Sarah Schulz

Education

01/2018	**Postdoctoral Researcher**, *University of Stuttgart*, Stuttgart, Institute for Natural Language Processing (IMS).
05/2015 –12/2017	**PhD student**, *University of Stuttgart*, Stuttgart, Institute for Natural Language Processing (IMS).
2013–2015	**PhD student**, *Ghent Univerity*, Ghent, Applied Linguistics – Langugage and Translation Technology Team (LT3).
2010–2013	**Master of Arts**, *Eberhard-Karls-Universität*, Tübingen, *1,3.* International Studies in Computational Linguistics
2007–2010	**Bachelor of Arts**, *Friedrich-Alexander-Universität*, Erlangen-Nürnberg, *1,7.* Studies in Theatre and Media Science and German
2004–2007	**Abitur**, *Fritz-Ruoff-Schule*, Nürtingen, *1,4.*
1998–2004	**Mittlere Reife**, *Realschule*, Neuffen, *1,8.*

PhD thesis

title **The Taming of the Shrew** - *Non-Standard Text Processing in the Digital Humanities*

supervisors Prof. Dr. Jonas Kuhn

description Digital Humanities as a research field that offers a variety of texts that are different from the standard on a lexical and syntactic level, challenges NLP to develop more flexible ways for automatic processing.

Master thesis

title **On Tokenization** - *Automatic extraction of semantic multiword units for meaning-based tasks*

supervisors Prof. Dr. Walt Detmar Meurers & PD. Dr. Frank Richter

description An approach is presented to automatically find meaning units in a text to enable semantic tokenization. It uses statistical methods borrowed from multiword expression extraction and combines it with a new approach based on unity checking of meaning of phrases.

Bachelor thesis

title **Zwischen Latein und Volkssprache** - *Der Deutsche Ptolemäus auf dem Weg zu einer vielschichtigen digitalen Edition durch die Implementierung von XLink*

supervisors Prof. Dr. Mechthild Habermann & Prof. Dr. Günther Görz description In this thesis, the digital edition of the Early New High German text “Der deutsche Ptolemäus” was extended by extra-information about the connection between the German and Latin vocabulary. This information was included in a glossary that was implemented with the help of Xlink.

Experience

05/2015–12/2017 **PhD student**, *University of Stuttgart - Instiute for Natural Language Processing* , Stuttgart.
PhD student - project: Center for Reflected Text Analytics (CRETA)
Task:
- processing of mainly historical text in a Digital Humanities context
- interdisciplinary work
- machine learning

12/2014–01/2015 **intern**, *The Swiss AI Lab - IDSIA (Istituto Dalle Molle di Studi sull'Intelligenza Artificiale)*, Lugano.
Reseach Internship Neural Networks and Language Processing

04/2013–04/2015 **PhD student**, *Ghent University - Language and Translation Technology Team* , Ghent.
PhD student - project: Automatic Monitoring for Cyberspace Applications
Task:
- automatic text normalization of user-generated content
- machine translation
- machine learning

03/2012–08/2012 **intern**, *European Academy of Bolzano - Institute for Specialised Communication and Multilingualism*, Bolzano.
Intern at the project 'Korpus Südtirol'
Tasks:
- acquisition and digitalization of data for Korpus Südtirol
- actualization and modification of Wortwarte developed by Lothar Lemnitzer and adaption for South Tyrolean German
- compilation of a web-based corpus for South Tyrolean German

2010 –2012 **student assistent**, *Department of Linguistics*, Tübingen.
Proceeding with the project of the German wordnet ‘GermaNet’

2009–2010 **student assistent**, *Department of Artificial Intelligence*, Erlangen.
Assistent at a course about XML and digital editions

09/2009–10/2009 **intern**, *Department of Artificial Intelligence*, Erlangen.
Assistance at different projects
Tasks:
- digital data acquisition and evaluation
- preparation of a first test run of a navigation and recommendation system for mobile phones

2008–2009 **student assistant**, *Department of German Linguistics*, Erlangen.
Investigation of language change by reference to editions of ‘Melusine’

Main research interests

- Machine translation
- Machine learning
- Digital Humanities

- Non-standad text

Languages

German	fluent	*native language*
English	fluent	*9 years at school, master's program in English*
Dutch	fluent	*private lessons, 2 years in Belgium*
French	basics	*3 years at school*
Latin	basics	*2 semesters at university*
Italian	basics	*1 semester at university, 6 months in South Tyrol*
Portuguese	basics	*1 semester at community college*

Computer skills

Programming	Python	*excellent*	Mark-up	LaTex	*excellent*
	Java	*basic*		XML	*good*
	Perl	*basics*		HTML	*basics*
	Bashscript	*basics*			
	Gnu R	*basics*			
Office	Linux				
	OpenOffice				

Memberships and Reviewing

- Digital Humanities im deutschsprachigen Raum (DHd) 2018 in Köln
- Programme Committee 12th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) 2018
- Scientific Committee LREC 2018
- Workshop CEx@AI*IA 2017 in Bari
- Digital Humanities (DH) 2017 in Montreal
- Programme Committee 11th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) 2017
- Digital Humanities im deutschsprachigen Raum (DHd) 2017 in Bern
- Scientific Committee LREC 2016
- Digital Humanities (DH) 2016 in Krakow
- Programme Committee 10th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) 2016
- Journal of Artificial Intelligence Research (JAIR)

Organizing/Managing

- Co-Manager of Interdisciplinary College 2018 (9th to 16th of March 2018, Günne)
- Co-Manager of Interdisciplinary College 2017 (10th to 17th of March 2017, Günne)

Interests

sports	pilates, martial arts, hiking, mountain biking,running, climbing
others	theater, old cars

Publications

[1] Özlem Çetinoğlu, Sarah Schulz, and Ngoc Thang Vu. Challenges of Computational Processing of Code-Switching. In *Proceedings of EMNLP Workshop on Computational Approaches to Linguistic Code Switching (CALCS 2016) @EMNLP*, Austin, Texas, USA, November 2016.

[2] Orphée De Clercq, Sarah Schulz, Bart Desmet, Els Lefever, and Véronique Hoste. Normalization of Dutch User-Generated Content. In *Proceedings of the 9th International Conference on Recent Advances in Natural Language Processing*, Hissar, Bulgaria, 2013.

[3] Orphée De Clercq, Schulz Schulz, Bart Desmet, and Véronique Hoste. Towards Shared Datasets for Normalization Research. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[4] Marisa Delz, Benjamin Layer, Sarah Schulz, and Johannes Wahle. Overgeneralization of verbs — The change of the German verb system. In *Proceedings of the 9th International Conference on the Evolution of Language*, Evolang IX, pages 96–103, Kyoto, Japan, 3 2012.

[5] Bart Desmet, Orphée De Clercq, Marjan Van de Kauter, Sarah Schulz, Cynthia Van Hee, and Veronique Hoste. Taaltechnologie 2.0: sentimentanalyse en normalisatie. In Stefaan Evenepoel, Patrick Goethals, and Lieve Jooken, editors, *Beschouwingen uit een talenhuis : opstellen over onderwijs en onderzoek in de vakgroep Vertalen, Tolken en Communicatie aangeboden aan Rita Godyns*, pages 157–161. Academia Press, 2014.

[6] Derek Doran, Sarah Schulz, and Tarek Besold. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. In *Proceedings of CeX – Comprehensibility and Explanation in AI and ML*, 2017.

[7] Nora Echelmeyer, Nils Reiter, and Sarah Schulz. Ein PoS-Tagger für "das" Mittelhochdeutsche. In *Book of Abstracts of DHd 2017*, Bern, Switzerland, February 2017.

[8] Nils Reiter, Sarah Schulz, Gerhard Kremer, Roman Klinger, Gabriel Viehhauser, and Jonas Kuhn. Teaching Computational Aspects in the Digital Humanities Program at University of Stuttgart – Intentions and Experiences. In *Proceedings of the Workshop on Teaching NLP for Digital Humanities (Teach4DH 2017) co-located with GSCL 2017*, pages 43–48, Berlin, Germany, September 2017.

[9] Sarah Schulz. Named-Entity Recognition for User-Generated Content. `http://www.kr.tuwien.ac.at/drm/dehaan/stus2014/proceedings.pdf`, 2014.

[10] Sarah Schulz and Mareike Keller. Code-Switching Ubique Est - Language Identification and Part-of-Speech Tagging for Historical Mixed Text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 43–51, Berlin, Germany, August 2016. Association for Computational Linguistics.

[11] Sarah Schulz and Nora Ketschik. From 0 to 10 Million Annotated Words – Part-of-Speech Tagging for Middle High German. *Language Resources and Evaluation*, subm.

[12] Sarah Schulz and Jonas Kuhn. Learning from Within? Comparing PoS Tagging Approaches for Historical Text. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *LREC*. European Language Resources Association (ELRA), 2016.

[13] Sarah Schulz and Jonas Kuhn. Multi-modular domain-tailored OCR post-correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. The Association for Computational Linguistics, 2017.

[14] Sarah Schulz, Jonas Kuhn, and Nils Reiter. Authorship Attribution of Mediaeval German Text: Style and Contents in Apollonius von Tyrland. In *Digital Humanities 2016: Conference Abstracts*, pages 883–885, Kraków, Poland, 2016.

[15] Sarah Schulz, Verena Lyding, and Lionel Nicolas. Compiling a diverse web corpus for South Tyrolean German - STirWaC. In *Proceedings of the 8th Web as Corpus Workshop*, pages 37–45, Lancaster, UK, 2013.

[16] Sarah Schulz, Guy Pauw, Orphée De Clercq, Bart Desmet, Véronique Hoste, Walter Daelemans, and Lieve Macken. Multimodular Text Normalization of Dutch User-Generated Content. *ACM Trans. Intell. Syst. Technol.*, 7(4), 2016.