


# 1 Linguistic Annotations on the Web

2 **Christian Chiarcos** 

3 Applied Computational Linguistics Lab, Goethe University Frankfurt, Germany

4 <http://www.acoli.informatik.uni-frankfurt.de/>

5 [chiarcos@informatik.uni-frankfurt.de](mailto:chiarcos@informatik.uni-frankfurt.de)

6 **Milan Dojchinovski**

7 InfAI/DBpedia Association, Germany

8 CTU in Prague, Czech Republic

9 <http://www.dojchinovski.mk>

10 **Fahad Khan**

11 Istituto di Linguistica Computazionale 'A. Zampolli', CNR, Pisa, Italy

12 <http://www.ilc.cnr.it/en/content/anas-fahad-khan>

13 **Bridget Almas**

14 The Alpheios Project, Ltd.

15 Niskayuna, NY USA

16 <https://www.linkedin.com/in/bridget-almas-43a7762/>

17 **Giedrė Valūnaitė Oleškevičienė**

18 Institute of Humanities, Mykolas Romeris University, Vilnius, Lithuania

19 [https://www.researchgate.net/profile/Giedre\\_Valunaite\\_Oleskeviciene/research](https://www.researchgate.net/profile/Giedre_Valunaite_Oleskeviciene/research)

## 20 — Abstract —

21 The workshop is focused on challenging issues of web annotation and to share and enhance the  
22 state-of-the-art of linguistic annotation on the web involving cross-discipline and cross-linguistic  
23 audiences. We aim to provide a general introduction into the topic, to consolidate this discussion,  
24 and to discuss directions, goals, and concrete strategies.

25 **2012 ACM Subject Classification** Applied computing → Annotation; Applied computing → Format  
26 and notation; Applied computing → Document management and text processing; Software and its  
27 engineering → Interoperability

28 **Keywords and phrases** interoperability, linguistic annotation, linked data, web standards, knowledge  
29 graphs, natural language processing, digital methods in linguistics, Digital Humanities

30 **Digital Object Identifier** 10.4230/OASICS...

31 **Category** workshop proposal, community meeting

## 32 **1 Detailed Description**

33 The workshop is focused on challenging issues of web annotation and to share and enhance  
34 the state-of-the-art of linguistic annotation on the web involving cross-discipline and cross-  
35 linguistic audiences. It provides background information on major community standards,  
36 their benefits and shortcomings, with the specific aim to contribute to and to consolidate  
37 an on-going discussion within the W3C Community Group Linked Data for Language  
38 Technology (LD4LT) on developing a consolidated LOD vocabulary for linguistic annotations  
39 for applications across language technology, empirical linguistics, computational lexicography,  
40 digital humanities, etc.

41 The numerous existing vocabularies that exist for the purpose are neither interoperable  
42 with each other, nor do they cover all relevant use cases. Since 2019, LD4LT is thus working  
43 towards the harmonization and extension of existing standards for creating, publishing,  
44 sharing, accessing and processing linguistic annotations on the web. The goals are to (a)



© Author: Please provide a copyright holder;  
licensed under Creative Commons License CC-BY  
OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

45 provide a survey about standards, challenges and requirements, to (b) work towards a W3C  
46 community report that provides either best practices for annotations or extensions to existing  
47 standards, and, ultimately, to (c) inform subsequent standardization efforts.

48 We aim to provide a general introduction into the topic, to consolidate this discussion,  
49 and to discuss directions, goals, and concrete strategies.

## 50 **2 List of topics**

- 51 ■ interoperability
- 52 ■ linguistic annotation
- 53 ■ linked data
- 54 ■ web standards
- 55 ■ knowledge graphs
- 56 ■ natural language processing
- 57 ■ digital methods in linguistics
- 58 ■ Digital Humanities

## 59 **3 Past editions**

60 As an LD4LT community meeting, this workshop is the first of its kind. However, the  
61 organizers have been involved in organizing numerous workshops, summer schools and  
62 conference on the topic, including:

- 63 ■ Seven international workshops on Linked Data in Linguistics (LDL-2013, 2014, 2015,  
64 2016, 2018, 2020): 40-90 participants each
- 65 ■ Two Summer Datathons on Linguistic Linked Open Data (SD-LLOD 2017, SD-LLOD  
66 2019): 40-50 participants
- 67 ■ Language Resources and Linked Data tutorial (EKAW-2014)
- 68 ■ Two conferences on Language, Data and Knowledge (LDK-2017, LDK-2019): 100-120  
69 participants

70 This meeting builds on this experience, but is dedicated to a more narrowly defined aspect.  
71 The LD4LT community group currently has more than 100 members, but for a presence  
72 meeting, we adopt a conservative estimate about the expected number of participants (see  
73 below).

## 74 **4 Format**

75 The workshop is planned to be a general assembly of the W3C community group Linked  
76 Data for Language Technology, and organized in conjunction with the COST Action Nexus  
77 Linguarum. In that sense, the format is less a classical workshop with formal paper submission  
78 but rather an informal discussion round with invited presentations, but focusing on discussion.  
79 It will be a mixture of presentations, use cases and discussion. We will begin with background  
80 presentations and discussion of specific standards for linguistic annotation on the web (60  
81 minutes). This will be followed by a period of general open discussion on topics including  
82 additional vocabularies, formats and their interrelationships; goals and requirements for  
83 interoperability; and challenges to community adoption of annotation standards (30 minutes).  
84 Subsequent to this discussion, there will be a short presentation and summary of the activities  
85 and progress of the LD4LT W3C community group (20 minutes).

86 The second half of the workshop will begin with presentation of specific use cases on topics  
87 such as tools for using and producing Linked Open Linguistic Annotation Data; annotating  
88 pragmatics; language resource transformation; the “Explicit Citations”: how to provide  
89 annotation of PDF documents, e.g., with explicit citation information, etc. It should be  
90 stressed that the workshop is open to any challenging topics related to web annotation and  
91 it is expecting heated, inspiring and productive discussions envisioning the future research.  
92 At the moment, four use case presentations have been confirmed, we expect this number to  
93 increase. Use case providers will give live lightning talks and also pre-prepare posters or  
94 other digital form of presentation such as a power-point or pre-recorded demo (80 minutes).  
95 (Note that the exact format and mechanism for presentation of the use cases will depend  
96 upon whether the workshop is to be on-site or partially or fully virtual, and if the latter, the  
97 virtual technology to be used). The workshop will conclude with a discussion of strategies  
98 and possible milestones for consolidating linguistic annotations (40 minutes).

## 99 **5 Expected audience**

100 We expect the mixed audience coming from the LD4LT, from Nexus Linguarum and also from  
101 outside these networks. Anyone willing to present their state-of-the-art research or participate  
102 in the discussions is kindly welcome to join as the workshop expects cross-fertilization both  
103 across the project Nexus Linguarum domains and the research coming outside the project  
104 with the view joining the project if there is an interest. Depending on Covid-19 situation we  
105 expect 15-20 on-site participants and about 50 online participants if the workshop has to be  
106 organized in a hybrid mode.

## 107 **6 Duration**

108 The workshop is envisioned to take half a day (4 hours).

## 109 **7 Technical requirements**

110 It is preferable to have video-conferencing service, audio-visual equipment and poster boards.

## 111 **8 Committee members**

112 As there will not be a call for papers, but a general call for participation, we did not appoint  
113 a program committee. The organization committee includes five people that reflect the  
114 expected composition of the audience:

- 115 ■ LD4LT
  - 116 ■ Christian Chiarcos, Applied Computational Linguistics, Goethe Universität Frankfurt,  
117 Germany
- 118 ■ Nexus Linguarum
  - 119 ■ Milan Dojchinovski, InfAI/DBpedia Association, Germany / CTU in Prague, Czech  
120 Republic
  - 121 ■ Fahad Khan, Istituto di Linguistica Computazionale ‘A. Zampolli’, CNR, Pisa, Italy
- 122 ■ The user community
  - 123 ■ Bridget Almas, The Alpheios Project, Ltd., Niskayuna, NY USA
  - 124 ■ Giedre Valunaite Oleskeviciene, Institute of Humanities, Mykolas Romeris University,  
125 Vilnius, Lithuania

## 126 **9** References

127 Almas, B., A. Babeu, A. Krohn, Linked Data in the Perseus Digital Library. In ISAW Papers  
128 7: Current Practice in Linked Open Data for the Ancient World, New York : Institute for  
129 the Study of the Ancient World, New York University, 2014.

130 Chiarcos, C., Nordhoff, Sebastian, Hellmann, Sebastian (Eds., 2013), *Linked Data in*  
131 *Linguistics – Representing and Connecting Language Data and Language Metadata*, Springer,  
132 Heidelberg

133 Cimiano, P., Chiarcos, C., McCrae, J.P., Gracia, J. (2020), *Linguistic Linked Data –*  
134 *Representation, Generation and Applications*, Springer, Cham

135 Dobrovoljc, K. (2017). Multi-word discourse markers and their corpus-driven identification:  
136 The case of MWDM extraction from the reference corpus of spoken Slovene. *International*  
137 *Journal of Corpus Linguistics*, 22(4), 551–582.

138 Dupont, M., Zufferey, S. (2017). Methodological issues in the use of directional parallel  
139 corpora: A case study of English and French concessive connectives. *International Journal of*  
140 *Corpus Linguistics*, 22(2), 270–297.

141 Hellmann, S., J. Lehmann, S. Auer, M. Brümmer (2013), *Integrating NLP using Linked*  
142 *Data*, in Proc. 12th International Semantic Web Conference, 21-25 October 2013 (Sydney,  
143 Australia).

144 Ide, N., Chiarcos, C., Stede, M., Cassidy, S. (2017). Designing annotation schemes: from  
145 model to representation. In *Handbook of Linguistic Annotation* (pp. 73-111). Springer,  
146 Dordrecht.

147 Oleskeviciene, G. V., Zeyrek, D., Mazeikiene, V., Kurfali, M. (2018). Observations on the  
148 annotation of discourse relational devices in TED talk transcripts in Lithuanian. *Proceedings*  
149 *of the Workshop on Annotation in Digital Humanities Co-Located with ESSLLI*, 2155, 53–58.

150 Pareja-Lora, A., María Blume, Barbara C. Lust and Christian Chiarcos (Eds., 2020),  
151 *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive*  
152 *Research in the Language Sciences*, MIT Press, Cambridge, MA

153 Snyder, B., Barzilay, R., Knight, K. (2010). A statistical model for lost language  
154 decipherment.

155 TEI Consortium (2020). 15.4 Linguistic Annotation of Corpora. In: *TEI P5: Guidelines*  
156 *for Electronic Text Encoding and Interchange Version 4.1.0*. Last updated on 19th August  
157 2020. TEI Consortium. <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/CC.html#CCAN> (Date of Access: 04/12/20)

159 Wei, N., Li, J. (2013). A new computing method for extracting contiguous phraseological  
160 sequences from academic text corpora. *International Journal of Corpus Linguistics*, 18(4),  
161 506–535.

162 Zufferey, S., Degand, L. (2017). Annotating the meaning of discourse connectives in  
163 multilingual corpora. *Corpus Linguistics and Linguistic Theory*, 13(2), 399–422.

## 164 **10** Organizers and presenters

165 Dr. Christian Chiarcos  
166 Goethe University Frankfurt, Germany  
167 [https://www.researchgate.net/profile/Christian\\_Chiarcos](https://www.researchgate.net/profile/Christian_Chiarcos)

169 Christian Chiarcos is Assistant Professor of Computer Science at the Goethe University  
170 Frankfurt, where he is heading the Applied Computational Linguistics lab. In 2010, he

171 received a doctoral degree on the topic of Natural Language Generation from the University  
172 Potsdam, Germany, he worked subsequently at the Information Sciences Institute of the  
173 University of Southern California (ISI/USC), before joining Goethe University Frankfurt in  
174 2013. His research focuses on semantic technologies, including both computational semantics  
175 as well as the innovative application of Semantic Web standards to NLP problems. As a  
176 computational linguist, Christian Chiarcos explored Semantic Web and Linked Data from  
177 an NLP and DH perspective and contributed to the emergence of a community at the  
178 intersection of NLP and Semantic Web: He has been co-founder of the Open Linguistics  
179 Working Group of the Open Knowledge Foundation (OWLG, since 2010), he initiated and  
180 co-organized the Linked Data in Linguistics workshop series (since 2012), the Language,  
181 Data and Knowledge conference series (since 2017), and the accompanying development of  
182 a Linguistic Linked Open Data (LLOD) cloud. He is a chair of the W3C Linked Data for  
183 Language Technology community group and leads a task on Data Modelling for Linguistic  
184 Linked Open Data in the COST action Nexus Linguarum.

185

186 Dr. Milan Dojchinovski  
187 InfAI/DBpedia Association, Germany  
188 CTU in Prague, Czech Republic  
189 <http://www.dojchinovski.mk>

190

191 Milan holds a Research Associate position at the Institute for Applied Informatics (InfAI)  
192 and an Assistant Professor position at the Czech Technical University in Prague. He has 10+  
193 years experience in the computer industry in Germany, Czech Republic and Slovenia. His  
194 research interests are in Semantic Web, NLP and Knowledge Graph technologies. Milan leads  
195 a working group on Linguistic Linked Open Data in the COST NexusLinguarum project.  
196 He has strong interests in the cross-section of linguistics and semantic web technologies. In  
197 recent years, he has also contributed to the development and dissemination of the NIF format.  
198 He was working on several European and national projects funded by the FP7 and H2020  
199 programmes. Since 2013 Milan has been an active member of the DBpedia community pro-  
200 ject. Milan holds a PhD in Information Science from the Czech Technical University in Prague.

201

202 Dr. Fahad Khan  
203 Istituto di Linguistica Computazionale 'A. Zampolli', CNR, Pisa, Italy  
204 <http://www.ilc.cnr.it/en/content/anas-fahad-khan>

205

206 Fahad Khan is a full time researcher at the Istituto di Linguistica Computazionale in Pisa.  
207 He has a PhD from the University of Nottingham in Computer Science. His research interests  
208 include the modelling, creation and publication of lexical resources, the use of ontologies and  
209 conceptual modelling techniques in linguistics and the digital humanities, and the use of  
210 standards in the social sciences and humanities. He is a member of the British Standard  
211 Institute as well as the ISO/TC 37/SC 4 Working Group, and a co-leader of the ISO Standard  
212 ISO/FDIS 24613-3. Fahad co-leads a task on Data Modelling for Linguistic Linked Open  
213 Data in the COST action Nexus Linguarum.

214

215 Bridget Almas  
216 The Alpheios Project, Ltd.

217 Niskayuna, NY USA

218 <https://www.linkedin.com/in/bridget-almas-43a7762/>

219

220 Bridget Almas has over 25 years of experience working in software development, in com-  
221 mercial, academic and non-profit environments. She is currently Executive Director and  
222 Software Architect for the non-profit Alpheios Project, building evidence-based, open-source  
223 software to support worldwide study of classical languages and literatures. This software  
224 makes extensive reuse of open source linguistic data sets and software. In her prior role  
225 at Tufts University, Bridget was the technical lead on the Perseids Project and before that  
226 the Perseus Digital Library. She has previously served on the Technical Advisory Board of  
227 the Research Data Alliance (RDA), and as co-chair of the RDA Research Data Collections  
228 Working Group and Data Fabric Interest Group. Bridget is a co-founder of the Distributed  
229 Text Services Specification effort and is currently a member of its Technical Committee.

230

231 Dr. Giedrė Valūnaitė Oleškevičienė

232 Institute of Humanities, Mykolas Romeris University, Vilnius, Lithuania

233 [https://www.researchgate.net/profile/Giedre\\_Valunaite\\_Oleskeviciene/research](https://www.researchgate.net/profile/Giedre_Valunaite_Oleskeviciene/research)

234

235 Giedrė is an associate professor at the Institute of Humanities, Mykolas Romeris University.  
236 Her scientific interests in the domain of humanities include discourse analysis, discourse  
237 annotated corpora, professional English and legal English, and in the domain of social  
238 sciences, educational science her scientific interests include social research methodology,  
239 modern education, philosophical issues, creativity development in modern education systems,  
240 etc. She is also engaged in second language teaching and learning research, linguistics and  
241 translation research. She is actively involved in participating in international project research  
242 activities. She was a Managing Committee (MC) member representing Lithuania in an  
243 international research project funded by the COST action no. IS1312 “Structuring Discourse  
244 in Multilingual Europe (TextLink)” 2013-2018. She also completed her postdoctoral research  
245 project “Discourse Annotated Corpus Based on Social Media Texts for Second Language  
246 Teaching and Learning” financed by Lithuanian Research Council.

247