelevant linguistic features (i.e. features which should not contribute to the classification) and even inconsistent features (which do not make sense for specific languages). We introduce a metric distance, based on the generalized Steinhaus transform, which allows one to deal jointly with irrelevance and inconsistency. To evaluate our methods, we test them on a syntactic data set, due to the linguist G. Longobardi and his school. We obtain phylogenetic trees which sometimes outperform the ones obtained by Atkinson and Gray (Gray and Atkinson, 2003; Bouckaert et al., 2012).

## 1 Introduction

According to Ethnologue (Eth, 2018), there are around 7000 living natural languages in the world, and one of the most interesting topics (not only in the academic field, but also in the general public) is their classification. While the comparative method was the main method of classifying natural languages until the 90s, the last decades brought an increasing number of computational approaches for estimating the historical evolution of languages and their relationships. Most of the computational historical linguistics approaches rely on the use of lexical items. In contrast, very few of them take into account syntactic aspects. Moreover, fuzzy tools and information theory were employed quite sparsely in language classification tasks (Ciobanu et al., 2018), in spite of the inherent fuzzy nature of the natural language data.

This paper is based on previous work on fuzzy string distances and linguistic classification started in (Franzoi and Sgarro, 2017a,b; Franzoi, 2017), and inspired by the path-breaking ideas put forward back in 1967 (Muljačić, 1967) by the Croat linguist Ž., Muljačić. The technical tool which will be used in this paper is the *general Steinhaus transform*, or *biotope transform*, applied to crisp strings which are however affected by irrelevance and inconsistency, as happens with data due to the linguist G. Longobardi and his school. Fuzziness in linguistics has been seldomly treated (Franzoi and Sgarro, 2017a,b; Dinu et al., 2018), as compared to crisp approaches.

In his 1967 paper Muljačić, even if only rather implicitly, had introduced what appears to us as a natural *fuzzy* generalization of crisp Hamming distances between binary strings of fixed length  $n$ , and this only two years after Zadeh's seminal work (Zadeh, 1965): the aim was showing that Dalmatic, now an extinct language, is a bridge between the Western group of Romance languages and the Eastern group, mainly Romanian. The situation is the following: Romance languages  $L, \Lambda, \dots$  are each described by means of  $n$  features, which can be present or absent, and so are encoded by string  $s(L) = \underline{x} = x_1 \dots x_n$ , where  $x_i$  is the truth value of the proposition *feature  $i$  is present in language  $L$* ; however, presence/absence is sometimes only vaguely defined and so each  $x = x_i$  is rather a truth value  $x \in [0, 1]$  in a multi-valued logic as is fuzzy logic;  $x = x_i$  is *crisp* only when either  $x = 0 = \text{false} = \text{absent}$  or  $x = 1 = \text{true} = \text{present}$ , else  $x$  is *strictly fuzzy*. So, the mathematical objects one deals with are *strings*  $\underline{x}, \underline{y}, \dots$  of length  $n$ , each of the  $n$  components being a real number in the interval  $[0, 1]$ , and moreover *distances* between such objects, since

the classifications are all distance-based. In what follows, rather than Muljačić distance, we need string distances obtained by use of the *Steinhaus transform*, cf. (Dinu et al., 2018), and the *generalized Steinhaus transform*; they are all *metric* distances, in particular they verify the triangle equality. Unlike the case of Muljačić distances, which span the interval  $[0, n]$ , these distances are *normalized* to the interval  $[0, 1]$ . Steinhaus transforms allow one to deal with *irrelevance* and *inconsistency* in linguistics, as we already argued in (Dinu et al., 2018), and not only with vagueness, or fuzziness, as in Muljačić case, cf. (Muljačić, 1967; Franzoi and Sgarro, 2017a); the reason to use the *generalized Steinhaus transform*, as we do here, is that it allows one to deal *jointly* with both irrelevance and inconsistency.

Based on arguments defended by the linguist G. Longobardi and his school, cf. (Bortolussi et al., 2011; Longobardi et al., 2016, 2013, 2015), if a feature  $i$  has a low truth value in two languages  $L$  and  $\Lambda$ , then that feature is scarcely relevant: in fact, in the practice of linguistics the values 0 and 1 have a very *asymmetric* use, and the fact that languages  $L$  and  $\Lambda$  both have zero in a position  $i$  means that such an irrelevant feature  $i$  should *not* really contribute to the distance between the two languages. Technically, one should move from Hamming distances to (normalized) Jaccard distances. To achieve the goal, the convenient tool we have used was the *Steinhaus transform*, cf. (Dinu et al., 2018), which is known to preserve metricity and which is general enough so as to amply cover also the fuzzy situation: one starts from a distance like Muljačić distance  $d_M(\underline{x}, \underline{y})$ , and obtains its Steinhaus transform, in this case a *fuzzy Jaccard distance*  $d_J(\underline{x}, \underline{y})$  for fuzzy strings  $\underline{x}$  and  $\underline{y}$ ; starting from the usual *crisp* Hamming distance the transform gives the usual *crisp* Jaccard distance.

In general, to apply a Steinhaus transformation one needs a *pivot string*, which in the Jaccard case is the all-0 string  $\underline{z} = \underline{0} = (0, \dots, 0)$ . In the transform, actually, any other string  $\underline{z}$  might be used, cf. (Dinu et al., 2018), as we do here so as to cover the case of *logical inconsistency*, as appears in the data due to G. Longobardi: his school is involved in an ambitious and innovative project on language classification based on *syntax*, cf. (Bor-

tolussi et al., 2011; Longobardi et al., 2016); languages are represented through yes-no strings of length 53, each string position corresponding to a syntactic feature which can be present or absent. In his notation Longobardi uses + if a feature is present, - if it is absent, 0 if it is undefined; in our case, cf. Tables 1, 2, we write 1 if a feature is present, 0 if it is absent, \* if it is undefined. Actually, due to a complex network of logical implications which constrain features, some positions might be undefined (logically inconsistent). For example, in Longobardi’s classification, feature 34 is defined if and only if feature 8 is set to + and either feature 9 is set to + or feature 18 is not set to + (or both); otherwise it will be “neutralized” (*inconsistent*)<sup>1</sup>. This property does not hold true for Ptg (Portuguese), OE (Old English) and Ice (Icelandic).

All this establishes an extremely complex network of logical dependencies in Longobardi’s data, and makes it necessary, if one wants to cover also this new intriguing facet, to suitably generalize crisp Hamming distances, or crisp Jaccard distances, respectively: in Longobardi’s approach, cf. (Bortolussi et al., 2011; Longobardi et al., 2016, 2013, 2015), the two distances for ternary strings one defines and uses are quite useful, but unfortunately they violate the triangle property, and so are not metric. In this paper we propose one *metric* alternative based on the generalized Steinhaus transform (or generalized biotope transform): the star \* will be replaced by the totally ambiguous truth value  $\frac{1}{2}$ , and the pivot strings in the transform will be given by the set compound by the all- $\frac{1}{2}$  string, i.e. the totally ambiguous string  $\underline{z} = (\frac{1}{2}, \dots, \frac{1}{2})$  (which stands for inconsistency) and all-0 string  $\underline{z} = (0, \dots, 0)$ , i.e. the totally false string, which

<sup>1</sup>Feature 34 stands for *checking possessives*: it opposes languages like French, wherein possessives occur without any visible article (*mon livre* vs. *le mon livre*), to those like Italian, in which a visible determiner is possible and normally required instead (*il mio libro* vs. *mio libro*). This feature seems to conceptually and typologically depend on full grammaticalization of definiteness (feature 8). Also, it is relevant only in languages with strong Person in D (feature 9) or without strong article (feature 18), because otherwise the language would have GenS with determiner-like function, cf. (Longobardi et al., 2013). Feature 8 asks if a language generalizes the overt marking of definiteness to all relevant cases. Feature 9 (*Strong Person*) defines whether attraction to the D area of referential nominal material (e.g. proper names) is overt (e.g. Romance) or not (e.g. English). Feature 18 (*Strong Article*) is presence of an indefinite article, i.e. of an obligatory marker on singular indefinite count argument nominals, distinct from those used for definite and mass indefinite, cf. (Longobardi et al., 2013).

stands for irrelevance. The idea is to play down not only the contribution of 0's and  $\frac{1}{2}$ 's separately, as we have done in (Dinu et al., 2018), but rather the contribution of both 0's and  $\frac{1}{2}$ 's *jointly*. It will turn out that in this case, which is not genuinely fuzzy, rather than to Muljačić distances, the generalized Steinhaus transform had been better applied to the usual *taxicab distance* (Manhattan distance, Minkowski distance), re-found when the standard fuzzy logical operators of *min* and *max* for conjunctions and disjunctions are replaced by Łukasiewicz T-norms and T-conorms, cf. (Franzoi and Sgarro, 2017b; Dinu et al., 2018).

The paper is divided as follow: in Section 2 we shortly re-take both fuzzy Hamming distances, or Muljačić distances, and taxicab distances stressing how the latter relate to Łukasiewicz T-norms; in Section 3 we introduce Steinhaus transform and we apply it to taxi-cab or Łukasiewicz distances; in Section 4 we introduce the general Steinhaus transform to deal with irrelevance and inconsistency *jointly* and we comment on our linguistic results; in Section 5 we sum up our results.

## 2 Fuzzy Hamming Distances vs. Łukasiewicz or Taxicab Distances

We need some notations and definitions: we set  $x \wedge y \doteq \min[x, y]$ ,  $x \vee y \doteq \max[x, y]$  and  $\bar{x} \doteq 1 - x$ ; these are the truth values of conjunction AND, disjunction OR and negation NOT, w.r. to propositions with truth values  $x$  and  $y$  in *standard fuzzy logic*, a relevant form of multi-valued logic;  $x \in [0, 1]$ . Define the *fuzziness* of the truth value  $x$  to be  $f(x) \doteq x \wedge (1 - x)$ . For the truth values  $x$  and  $y$  in  $[0, 1]$  we say that  $x$  and  $y$  are *consonant* if either  $x \vee y \leq \frac{1}{2}$  or  $x \wedge y \geq \frac{1}{2}$ , else they are *dissonant*; let  $\mathcal{D}$  and  $\mathcal{C}$  denote the set of dissonant and consonant positions  $i$ , respectively. We define the following distance for strings  $\underline{x}, \underline{y} \in [0, 1]^n$ :

$$d_M(\underline{x}, \underline{y}) \doteq \sum_{i \in \mathcal{D}} [1 - [f(x_i) \vee f(y_i)]] + \sum_{i \in \mathcal{C}} [f(x_i) \vee f(y_i)] \quad (1)$$

This expression stresses the link with *crisp* Hamming distances for binary strings  $\in \{0, 1\}^n$ , but its meaning is better understood due to the following fact: each of the  $n$  *additive* terms summed is the truth value of the statement:

*[(feature  $f_i$  is present in  $L$  and absent in  $\Lambda$ ) or (feature  $f_i$  is absent in  $L$  and present in  $\Lambda$ )]*

since, as soon proved, cf. e.g. (Franzoi and Sgarro, 2017a), for two truth values  $x$  and  $y$  one has  $(x \wedge \bar{y}) \vee (\bar{x} \wedge y)$  equal to  $f(x_i) \vee f(y_i)$  or to  $1 - [f(x_i) \vee f(y_i)]$  according whether there is consonance or dissonance. This distance, called henceforth Muljačić distance (and called *Sgarro distance* in (Deza and Deza, 2009), cf. also (Sgarro, 1977)) is simply a natural generalization of crisp Hamming distances to a fuzzy setting. As for alternative logical operators for conjunctions and disjunctions (different T-norms and T-conorms, for which cf. e.g. (Dubois et al., 2000)), they have been discussed in (Franzoi and Sgarro, 2017b). From a metric point of view, the only attractive choice, beside fuzzy Hamming distances, turned out to be Łukasiewicz T-norms for conjunctions and the corresponding T-conorms for disjunctions:

$$x \top y \doteq (x + y - 1) \vee 0, \quad x \perp y \doteq (x + y) \wedge 1$$

One soon checks that in this case, rather curiously,  $(x \top \bar{y}) \perp (\bar{x} \top y)$  turns out to be simply  $|x - y|$ , and so the string distance one obtains is nothing else but the very well-known taxicab distance  $d_T(\underline{x}, \underline{y}) = \sum_i |x_i - y_i|$ , which in our context, when it is applied to fuzzy strings of length  $n$ , might be also legitimately called *Łukasiewicz distance*.

If we consider the fuzziness  $f(x) \doteq d(x, x)$  of a logical value  $x$  and if we use the Muljačić distance, then we get  $f_M(x) = x \wedge (1 - x)$ ; if we use instead the Łukasiewicz distance, then the fuzziness is always 0.

However, if we consider another equally legitimate definition of fuzziness, namely “ambiguity - crispness”, which can be formalized as  $\frac{1}{2} - d(x, \frac{1}{2})$ , then if we use the Muljačić distance the new fuzziness is 0, but if we use the Łukasiewicz distance it is  $f_T(x) = \frac{1}{2} - d_T(x, \frac{1}{2}) = x \wedge (1 - x)$ : the result of the competition Muljačić distance vs. Łukasiewicz distance turns out to be a tie. In the next Section we explain why, with Longobardi's data, we decided to resort to taxicab distances.

The distance in (1) is a *fuzzy metric distance*, cf. (Sgarro, 1977; Franzoi and Sgarro, 2017a), from which a standard metric distance is soon obtained by imposing that self-distances  $d_M(\underline{x}, \underline{y})$  should be 0, while, unless  $\underline{x}$  is crisp (i.e. belong to  $\{0, 1\}^n$ , the set of the  $2^n$  binary strings of length  $n$ ), the value given by (1) would be strictly positive.

As for taxicab or Łukasiewicz distances, the self-distance  $d_T(\underline{x}, \underline{y})$  is always zero even when the argument  $\underline{x}$  is not crisp, a possibly unpleasant fact in a fuzzy context (but not in ours), as argued in (Franzoi and Sgarro, 2017b).

### 3 Steinhaus Transforms

In the general situation, one has objects  $x, y, \dots$ , not necessarily strings, a *metric* distance  $d(x, y)$  between objects, and a special object  $z$  called the “pivot-object”. The Steinhaus transform, cf. (Deza and Deza, 2009), itself proven to be a metric distance, is:

$$S_d(x, y) \doteq \frac{2d(x, y)}{d(x, y) + d(x, z) + d(y, z)}$$

set equal to zero when  $x = y = z$ .

In our case the objects are strings and *pivots*  $\underline{z}$  will always be *constant* strings  $\underline{z} = (z, \dots, z)$ ,  $z_i = z$ ,  $\forall i, z \in [0, 1]$ .

If one starts with the crisp Hamming distance, one obtains the usual crisp Jaccard distance (distances from the pivot are then Hamming *weights*); starting with the more general fuzzy Hamming distance, or Muljačić distance, one has an appropriate Jaccard-like generalization, which weighs only “little” a position where both  $x$  and  $y$  are “almost 0”, and which accounts for irrelevance in itself, but not for inconsistency, as instead we need.

If the term  $d_M(\underline{x}, \underline{z})$  is equal to the *fuzzy Hamming weight*  $w(\underline{x}) \doteq \sum_i x_i$  for  $\underline{z} = \underline{0}$ , it is equal to  $\frac{n}{2}$  independent of  $\underline{x}$  when  $\underline{z} = \underline{\frac{1}{2}}$ , a constant pivot string which we shall need to deal with inconsistency. The fact that  $d_M(\underline{x}, \underline{z})$  with  $\underline{z} = \underline{\frac{1}{2}}$  is independent of  $\underline{x}$  is a serious drawback, indeed. This is why in the case of Longobardi’s data, we have applied the Steinhaus transform, rather than to the fuzzy Hamming distance or Muljačić distance, directly to the taxicab distance or Łukasiewicz distance  $d_T(\underline{x}, \underline{y})$ . In this case, in the denominator of the corresponding Steinhaus transform, the fuzzy Hamming weight  $w(\underline{x})$  is replaced by  $d_T(\underline{x}, \underline{z}) = \sum_i |x_i - \frac{1}{2}|$ . In the next Section, more ambitiously, we shall deal *jointly* with both irrelevance and inconsistency.

### 4 Dealing with Irrelevance and Inconsistency

In (Franzoi and Sgarro, 2017a,b; Franzoi, 2017; Dinu et al., 2018) one has presented new methods for language classification, testing them on data

sets due to Muljačić and Longobardi. So far we have dealt separately with irrelevance and inconsistency, but a question arises spontaneously: can we consider jointly both irrelevance and inconsistency? Does a mathematical tool which takes into account both of them exist? The answer is yes and the tool we are looking for is the *generalized Steinhaus transform* or *generalized biotope transform*, cf. (Deza and Deza, 2009).

Prompted by arguments defended by G. Longobardi and his school, cf. (Bortolussi et al., 2011; Longobardi et al., 2016, 2013, 2015), the novelty of this section is that, since in the language classifications features can be irrelevant or inconsistent, we want to consider both aspects together.

As we said above the idea is to play down not only the contribution of 0’s, as in the case of irrelevance, but also the contribution of the  $\frac{1}{2}$ -positions. Unlike ours, Longobardi’s non-metric distance gets rid of irrelevant and inconsistent positions in quite a drastic way, possibly a serious draw-back, as we comment in our Conclusions.

The generalized Steinhaus transform, or generalized biotope transform, is:

$$S_d(x, y) = \frac{2d(x, y)}{d(x, y) + \inf_{z \in M} (d(x, z) + d(y, z))} \quad (2)$$

where  $M$  is the set of pivots we are considering, cf. (Deza and Deza, 2009).

We tackle Longobardi’s data (or rather to a sample of his languages, since the data he and his school are providing are steadily improving and extending), data which are not really fuzzy, even if we have decided to “simulate” logical inconsistency by *total fuzziness*. In this case the number of features is 53, and the languages are: Sic = Sicilian, Cal = Calabrese as spoken in South Italy, It = Italian, Sal = Salentin as spoken in Salento, South Italy, Sp = Spanish, Fr = French, Ptg = Portuguese, Rm = Romanian, Lat = Latin, CIG = Classical Attic Greek, NTG = New Testament Greek, BoG = Bova Greek as spoken in the village of Bova, Italy, Gri = Grico, a variant of Greek spoken in South Italy, Grk = Greek, Got = Gothic, OE = Old English, E = English, D = German, Da = Danish, Ice = Icelandic, Nor = Norwegian, Blg = Bulgarian, SC = Serbo Croatian, Slo = Slovenian, Po = Polish, Rus = Russian, Ir = Gaelic, Wel = Welsh, Far = Farsi, Ma = Marathi, Hi = Hindi, Ar = Arabic, Heb = Hebrew or ’ivrit, Hu = Hungarian, Finn = Finnish, StB = Standard Basque, WB = Western

Basque, Wo = Wolof as spoken mainly in Senegal. For comparison reasons, we have selected a part of Longobardi's data set compound by 38 languages; taking  $M = \left\{0, \frac{1}{2}\right\}$  in (2), the UPGMA tree we obtain is given in the following figure:

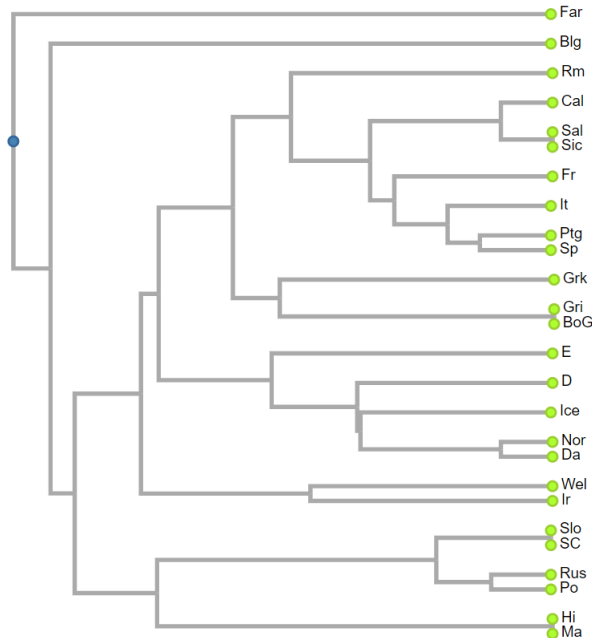


Figure 1: Generalized Steinhaus transform with taxi-cab distance and Longobardi's data

while the Longobardi's original tree is the following one:

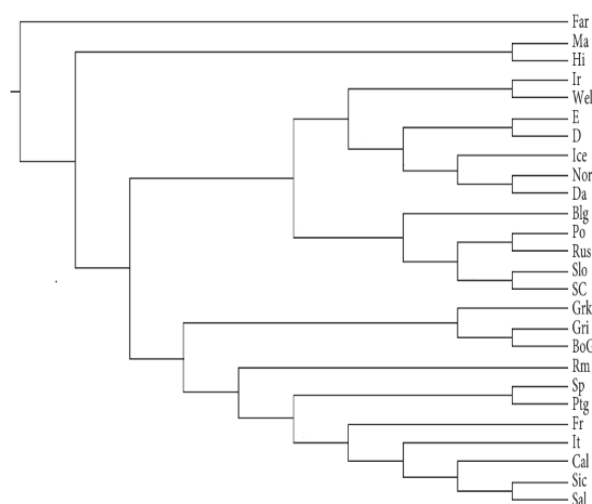


Figure 2: Longobardi's classification tree

We can observe that the Romance languages are grouped together. However there are some differences between the two trees: in our tree (Fig. 1) the big Romance languages (i.e. Italian, Spanish, Portuguese and French) are grouped together and

Italian is more integrated with the Ibero-Romance languages (i.e. Portuguese and Spanish), which are clustered together like in the standard language classifications. The three Italian dialects (i.e. Salentine, Sicilian and Calabrese) are external to this cluster in our case in Fig. 1, while in the original Longobardi's tree (Fig. 2) they are integrated with Italian and then the entire group is linked with French and after with the Ibero-Romance group. In both trees the Romanian is grouped with Romance languages, but is the most exterior with the languages from this group. In both trees the Celtic languages Gaelic (Ir) and Welsh (Wel) and Germanic languages are grouped together, but in the Longobardi's tree in Fig. 2 the Celtic group is more integrated with the Germanic group. There are two main differences between the two trees: the first one is that in Longobardi's tree in Fig. 2 Bulgarian is grouped with Slavic languages; the second one is the moving of the entire Slavic group from a closet proximity with the Germanic group (in the Longobardi's tree) to a more distance linkage with them in our case.

Our classification compares with the one obtained by Longobardi's school with these data, cf. comments in the Conclusion, where we argue why our distance is quite promising for the new and ambitious data Longobardi's school are now providing. Actually, our distance compares rather well also with the classification obtained by Q. D. Atkinson and R. D. Gray, cf. (Gray and Atkinson, 2003; Bouckaert et al., 2012).

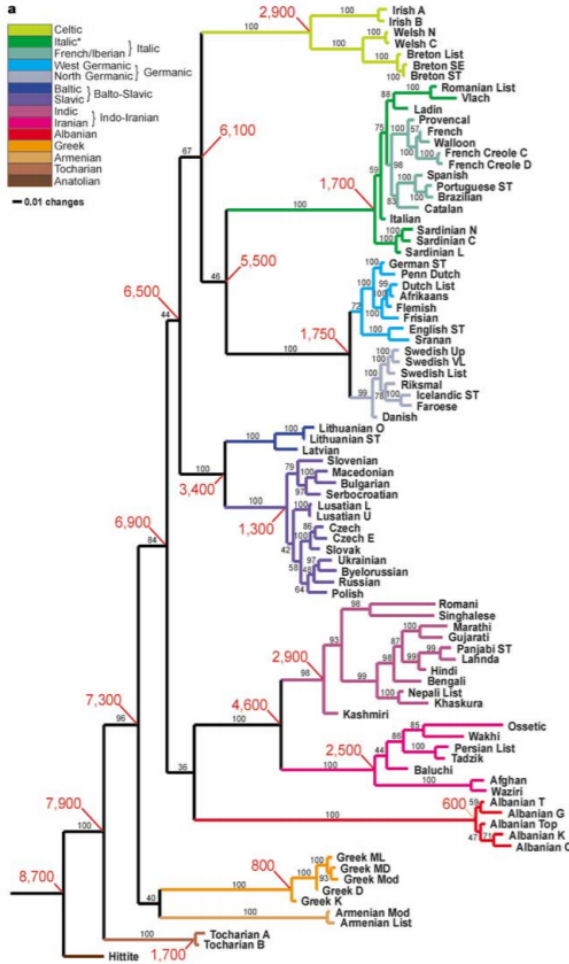


Figure 3: Q. D. Atkinson and R. D. Gray classification tree, cf. (Gray and Atkinson, 2003; Bouckaert et al., 2012)

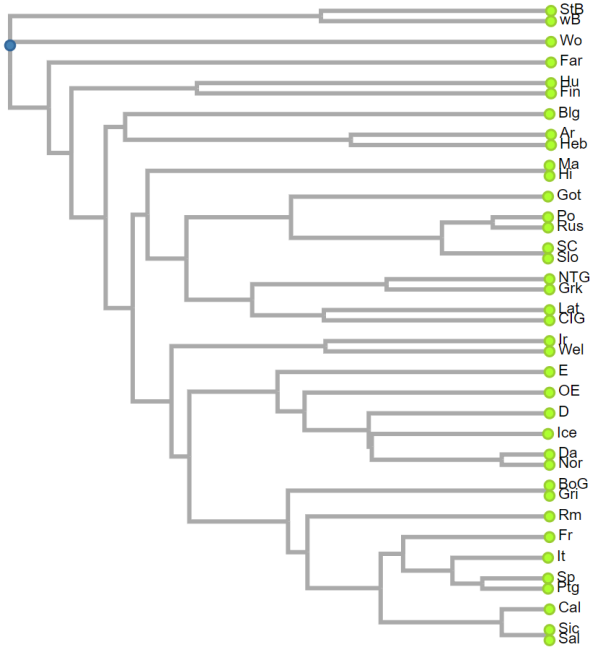


Figure 4: Classification obtained with the generalized Steinhaus transform applied to the taxi-cab distance with Longobardi's data

First of all for the classification we have used Longobardi's dataset, while Atkinson and Gray have used their own dataset. If we look to Marathi and Hindi we can notice that they are grouped together in both trees; also Polish, Russian, Serbo Croatian and Slovenian are grouped together in both trees; the same is for New Testament Greek, Greek and Classical Attic Greek. Also the Celtic languages (i.e. Gaelic and Welsh) and Germanic languages are grouped together. Our misclassification of Bulgarian is not that worrying, since Longobardi covers only the syntax of the noun, and the Bulgarian noun is well-known to behave in quite a non-Slavic way, due possibly to its Balcanian substratum.

## 5 Conclusions

In this paper we have investigated the language classification problem by using original tools inspired by fuzzy logic. In the literature fuzzy tools and information theory have been used only quite sparsely. We have exhibited a metric distance which allows one to deal jointly with both irrelevance and inconsistency, and which is based on the generalized Steinhaus transform. Our classification compares quite well both with the one obtained by Longobardi and the one obtained by Atkinson and Gray. The merits of our metric proposal should not be underestimated, as we now comment. In more recent datasets, Longobardi and his school introduce families and macrofamilies which are quite apart. Now, think of two languages  $L$  and  $\Lambda$  such that the following occurs (and this does occur with "remote" languages): in most position  $i$  at least one of the two languages has a star signalling non-definition of the corresponding features. Since such positions are totally ignored by Longobardi's non-metric distance, the value obtained for the distance relies on a handful of positions only, and it is no surprise that the two languages end up being poorly classified, a source of worry, indeed. Now, our metric distances are not that drastic, and so might be used as a sort of companion to Longobardi's non-metric distances, useful when the latter have a low significance due to the fact that only few features "survive".

We are confident that the fuzzy ideas and methods discussed in this paper and in (Franzoi and Sgarro, 2017a; Franzoi, 2017; Dinu et al., 2018) will prove to be useful not only in linguistic classification and linguistic phylogeny, but also outside

linguistic, first of all in coding theory cf. (Franzoi and Sgarro, 2017a), or even in bioinformatics.

Irrelevance and inconsistency appear to be features which are dealt with quite sparsely, if ever, outside Longobardi’s school; actually, these flexible features might prove to be quite useful not only in linguistic classification phylogeny, cf. (Franzoi and Sgarro, 2017a,b), but also in the investigation of the history of texts. So far, we are just providing technical tools to be used in Longobardi’s research, which, in its turn, is methodically matched with the *current state of the art*, cf. (Bortolussi et al., 2011; Longobardi et al., 2016, 2013, 2015; Longobardi, 2017; Kazakov et al., 2017).

Table 1: Longobardi original data

| ft. | Sic | Cal | It | Sal | Sp | Fr | Ptg | Rm | Lat | CIG | NtG | BoG | Gri | Grk | Got | OE | E | D | Da | Ice | Nor |
|-----|-----|-----|----|-----|----|----|-----|----|-----|-----|-----|-----|-----|-----|-----|----|---|---|----|-----|-----|
| 1.  | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 2.  | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 3.  | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 4.  | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 5.  | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 6.  | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 7.  | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 8.  | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 9.  | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 10. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 11. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 12. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 13. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 14. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 15. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 16. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 17. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 18. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 19. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 20. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 21. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 22. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 23. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 24. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 25. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 26. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 27. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 28. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 29. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 30. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 31. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 32. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 33. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 34. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 35. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 36. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 37. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 38. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 39. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 40. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 41. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 42. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 43. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 44. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 45. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 46. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 47. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 48. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 49. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 50. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |
| 51. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 52. | 1   | 1   | 1  | 1   | 1  | 1  | 1   | 1  | 1   | 1   | 1   | 1   | 1   | 1   | 1   | 1  | 1 | 1 | 1  | 1   | 1   |
| 53. | 0   | 0   | 0  | 0   | 0  | 0  | 0   | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0 | 0 | 0  | 0   | 0   |

Table 2: Longobardi original data

| ft. | Blg | SC | Slo | Po | Rus | Ir | Wel | Far | Ma | Hi | Ar | Heb | Hu | Fin | StB | wB | Wo |
|-----|-----|----|-----|----|-----|----|-----|-----|----|----|----|-----|----|-----|-----|----|----|
| 1.  | 1   | 1  | 1   | 1  | 1   | 1  | 1   | 1   | 1  | 1  | 1  | 1   | 1  | 1   | 1   | 1  | 1  |
| 2.  | 1   | 1  | 1   | 1  | 1   | 1  | 1   | 1   | 1  | 1  | 1  | 1   | 1  | 1   | 1   | 1  | 1  |
| 3.  | 1   | 1  | 1   | 1  | 1   | 1  | 1   | 0   | 0  | 1  | 1  | 1   | 0  | 0   | 0   | 1  | 1  |
| 4.  | 0   | 0  | 0   | 0  | 0   | 0  | 0   | 0   | 0  | 0  | 0  | 0   | 0  | 0   | 1   | 1  | 1  |
| 5.  | 1   | 1  | 1   | 1  | 1   | 1  | 1   | 1   | 1  | 1  | 1  | 1   | 1  | 1   | 0   | 0  | 0  |
| 6.  | 1   | 1  | 1   | 1  | 1   | 1  | 1   | 1   | 1  | 1  | 1  | 1   | 1  | 1   | *   | *  | *  |
| 7.  | 1   | 0  | 0   | 0  | 0   | 1  | 1   | 0   | 0  | 0  | 1  | 1   | 1  | 0   | 0   | 0  | 1  |
| 8.  | 1   | *  | *   | *  | *   | 1  | 1   | *   | *  | *  | 1  | 1   | 1  | *   | *   | *  | 1  |
| 9.  | 1   | *  | *   | *  | *   | 0  | 0   | *   | *  | *  | 1  | 1   | 1  | *   | *   | *  | *  |
| 10. | 0   | 0  | 0   | 0  | 0   | 0  | 0   | 0   | 0  | 0  | 0  | 0   | 0  | 1   | *   | *  | *  |
| 11. | 0   | *  | *   | *  | *   | 0  | 0   | *   | *  | *  | 0  | 0   | 0  | *   | 0   | 1  | 1  |
| 12. | 1   | *  | *   | *  | *   | 0  | 0   | *   | *  | *  | 0  | 0   | 0  | *   | *   | *  | 0  |
| 13. | *   | *  | *   | *  | *   | *  | *   | *   | *  | *  | *  | *   | *  | *   | *   | *  | *  |
| 14. | 0   | *  | *   | *  | *   | 0  | 0   | *   | *  | *  | 1  | 0   | 0  | *   | *   | *  | 1  |
| 15. | 0   | 0  | 0   | 0  | 0   | 0  | 0   | 0   | 0  | 0  | 1  | 0   | 0  | 0   | *   | *  | *  |
| 16. | 1   | 1  | 1   | 1  | 1   | 1  | 0   | 0   | 1  | 1  | 1  | 1   | 0  | 0   | *   | *  | *  |
| 17. | 0   | 0  | 0   | 0  | 0   | 0  | 0   | 1   | 1  | 1  | 0  | 0   | 1  | 0   | 0   | 0  | 1  |
| 18. | 0   | *  | *   | *  | *   | 0  | 0   | *   | *  | *  | 0  | 0   | *  | *   | *   | *  | *  |
| 19. | *   | *  | *   | *  | *   | *  | *   | 1   | 0  | 0  | *  | *   | 0  | *   | *   | *  | 0  |
| 20. | *   | *  | *   | *  | *   | *  | *   | *   | *  | *  | 0  | 0   | 0  | *   | 1   | 1  | *  |
| 21. | 1   | 1  | 1   | 1  | 1   | 1  | 1   | 1   | 1  | 1  | 1  | 1   | 1  | 1   | 1   | 1  | 0  |
| 22. | 1   | 1  | 1   | 1  | 1   | 1  | 0   | 0   | 1  | 1  | 1  | 1   | 0  | 1   | *   | *  | *  |
| 23. | 1   | 1  | 1   | 1  | 1   | 1  | 0   | 0   | 1  | 1  | 1  | 1   | 1  | 1   | 1   | 1  | 1  |
| 24. | *   | *  | *   | *  | *   | 0  | *   | *   | *  | *  | *  | *   | *  | *   | *   | *  | *  |
| 25. | 1   | 1  | 1   | 1  | 1   | 1  | 1   | 1   | 0  | 0  | 1  | 1   | 1  | 1   | 0   | 0  | 1  |
| 26. | *   | *  | *   | *  | *   | *  | *   | *   | 1  | 1  | *  | *   | *  | *   | 0   | 0  | *  |
| 27. | 0   | 0  | 0   | 0  | 0   | 0  | 0   | 0   | 1  | 1  | 1  | 0   | 0  | 0   | 0   | 0  | *  |
| 28. | 0   | 0  | 0   | 0  | 0   | 0  | 0   | 0   | 0  | 0  | 1  | 1   | 0  | 0   | 1   | 1  | *  |
| 29. | 1   | 0  | 0   | 0  | 0   | 1  | 1   | 1   | 1  | 1  | 1  | 1   | 0  | 0   | 1   | 1  | 0  |
| 30. | 0   | *  | *   | *  | *   | 0  | 0   | 0   | 0  | 0  | 0  | 0   | *  | *   | 0   | 0  | *  |
| 31. | 1   | *  | *   | *  | *   | 1  | 1   | 1   | *  | *  | 1  | 1   | *  | *   | *   | *  | *  |
| 32. | 0   | 1  | 1   | 1  | 1   | 1  | 1   | 1   | 0  | 0  | 1  | 1   | 0  | 0   | 0   | 0  | 1  |
| 33. | 0   | 0  | 0   | 0  | 0   | 0  | 0   | 0   | 0  | 0  | 0  | 0   | 1  | 1   | 0   | 0  | 0  |
| 34. | 0   | *  | *   | *  | *   | 1  | 1   | *   | *  | *  | 0  | 0   | 0  | *   | *   | *  | 0  |
| 35. | 1   | 1  | 1   | 1  | 1   | 0  | *   | *   | 1  | 1  | 0  | 0   | *  | 0   | *   | *  | 0  |
| 36. | 0   | *  | *   | *  | *   | 0  | *   | *   | 1  | 1  | *  | *   | *  | *   | *   | *  | *  |
| 37. | 1   | 1  | 1   | 0  | 1   | *  | 0   | 0   | 0  | 0  | *  | *   | *  | *   | 0   | 0  | *  |
| 38. | 0   | 0  | 0   | 0  | 0   | 0  | 0   | 0   | 0  | 0  | 1  | 1   | *  | *   | 0   | 0  | *  |
| 39. | 1   | 1  | 1   | 1  | 1   | 0  | 0   | 1   | 1  | 1  | 1  | 0   | 1  | 1   | 1   | 0  | 1  |
| 40. | 0   | 0  | 0   | 0  | 0   | *  | *   | 0   | 0  | 0  | 1  | *   | 0  | 0   | 0   | *  | 0  |
| 41. | 1   | *  | *   | *  | *   | *  | *   | *   | *  | *  | 0  | *   | 0  | *   | 1   | *  | 1  |
| 42. | 0   | 0  | 0   | 0  | 0   | 0  | 0   | 0   | 0  | 0  | 1  | 1   | 0  | 0   | 0   | 0  | 0  |
| 43. | 0   | 0  | 0   | 0  | 0   | 0  | 0   | 1   | 0  | 0  | *  | *   | 0  | 0   | 0   | 0  | 0  |
| 44. | 0   | 0  | 0   | 0  | 0   | 1  | 1   | *   | 0  | 0  | *  | *   | 0  | 0   | *   | *  | *  |
| 45. | 0   | 0  | 0   | 0  | 0   | *  | *   | *   | 0  | 0  | *  | *   | 0  | 0   | *   | *  | *  |
| 46. | 0   | 0  | 0   | 0  | 0   | *  | *   | *   | 0  | 0  | *  | *   | 0  | 0   | *   | *  | *  |
| 47. | *   | 1  | 1   | 1  | 1   | *  | *   | *   | *  | *  | *  | *   | *  | *   | *   | *  | 1  |
| 48. | 1   | *  | *   | *  | *   | *  | *   | *   | 1  | 1  | *  | *   | 1  | 1   | *   | *  | *  |
| 49. | 0   | 0  | 0   | 0  | 0   | *  | *   | *   | 0  | 0  | *  | *   | 0  | 0   | *   | *  | *  |
| 50. | 0   | 0  | 0   | 1  | 1   | *  | *   | *   | 0  | 0  | *  | *   | 0  | 0   | *   | *  | *  |
| 51. | *   | *  | *   | *  | *   | 0  | 0   | *   | *  | *  | 1  | 1   | *  | *   | *   | *  | 0  |
| 52. | *   | *  | *   | 0  | 0   | 0  | 0   | 1   | *  | *  | 0  | 0   | *  | *   | 0   | 0  | 1  |
| 53. | 0   | 0  | 0   | 0  | 0   | *  | *   | 0   | *  | *  | *  | *   | 1  | 1   | 1   | 1  | *  |

- M. M. Deza and E. Deza. 2009. *Encyclopedia of Distances*. Springer Dordrecht Heidelberg, London New York.
- A. Dinu, L. P. Dinu, L. Franzoi, and A. Sgarro. 2018. Steinhaus transforms of fuzzy string distances in computational linguistics. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations - 17th International Conference, IPMU 2018, Cádiz, Spain, June 11-15, 2018*. volume Proceedings, Part I, pages 171–182.
- D. Dubois, H. T. Nguyen, and H. Prade. 2000. Possibility theory, probability and fuzzy sets: Misunderstanding, bridges and gaps. In *Fundamentals of Fuzzy Sets*. Kluwer Academic Publishers, pages 343–438.
- L. Franzoi. 2017. Jaccard-like fuzzy distances for computational linguistics. In *19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2017, Timișoara, Romania, September 21-24, 2017*. pages 196–202.
- L. Franzoi and A. Sgarro. 2017a. Fuzzy hamming distinguishability. In *2017 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2017, Naples, Italy, July 9-12, 2017*. pages 1–6.
- L. Franzoi and A. Sgarro. 2017b. Linguistic classification: T-norms, fuzzy distances and fuzzy distinguishabilities. In *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference KES-2017, Marseille, France, 6-8 September 2017*. pages 1168–1177.
- R. D. Gray and Q. D. Atkinson. 2003. Language-tree divergence times support the anatolian theory of indo-european origin. *Nature* 426:435–439.
- D. Kazakov, G. Cordonì, A. Ceolin, M. A. Irìmia, S. Kim, D. Michelioudakis, N. Radkevich, C. Guardiano, and G. Longobardi. 2017. Machine learning models of universal grammar parameter dependencies. *Proceedings of Knowledge Resources for the Socio-Economic Sciences and Humanities associated with RANLP-17* pages 31–37.
- G. Longobardi. 2017. Principles, parameters, and schemata: A radically underspecified ug. *Linguistic Analysis* 41(3–4):517–557.
- G. Longobardi, A. Ceolin, L. Bortolussi, C. Guardiano, M. A. Irìmia, D. Michelioudakis, N. Radkevich, and A. Sgarro. 2016. Mathematical modeling of grammatical diversity supports the historical reality of formal syntax. *University of Tübingen, online publication system Tübingen DEU* pages 1–4.
- G. Longobardi, S. Ghirotto, C. Guardiano, F. Tassi, A. Benazzo, A. Ceolin, and G. Barbujan. 2015. Across language families: Genome diversity mirrors language variation within europe. *American Journal of Physical Anthropology* 157:630–640.
- G. Longobardi, C. Guardiano, G. Silvestri, A. Boatini, and A. Ceolin. 2013. Toward a syntactic phylogeny of modern indo-european languages. *Journal of Historical Linguistics* 3:11:122–152.
- Ž. Muljačić. 1967. Die Klassifikation der romanischen Sprachen. *Rom. Jahrbuch* 18 pages 23–37.
- A. Sgarro. 1977. A fuzzy hamming distance. *Bullettin Math. de la Soc. Sci. Math. de la R. S. de Roumanie* 69(1-2):137–144.
- L. A. Zadeh. 1965. Fuzzy sets. *Information and Control* 8(3):338–353.