

TLT 2019

**18th International Workshop on  
Treebanks and Linguistic Theories  
(TLT, SyntaxFest 2019)**

**Proceedings**

28–29 August, 2019  
held within the **SyntaxFest** 2019, 26–30 August  
Paris, France

©2019 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)  
209 N. Eighth Street  
Stroudsburg, PA 18360  
USA  
Tel: +1-570-476-8006  
Fax: +1-570-476-0860  
[acl@aclweb.org](mailto:acl@aclweb.org)

ISBN 978-1-950737-64-2

## Preface

The 18th edition of the International Workshop on Treebanks and Linguistics (TLT) follows an annual series that started in 2002, in Sozopol, Bulgaria. TLT addresses all aspects of treebank design, development, and use, "treebank" taken in a broad sense of any spoken or written data augmented with computationally processable annotations of linguistic structure at various levels.

This year's edition is special as TLT is part of the first SyntaxFest, a grouping of four events, which took place in Paris, France, during the last week of August:

- the Fifth International Conference on Dependency Linguistics (Depling 2019)
- the First Workshop on Quantitative Syntax (Quasy)
- the 18th International Workshop on Treebanks and Linguistic Theories (TLT 2019)
- the Third Workshop on Universal Dependencies (UDW 2019)

The use of corpora for NLP and linguistics has only increased in recent years. In NLP, machine learning systems are by nature data-intensive, and in linguistics there is a renewed interest in the empirical validation of linguistic theory, particularly through corpus evidence. While the first statistical parsers have long been trained on the Penn treebank phrase structures, dependency treebanks, whether natively annotated with dependencies, or converted from phrase structures, have become more and more popular, as evidenced by the success of the Universal Dependency project, currently uniting 120 treebanks in 80 languages, annotated in the same dependency-based scheme. The availability of these resources has boosted empirical quantitative studies in syntax. It has also led to a growing interest in theoretical questions around syntactic dependency, its history, its foundations, and the analyses of various constructions in dependency-based frameworks. Furthermore, the availability of large, multilingual annotated data sets, such as those provided by the Universal Dependencies project, has made cross-linguistic analysis possible to an extent that could only be dreamt of only a few years ago.

In this context it was natural to bring together TLT (Treebanks and Linguistic Theories), the historical conference on treebanks as linguistic resources, Depling (The international conference on Dependency Linguistics), the conference uniting research on models and theories around dependency representations, and UDW (Universal Dependency Workshop), the annual meeting of the UD project itself. Moreover, in order to create a point of contact with the large community working in quantitative linguistics it seemed expedient to create a workshop dedicated to quantitative syntactic measures on treebanks and raw corpora, which gave rise to Quasy, the first workshop on Quantitative Syntax. And this led us to the first SyntaxFest.

Because the potential audience and submissions to the four events were likely to have substantial overlap, we decided to have a single reviewing process for the whole SyntaxFest. Authors could choose to submit their paper to one or several of the four events, and in case of acceptance, the program co-chairs would decide which event to assign the accepted paper to.

This choice was found to be an appropriate one, as most submissions were submitted to several of the events. Indeed, there were 40 long paper submissions, with 14 papers submitted to Quasy, 31 to DepLing, 13 to TLT and 16 to UDW. Among them, 28 were accepted (6 at Quasy, 10 at DepLing, 6 at TLT, 6 at UDW). Note that due to multiple submissions, the acceptance rate is defined at the level of the whole

SyntaxFest (around 70%). As far as short papers are concerned, 62 were submitted (24 to Quasy, 41 to DepLing, 35 to TLT and 37 to UDW), and 41 were accepted (8 were presented at Quasy, 14 at DepLing, 9 at TLT and 9 at UDW), leading to an acceptance rate for short papers of around 66%.

We are happy to announce that the first SyntaxFest has been a success, with over 110 registered participants, most of whom attended for the whole week.

SyntaxFest is the result of efforts from many people. Our sincere thanks go to the reviewers who thoroughly reviewed all the submissions to the conference and provided detailed comments and suggestions, thus ensuring the quality of the published papers.

We would also like to warmly extend our thanks to the five invited speakers,

- Ramon Ferrer i Cancho - Universitat Politècnica de Catalunya (UPC)
- Emmanuel Dupoux - ENS/CNRS/EHESS/INRIA/PSL Research University, Paris
- Barbara Plank - IT University of Copenhagen
- Paola Merlo - University of Geneva
- Adam Przepiórkowski - University of Warsaw / Polish Academy of Sciences / University of Oxford

We are grateful to the Université Sorbonne Nouvelle for generously making available the Amphithéâtre du Monde Anglophone, a very pleasant venue in the heart of Paris. We would like to thank the ACL SIGPARSE group for its endorsement and all the institutions who gave financial support for SyntaxFest:

- the "Laboratoire de Linguistique formelle" (Université Paris Diderot & CNRS)
- the "Laboratoire de Phonétique et Phonologie" (Université Sorbonne Nouvelle & CNRS)
- the Modyco laboratory (Université Paris Nanterre)
- the "École Doctorale Connaissance, Langage, Modélisation" (CLM) - ED 139
- the "Université Sorbonne Nouvelle"
- the "Université Paris Nanterre"
- the Empirical Foundations of Linguistics Labex (EFL)
- the ATALA association
- Google
- Inria and its Almanach team project.

Finally, we would like to express special thanks to the students who have been part of the local organizing committee. We warmly acknowledge the enthusiasm and community spirit of:

Danrun Cao, Université Paris Nanterre

Marine Courtin, Sorbonne Nouvelle

Chuanming Dong, Université Paris Nanterre

Yoann Dupont, Inria

Mohammed Galal, Sohag University  
Gaël Guibon, Inria  
Yixuan Li, Sorbonne Nouvelle  
Lara Perinetti, Inria et Fortia Financial Solutions  
Mathilde Regnault, Lattice and Inria  
Pierre Rochet, Université Paris Nanterre  
Chunxiao Yan, Université Paris Nanterre

Marie Candito, Kim Gerdes, Sylvain Kahane, Djamé Seddah (local organizers and co-chairs),  
and Xinying Chen, Ramon Ferrer-i-Cancho, Alexandre Rademaker, Francis Tyers (co-chairs)

September 2019

## Program co-chairs

The chairs for each event (and co-chairs for the single SyntaxFest reviewing process) are:

- Quasy:
  - Xinying Chen (Xi’an Jiaotong University / University of Ostrava)
  - Ramon Ferrer i Cancho (Universitat Politècnica de Catalunya)
- DepLing:
  - Kim Gerdes (LPP, Sorbonne Nouvelle & CNRS / Almanach, INRIA)
  - Sylvain Kahane (Modyco, Paris Nanterre & CNRS)
- TLT:
  - Marie Candito (LLF, Paris Diderot & CNRS)
  - Djamé Seddah (Paris Sorbonne / Almanach, INRIA)
  - with the help of Stephan Oepen (University of Oslo, previous co-chair of TLT) and Kilian Evang (University of Düsseldorf, next co-chair of TLT)
- UDW:
  - Alexandre Rademaker (IBM Research, Brazil)
  - Francis Tyers (Indiana University and Higher School of Economics)
  - with the help of Teresa Lynn (ADAPT Centre, Dublin City University) and Arne Köhn (Saarland University)

## Local organizing committee of the SyntaxFest

Marie Candito, Université Paris-Diderot (co-chair)  
Kim Gerdes, Sorbonne Nouvelle (co-chair)  
Sylvain Kahane, Université Paris Nanterre (co-chair)  
Djamé Seddah, University Paris-Sorbonne (co-chair)  
Danrun Cao, Université Paris Nanterre  
Marine Courtin, Sorbonne Nouvelle  
Chuanming Dong, Université Paris Nanterre  
Yoann Dupont, Inria  
Mohammed Galal, Sohag University  
Gaël Guibon, Inria  
Yixuan Li, Sorbonne Nouvelle  
Lara Perinetti, Inria et Fortia Financial Solutions  
Mathilde Regnault, Lattice and Inria  
Pierre Rochet, Université Paris Nanterre  
Chunxiao Yan, Université Paris Nanterre

## **Program committee for the whole SyntaxFest**

Patricia Amaral (Indiana University Bloomington)  
Miguel Ballesteros (IBM)  
David Beck (University of Alberta)  
Emily M. Bender (University of Washington)  
Ann Bies (Linguistic Data Consortium, University of Pennsylvania)  
Igor Boguslavsky (Universidad Politécnica de Madrid)  
Bernd Bohnet (Google)  
Cristina Bosco (University of Turin)  
Gosse Bouma (Rijksuniversiteit Groningen)  
Miriam Butt (University of Konstanz)  
Radek Čech (University of Ostrava)  
Giuseppe Giovanni Antonio Celano (University of Pavia)  
Çağrı Çöltekin (University of Tuebingen)  
Benoit Crabbé (Paris Diderot University)  
Éric De La Clergerie (INRIA)  
Miryam de Lhoneux (Uppsala University)  
Marie-Catherine de Marneffe (The Ohio State University)  
Valeria de Paiva (Samsung Research America and University of Birmingham)  
Felice Dell'Orletta (Istituto di Linguistica Computazionale "Antonio Zampolli" - ILC CNR)  
Kaja Dobrovoljc (Jožef Stefan Institute)  
Leonel Figueiredo de Alencar (Universidade federal do Ceará)  
Jennifer Foster (Dublin City University, Dublin 9, Ireland)  
Richard Futrell (University of California, Irvine)  
Filip Ginter (University of Turku)  
Koldo Gojenola (University of the Basque Country UPV/EHU)  
Kristina Gulordava (Universitat Pompeu Fabra)  
Carlos Gómez-Rodríguez (Universidade da Coruña)  
Memduh Gökirmak (Charles University, Prague)  
Jan Hajič (Charles University, Prague)  
Eva Hajičová (Charles University, Prague)  
Barbora Hladká (Charles University, Prague)  
Richard Hudson (University College London)  
Leonid Iomdin (Institute for Information Transmission Problems, Russian Academy of Sciences)  
Jingyang Jiang (Zhejiang University)  
Sandra Kübler (Indiana University Bloomington)  
François Lareau (OLST, Université de Montréal)  
John Lee (City University of Hong Kong)  
Nicholas Lester (University of Zurich)  
Lori Levin (Carnegie Mellon University)  
Haitao Liu (Zhejiang University)  
Ján Mačutek (Comenius University, Bratislava, Slovakia)  
Nicolas Mazziotta (Université)  
Ryan McDonald (Google)  
Alexander Mehler (Goethe-University Frankfurt am Main, Text Technology Group)

Wolfgang Menzel (Department of Informatik, Hamburg University)  
Paola Merlo (University of Geneva)  
Jasmina Milićević (Dalhousie University)  
Simon Mille (Universitat Pompeu Fabra)  
Simonetta Montemagni (ILC-CNR)  
Jiří Mírovský (Charles University, Prague)  
Alexis Nasr (Aix-Marseille Université)  
Anat Ninio (The Hebrew University of Jerusalem)  
Joakim Nivre (Uppsala University)  
Pierre Nugues (Lund University, Department of Computer Science Lund, Sweden)  
Kemal Oflazer (Carnegie Mellon University-Qatar)  
Timothy Osborne (independent)  
Petya Osenova (Sofia University and IICT-BAS)  
Jarmila Panevová (Charles University, Prague)  
Agnieszka Patejuk (Polish Academy of Sciences / University of Oxford)  
Alain Polguère (Université de Lorraine)  
Prokopis Prokopidis (Institute for Language and Speech Processing/Athena RC)  
Ines Rehbein (Leibniz Science Campus)  
Rudolf Rosa (Charles University, Prague)  
Haruko Sanada (Rissho University)  
Sebastian Schuster (Stanford University)  
Maria Simi (Università di Pisa)  
Reut Tsarfaty (Open University of Israel)  
Zdenka Uresova (Charles University, Prague)  
Giulia Venturi (ILC-CNR)  
Veronika Vincze (Hungarian Academy of Sciences, Research Group on Artificial Intelligence)  
Relja Vulcanovic (Kent State University at Stark)  
Leo Wanner (ICREA and University Pompeu Fabra)  
Michael White (The Ohio State University)  
Chunshan Xu (Anhui Jianzhu University)  
Zhao Yiyi (Communication University of China)  
Amir Zeldes (Georgetown University)  
Daniel Zeman (Univerzita Karlova)  
Hongxin Zhang (Zhejiang University)  
Heike Zinsmeister (University of Hamburg)  
Robert Östling (Department of Linguistics, Stockholm University)  
Lilja Øvrelid (University of Oslo)

## **Additional reviewers**

James Barry  
Ivan Vladimir Meza Ruiz  
Rebecca Morris  
Olga Sozinova  
He Zhou

## Table of Contents

SyntaxFest 2019 Invited talk - Quantitative Computational Syntax: dependencies, intervention effects and word embeddings . . . . .	1
<i>Paola Merlo</i>	
Are formal restrictions on crossing dependencies epiphenominal? . . . . .	2
<i>Himanshu Yadav, Samar Husain and Richard Futrell</i>	
A Surface-Syntactic UD Treebank for Naija . . . . .	13
<i>Bernard Caron, Marine Courtin, Kim Gerdes and Sylvain Kahane</i>	
Can Greenbergian universals be induced from language networks? . . . . .	25
<i>Kartik Sharma, Kaivalya Swami, Aditya Shete and Samar Husain</i>	
Parallel Dependency Treebank Annotated with Interlinked Verbal Synonym Classes and Roles . . . . .	38
<i>Zdeňka Urešová, Eva Fučíková, Eva Hajičová and Jan Hajič</i>	
Ordering of Adverbials of Time and Place in Grammars and in an Annotated English-Czech Parallel Corpus . . . . .	51
<i>Eva Hajičová, Jiří Mírovský and Kateřina Rysová</i>	
Weighted posets: Learning surface order from dependency trees . . . . .	61
<i>William Dyer</i>	
Linked Open Treebanks. Interlinking Syntactically Annotated Corpora in the XXX Knowledge Base of Linguistic Resources for Latin . . . . .	74
<i>Francesco Mambrini and Marco Passarotti</i>	
Challenges of Annotating a Code-Switching Treebank . . . . .	82
<i>Özlem Çetinoğlu and ÇağrıÇöltekin</i>	
Dependency Parser for Bengali-English Code-Mixed Data enhanced with a Synthetic Treebank . . . . .	91
<i>Urmi Ghosh, Dipti Sharma and Simran Khanuja</i>	
tweeDe –A Universal Dependencies treebank for German tweets . . . . .	100
<i>Ines Rehbein, Josef Ruppenhofer and Bich-Ngoc Do</i>	
Creating, Enriching and Valorizing Treebanks of Ancient Greek . . . . .	109
<i>Alek Keersmaekers, Wouter Mercelis, Colin Swaelens and Toon Van Hal</i>	
Syntax is clearer on the other side - Using parallel corpus to extract monolingual data . . . . .	118
<i>Andrea Dömötör</i>	
Improving Surface-syntactic Universal Dependencies (SUD): MWEs and deep syntactic features . . . . .	126
<i>Kim Gerdes, Bruno Guillaume, Sylvain Kahane and Guy Perrier</i>	
Artificially Evolved Chunks for Morphosyntactic Analysis . . . . .	133
<i>Mark Anderson, David Vilares and Carlos Gómez-Rodríguez</i>	
Challenges of language change and variation: towards an extended treebank of Medieval French . . . . .	144
<i>Mathilde Regnault, Sophie Prévost and Eric Villemonte de la Clergerie</i>	