egon w. stemle

Researcher

European Academy of Bozen/Bolzano (EURAC)
Viale Druso, 1
I-39100 Bolzano (BZ)
☎ *+39 0471 055.129*
✉ *egon.stemle@eurac.edu*
iiegn.eu/work

Egon Stemle is a Cognitive Scientist: he studies skills like perception, thinking, learning, motor function, and language by combining the humanistic and analytical methods of the arts and the formal sciences.
His research focus within this relatively new 'inter-discipline' lies in the area where Computational Linguistics and Artificial Intelligence converge. He works on computer aided fabrication of ontologies from large document repositories, the technological feasibility thereof and the utilization of cross-linked structured data in applications, as well as on tools for editing, processing, and annotating linguistic data.
His curiosity in research is driven by the question why humans handle incomplete and – more often than not – inconsistent structured concepts just fine, whereas computational processes are often of little avail or fail completely.

Research and Professional Experience

since 02.2012 **Researcher**, *Institute for Specialised Communication and Multilingualism at the European Academy of Bozen/Bolzano (EURAC).*

04.2009 – – 01.2012 **Research Fellow**, *LiveMemories sponsored grant, financed by the Provincia Autonoma of Trento, for research activities at the Center for Mind/Brain Sciences (CIMeC) of the University of Trento.*

since 2009 **Technology Consultant and Shareholder**, *Whitematter Labs.*
Provider of neuroscientific technologies for marketers.

04.2008 – – 2010 **Project Member**, *GoodGaze*, Institute of Cognitive Science, University of Osnabrück, Germany.
Research project to predict where people will look on Web pages.

2008 **Freelancer**, *ontoprise GmbH.*
Integrating in-house technology with IBM's OmniFind; Feasibility Study: evaluate to which extent very specific information extraction from large web data collections combined with semi-automatic cleaning is technically feasible, what benefits and efforts are to be expected – while also considering external alternatives.

10.2006 – – 03.2008 **Research Participant**, *DAAD funded research project (Project-linked exchange of academics and scientists, PPP)*, Collaboration between the Institute of Cognitive Science, U. Osnabrück, Germany, the CNRS UMR 8163 (Savoirs, Textes, Langages), U. of Lille (III), France, and the Departament de Traducció i Filologia, U. Barcelona (Pompeu Fabra), Spain.
Reference to Abstract Objects in Natural Language (OntoRef).

04.2006 – – 03.2008 **Research Assistant**, Institute of Applied Informatics and Formal Description Methods, University of Karlsruhe, Germany, Knowledge Management Research Group.
Maintenance and development of software and scientific counsel; e.g. evaluating CL tools for OmniFind+UIMA and hooking them together.

08.2004 – – 09.2005 **Research Participant**, *ASADO*, Cooperation between the Universities of Osnabrück and Hildesheim, and the aircraft manufacturer AIRBUS.
Project to research methodologies and technologies to analyze and structure the huge amount of documentation produced during aircraft construction.

08.2003 – – 12.2003 **Student Assistant**, *Institute of Cognitive Science, University of Osnabrück, Germany*, Artificial Intelligence Research Group.
Maintenance and development of software for the MiLCA (Media intensive learning units for courses in computational linguistics) project.

10.2001 – – 03.2009 **System Administrator**, *Institute of Cognitive Science, University of Osnabrück, Germany.*
Maintenance of computer hardware and software, server administration (Web Services, Databases, File Sharing, Printing, Mail, Calendaring, (D)VCS, CMS), user support; assisting work groups in making strategic decisions about hardware and software.

Education

02.2006 – – 03.2009 **MSc in Cognitive Science**, *with distinction*, University of Osnabrück, Germany.
Majors: *Linguistics and Computational Linguistics*, and *Artificial Intelligence*, One Year Study Project: Analysis and Structure of Aviation Documents (ASADO).

Thesis **Hybrid Sweeping: Streamlined Perceptual Structured-Text Refinement**.
Development of a perceptually driven content extraction architecture for Web pages – exemplified by a Web page cleaning system. Supervisors: Prof. Dr. Stefan Evert and Prof. Dr. Peter König.

03.2002 – – 09.2002 **Semester abroad**, *ILLC (Institute for Logic, Language and Computation)*, University of Amsterdam, The Netherlands.

10.2000 – – 01.2006 **BSc in Cognitive Science**, University of Osnabrück, Germany.

12.1998 – – 07.2000 **Apprenticeship as Travel Agent**, WORLDWIDE Touristik, Nürnberg, Germany.

06.1997 **Abitur**, Sigmund-Schuckert-Gymnasium, Nürnberg, Germany.

Teaching Experience

2009 – 2011 (yearly) **Lecture on Web Corpora**, *Course in Text Processing class*, Coordinated by Marco Baroni, University of Trento, Italy, Offered to students of the School of Humanities and Philosophy, of the International Master in Cognitive Science and of the Master in Human Language Technologies and Interfaces.

10.2003 – – 02.2006 (every other semester) **Artificial Intelligence Tutor**, *Methods of Artificial Intelligence class*, Institute of Cognitive Science, University of Osnabrück, Germany, Second year Cognitive Science class.
Weekly tutoring of the lecture material, assessment of programming assignments, homework corrections, and exam preparation.

04.2003 – – 07.2006 (every other semester) **Computational Linguistics Tutor**, *Introduction to Computational Linguistics class*, Institute of Cognitive Science, University of Osnabrück, Germany, First year Cognitive Science class.
Weekly tutoring of the lecture material, preparation and assessment of programming assignments, homework corrections, and exam preparation.

10.2002 – – 02.2003 **Theoretical Neuroscience Tutor**, *Introduction to Theoretical Neuroscience class*, Institute of Cognitive Science, University of Osnabrück, Germany, Second year Cognitive Science class.
Weekly tutoring of the lecture material, homework corrections.

10.2001 – – 02.2002 **Computer Science Tutor**, *Algorithms class*, Computer Science Department, University of Osnabrück, Germany, First year Computer Science class.
Weekly oral assessment of 10 2-people groups and homework corrections.

Publications

2015 **Egon Stemle and Alexander Onysko**, *Automated L1 identification in English learner essays and its implications for language transfer*, In Peukert, editor, *Transfer Effects in Multilingual Language Development*, pages 297–321.
John Benjamins, [link]

12.2014 **Aivars Glaznieks and Egon Stemle**, *Challenges of building a CMC corpus for analyzing writer's style by age: The DiDi project*, In Beißwenger et al., editors, *Journal for Language Technology and Computational Linguistics (JLCL)*, pages 31–57, [link].

12.2014 **Aivars Glaznieks, Andrea Abel, Verena Lyding, Lionel Nicolas, and Egon Stemle**, *Establishing a Standardised Procedure for Building Learner Corpora*, In Nikula et al., editors, *Apples - Journal of Applied Language Studies*, pages 5–20, [link].

12.2014 **Michel Généreux, Egon W. Stemle, Lionel Nicolas, and Verena Lyding**, *Correcting OCR errors for German in Fraktur font*, In Basili et al., editors, *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014)*, Pisa, Italy, [link].

10.2014 **Jennifer-Carmen Frey, Egon W. Stemle, and Aivars Glaznieks**, *Collecting language data of non-public social media profiles*, In Faaß and Ruppenhofer, editors, *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 11–15, Hildesheim, Germany, Universitatsverlag Hildesheim, Germany, [link].

05.2014 **Verena Lyding, Lionel Nicolas, and Egon Stemle**, *'interHist' - an interactive visual interface for corpus exploration*, In Calzolari et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 635–641, Reykjavik, Iceland, European Language Resources Association (ELRA), [link].

05.2014 **Andrea Abel, Aivars Glaznieks, Lionel Nicolas, and Egon Stemle**, *KoKo: An L1 Learner Corpus for German*, In Calzolari et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2414–2421, Reykjavik, Iceland, European Language Resources Association (ELRA), [link].

04.2014 **Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli**, *The PAISÀ Corpus of Italian Web Texts*, In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 36–43, Gothenburg, Sweden, Association for Computational Linguistics, [link].

12.2013 **Egon W. Stemle and Alexander Onysko**, *Language as a Detective Story*, Article in Academia (science magazine by EURAC and unibz), Bolzano, Italy.

11.2013 **Verena Lyding, Claudia Borghetti, Henrik Dittmann, Lionel Nicolas, and Egon Stemle**, *Open Corpus Interface for Italian Language Learning*, In *Proceedings of the International Conference ICT for Language Learning, 6th edition*, Florence, Italy, libreriauniversitaria.it, [link].

09.2013 **Lionel Nicolas, Egon W. Stemle, Klara Kranebitter, and Verena Lyding**, *High-Accuracy Phrase Translation Acquisition Through Battle-Royale Selection*, In Angelova et al., editors, *Proceedings of Recent Advances in Natural Language Processing, RANLP 2013*, pages 516–524, Hissar, Bulgaria, RANLP 2011 Organising Committee / ACL, [link].

07.2013 **Stefan Evert, Egon Stemle, and Paul Rayson, editors**, Proceedings of the 8th Web as Corpus Workshop (WAC-8), Workshop at the seventh international Corpus Linguistics conference (CL2013), Lancaster, UK.
WAC-8 Organising Committee, [link]

06.2013 (in press) **Klara Kranebitter and Egon W. Stemle**, *Constructing concept relation maps to support building concept systems in comparative legal terminology*, Terminologie & Ontologie: Théories et Applications (TOTh'2013), Chamébry, France.

05.2013 (in press) **Lionel Nicolas, Egon Stemle, Aivars Glaznieks, and Andrea Abel**, *A Generic Data Workflow for Building Annotated Text Corpora*, Compiling and Using Learner Corpora to Teach and Assess Productive and Interactive Skills in Foreign Languages at University Level, Learner Corpora, Padova, Italia.

09.2012 **Lionel Nicolas, Egon W. Stemle, and Klara Kranebitter**, *Towards high-accuracy bilingual phrase acquisition from parallel corpora*, In Jancsary, editor, *11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing*, pages 471–479, Vienna, Austria, ÖGAI, [link].

07.2012 **Francesca Bonin, Fabio Cavulli, Aronne Noriller, Massimo Poesio, and Egon W. Stemle**, *Annotating Archaeological Texts: An Example of Domain-Specific Annotation in the Humanities*, In *Proceedings of the Sixth Linguistic Annotation Workshop*, number July in LAW VI '12, pages 134–138, Jeju, Republic of Korea, Association for Computational Linguistics, [link].

11.2011 **Massimo Poesio, Eduard Barbu, Francesca Bonin, Fabio Cavulli, Asif Ekbal, Egon Stemle, and Christian Girardi**, *The Humanities Research Portal: Human Language Technology Meets Humanities Publication Archives*, In Maegaard, editor, *Proceedings of Supporting Digital Humanities (SDH2011): Answering the unaskable*, Copenhagen, Denmark.

11.2011 **Asif Ekbal, Francesca Bonin, Sriparna Saha, Egon Stemle, Eduard Barbu, Fabio Cavulli, Christian Girardi, and Massimo Poesio**, *Rapid Adaptation of NE Resolvers for Humanities Domains using Active Annotation*, In *Journal for Language Technology and Computational Linguistics (JLCL)*, pages 39–51, [link].

07.2011 **Brian Murphy and Egon W. Stemle**, *PaddyWaC: A Minimally-Supervised Web-Corpus of Hiberno-English*, In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 22–29, Edinburgh, Scotland, UK, Association for Computational Linguistics, [link].

06.2011 **Massimo Poesio, Eduard Barbu, Egon W. Stemle, and Christian Girardi**, *Structure-Preserving Pipelines for Digital Libraries*, In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2011)*, pages 54–62, Portland, OR, USA, Association for Computational Linguistics, [link].

05.2010 **Kepa Joseba Rodríguez, Francesca Delogu, Jannick Versley, Egon W. Stemle, and Massimo Poesio**, *Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus*, In Calzolari et al., editors, *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, European Language Resources Association (ELRA), [link].

09.2009 **Johannes Steger and Egon Stemle**, *KrdWrd: Architecture for Unified Processing of Web Content*, In Alegria et al., editors, *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*, pages 63–70, Donostia-San Sebastian, Basque Country, Spain, Elhuyar Fundazioa, [link].

09.2007 **Daniel Bauer, Judith Degen, Xiaoye Deng, Priska Herger, Jan Gasthaus, Eugenie Giesbrecht, Lina Jansen, Christin Kalina, Thorben Krüger, Robert Märtin, Martin Schmidt, Simon Scholler, Johannes Steger, Egon Stemle, and Stefan Evert**, *FIASCO: Filtering the Internet by Automatic Subtree Classification, Osnabrück*, In Fairon et al., editors, *Proceedings of the Third Web as Corpus Workshop (WAC3)*, Louvain-la-Neuve, Presses universitaires de Louvain, [link].

07.2007 **Sebastian Blohm, Philipp Cimiano, and Egon Stemle**, *Harvesting Relations from the Web - Quantifiying the Impact of Filtering Functions*, In *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI-07)*, pages 1316–1323, Association for the Advancement of Artificial Intelligence, [link].

11.2005 **Martin Bleichner, Eugenie Giesbrecht, Helmar Gust, Eva-Maria Leicht, Petra Ludewig, Sabine Möller, Wiebke Müller, Martin Schmidt, Moritz Stefaner, Egon Stemle, and Katja Wilke**, *ASADO: The Analysis and Structuring of Aviation Documents - Final Report*, Technical report, Institute of Cognitive Science at the University of Osnabrück and Institute of Applied Linguistics at the University of Hildesheim.

Presentations (talks, posters, etc. without proceedings)

10.2015 (upcoming) **Egon W. Stemle**, *The DiDi Project: Collecting, Annotating, and Analysing South Tyrolean Data of Computer-mediated Communication.*, Invited Talk at the first international research days (IRDs) on Social Media and CMC Corpora for the eHumanities, Rennes 2 University, Rennes, France, [link].

04.2014 **Egon W. Stemle* and Alexander Onysko***, *Automated L1 identification in English learner essays and its implications for language transfer*, Talk in the 'Work in Progress Series' at the Kompetenzzentrum Sprachen, Freie Universität Bozen, Bozen/Bolzano, Italy.

11.2013 **Andrea Abel, Aivars Glaznieks, and Egon W. Stemle**, *Automatische Annotation von Schülertexten - Herausforderungen und Lösungsvorschläge am Beispiel des Projekts KoKo*, Talk at the Workshop from the "Arbeitsgruppe: Korpusbasierte Linguistik" at the 40. Österreichische Linguistiktagung, Universität Salzburg, Salzburg, Austria, [link].

09.2013 **Aivars Glaznieks and Egon W. Stemle**, *Herausforderungen bei der automatischen Verarbeitung von dialektalen IBK-Daten*, Talk at the Workshop on "Verabeitung und Annotation von Sprachdaten aus Genres internetbasierter Kommunikation" at the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL 2013), TU Darmstadt, Darmstadt, German, [link].

06.2013 **Egon W. Stemle and Verena Lyding**, *The future of BootCaT: A Creative Commons License filter*, Talk at BootCaTters of the world unite! (BOTWU), A workshop (and a survey) on the BootCaT toolkit, Department of Interpreting and Translation, University of Bologna, Forlì, Italy, [link].

02.2013 **Egon W. Stemle and Aivars Glaznieks**, *(Technical Aspects of) Harvesting Data from Social Network Sites*, Talk at the international workshop "Building Corpora of Computer-Mediated Communication: Issues, Challenges, and Perspectives", Department of German Language and Literature, Faculty of Culture Studies, TU Dortmund University, Dortmund, Germany, [link].

07.2012 **Egon W. Stemle**, *Web Corpus Creation and Cleaning*, Plenary Talk at Student Research Workshop Computer Applications in Linguistics (CSRW2012), English Corpus Linguistics Group at the Institute of Linguistics and Literary Studies, Technische Universität Darmstadt, Darmstadt, German, [link].

05.2012 **Egon W. Stemle, Verena Lyding, and Lionel Nicolas**, *On visual Approaches towards Corpus Exploration*, Short presentation at the 3rd workshop of the academic network on "Internet Lexicography", EURAC research, Bozen/Bolzano, Italy, [link].

11.2011 **Massimo Poesio, Eduard Barbu, Egon Stemle, and Christian Girardi**, *Portale Ricerca Umanistica*, Live Demo with Poster at the LiveMemories Final Event - Internet, Memoria e Futuro and The Semantic Way, Povo di Trento, Italy.

Workshops and Conferences

10.2014 **Member of the Programme Committee**, *Workshop*, Natural Language Processing for Computer-Mediated Communication / Social Media (NLP4CMC), Pre-conference workshop at KONVENS 2014, Hildesheim, Germany.

04.2014 **Member of the Programme Committee**, *Workshop*, 9th Web as Corpus Workshop (WAC-9), Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014), Gothenburg, Sweden.

07.2013 **Member of the Organising Committee and Programme Committee**, *Workshop*, 8th Web as Corpus Workshop (WAC-8), Workshop day at the seventh international Corpus Linguistics conference (CL2013), Lancaster University, UK.

09.2006 **Student Worker at OTT06**, *Workshop*, Ontologies in Text Technology: Approaches to Extract Semantic Knowledge from Syntactic Information, jointly organized by the Institute of Cognitive Science at the University of Osnabrück and the project C2 of the distributed DFG-research group Text Technological Modelling of Information.

06.2006 **Student Worker at QITL-2**, *Workshop*, Second Workshop on Quantitative Investigations in Theoretical Linguistics, organized by the Computational Linguistics Group at the Institute of Cognitive Science.

Administrative Experience

since 08.2012 **Secretary of the Special Interest Group on the Web as Corpus (SIGWAC)**, SIG of the Association for Computational Linguistics.

04.2006 – – 03.2007 **Member of the Examination Board (Prüfungsausschuss)**, Cognitive Science Study Programme, University of Osnabrück, Germany.

12.2005 – – 02.2006 **Member of the Search Committee (Besetzungkommission)**, Institute of Cognitive Science, University of Osnabrück, Germany.
BAT IIa research associate in Artificial Intelligence.

04.2005 – – 03.2008 **Member of the Academic Studies Commission (Studienkommission)**, Cognitive Science Study Programme, University of Osnabrück, Germany.

02.2005 – – 01.2006 **Member of the Search Committee (Berufungskommission)**, Faculty of Humanities, University of Osnabrück, Germany.
W3 professor ship Artificial Intelligence and Cognitive Science.

10.2003 – – 03.2008 **Member of the Steering Committee (Vorstand)**, Institute of Cognitive Science, University of Osnabrück, Germany.

07.2003 – – 10.2003 **Member of the Search Committee (Besetzungskommission)**, Institute of Cognitive Science, University of Osnabrück, Germany.
BAT IIa research associate in Artificial Intelligence.

Other Activities

04.2006 **A for Alibi Symposium**, Uqbar Foundation, Utrecht University Museum [link].

12.2005 **Amsterdam 2.0 Exhibition**, Mediamatic, Amsterdam [link].
Technical counsel for Kasper Andreasen & Tine Melzer

03.2003 **Spring School**, Interdisciplinary College 2003 (IK2003), Günne at Lake Möhne.
Focus Theme: Applications, Brains and Computers

2003 **The Complete Dictionary**, Tine Melzer, 26 volumes, A–Z.
Programming and Processing [link]

since 11.2002 **Member of the German Cognitive Science Society (GK e.V.)**, [link].

Languages

German	**Native**	
English	**Excellent**	*Main education language at university.*
French	**Basic**	*5 years training during high-school.*
Italian	**Basic**	

Computer Skills

programming	Python, Java, JavaScript (ECMAScrip), Prolog, ML	scripting	Bash, Perl, PHP
scientific	R, CQP, Oracle/Sun Grid Engine, Apache Hadoop, Matlab, OmniFind+UIMA, PDP++, Octave, Protégé	design	LaTeX, gnuplot, Inkscape, Scribus, InDesign, Illustrator, Photoshop, GIMP
OS	Linux, OS X, Windows 2000/XP/Vista	services	LAMP Stack, PostgreSQL, Apache Tomcat, Exim, Mailman, SpamAssassin, Trac, ISC Bind, Cyrus, Dovecot, OpenVPN, Shorewall, Darwin Calendar Server, FreeIPA, OpenLDAP, NIS/YP, Kerberos+NFSv4