

# egon w. stemle

Researcher

European Academy of Bozen/Bolzano (EURAC)  
Viale Druso, 1  
I-39100 Bolzano (BZ)

✉ [egon.stemle@eurac.edu](mailto:egon.stemle@eurac.edu)  
🌐 [www.eurac.edu/staff/estemle/](http://www.eurac.edu/staff/estemle/)



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 is working on computer aided fabrication of ontologies from large document repositories, the technological feasibility thereof and the utilization of cross-linked structured data in applications, and on tools for editing, processing, and annotating linguistic data.

His driving force is 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 – **Research Fellow**, *LiveMemories sponsored grant, financed by the Provincia Autonoma of Trento,*  
– 01.2012 *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 – **Project Member**, *GoodGaze*, Institute of Cognitive Science, University of Osnabrück, Germany.  
– 2010 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 – **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.  
– 03.2008 Reference to Abstract Objects in Natural Language (OntoRef).
- 04.2006 – **Research Assistant**, Institute of Applied Informatics and Formal Description Methods, University of Karlsruhe, Germany, Knowledge Management Research Group.  
– 03.2008 Maintenance and development of software and scientific counsel; e.g. evaluating CL tools for OmniFind+UIMA and hooking them together.
- 08.2004 – **Research Participant**, *ASADO*, Cooperation between the Universities of Osnabrück and  
– 09.2005 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 – **Student Assistant**, *Institute of Cognitive Science, University of Osnabrück, Germany*, Artificial  
– 12.2003 Intelligence Research Group.  
Maintenance and development of software for the MiLCA (Media intensive learning units for courses in computational linguistics) project.

- 10.2001 – **System Administrator**, *Institute of Cognitive Science, University of Osnabrück, Germany.*
- 03.2009 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 – **MSc in Cognitive Science**, *with distinction*, University of Osnabrück, Germany.
- 03.2009 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 – **Semester abroad**, *ILLC (Institute for Logic, Language and Computation)*, University of Amsterdam, The Netherlands.
- 09.2002
- 10.2000 – **BSc in Cognitive Science**, University of Osnabrück, Germany.
- 01.2006
- 12.1998 – **Apprenticeship as Travel Agent**, WORLDWIDE Touristik, Nürnberg, Germany.
- 07.2000
- 06.1997 **Abitur**, Sigmund-Schuckert-Gymnasium, Nürnberg, Germany.

---

## Teaching Experience

- 2009 – 2011 **Lecture on Web Corpora**, *Course in Text Processing class*, Coordinated by Marco Baroni, University of Trento, 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.  
(yearly)
- 10.2003 – **Artificial Intelligence Tutor**, *Methods of Artificial Intelligence class*, Institute of Cognitive Science, University of Osnabrück, Second year Cognitive Science class.  
– 02.2006  
(every other semester) Weekly tutoring of the lecture material, assessment of programming assignments, homework corrections, and exam preparation.
- 04.2003 – **Computational Linguistics Tutor**, *Introduction to Computational Linguistics class*, Institute of Cognitive Science, University of Osnabrück, First year Cognitive Science class.  
– 07.2006  
(every other semester) Weekly tutoring of the lecture material, preparation and assessment of programming assignments, homework corrections, and exam preparation.
- 10.2002 – **Theoretical Neuroscience Tutor**, *Introduction to Theoretical Neuroscience class*, Institute of Cognitive Science, University of Osnabrück, Second year Cognitive Science class.  
– 02.2003  
Weekly tutoring of the lecture material, homework corrections.
- 10.2001 – **Computer Science Tutor**, *Algorithms class*, Computer Science Department, University of Osnabrück, First year Computer Science class.  
– 02.2002  
Weekly oral assessment of 10 2-people groups and homework corrections.

---

## Publications

- 2014 **Egon Stemle, Alexander Onysko**, *The role of transfer in automated L1 identification of English learner essays*, in *Hagen Peukert, editor: Transfer Effects in Multilingual Language Development*, John Benjamins.  
(in Press)
- 2014 **Aivars Glaznieks, Egon Stemle**, *Challenges of building a CMC corpus for analyzing writer's style by age: The DiDi project*, in *Michael Beisswenger, Nellek Oostdijk, Angelika Storrer, Henk van den Heuvel, editors: Journal for Language Technology and Computational Linguistics (JLCL), Special Issue.*  
(in Press)
- 10.2014 **Jennifer-Carmen Frey, Egon W Stemle, Aivars Glaznieks**, *Collecting language data of non-public social media profiles*, in *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 11–15, Hildesheim, Germany.  
Universitätsverlag Hildesheim
- 09.2014 **Michel Génèreux, Egon W Stemle, Verena Lyding, Lionel Nicolas**, *Correcting OCR errors for German in Fraktur font*, first Italian Computational Linguistics Conference (CLIC-it), Dec 9-10, Pisa, Italy.  
(Accepted)

- 05.2014 **Verena Lyding, Lionel Nicolas, Egon Stemle**, *'interHist' - an interactive visual interface for corpus exploration*, in *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, Egon Stemle**, *KoKo: An L1 Learner Corpus for German*, in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 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, 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, [\[link\]](#)
- 11.2013 **Verena Lyding, Claudia Borghetti, Henrik Dittmann, Lionel Nicolas, Egon Stemle**, *Open Corpus Interface for Italian Language Learning*, in *Conference proceedings. ICT for language learning*, Italy. [libreriauniversitaria.it](#), [\[link\]](#)
- 09.2013 **Lionel Nicolas, Egon W Stemle, Klara Kranebitter, Verena Lyding**, *High-Accuracy Phrase Translation Acquisition Through Battle-Royale Selection*, in *Proceedings of Recent Advances in Natural Language Processing, RANLP 2013*, pages 516–524, Hissar, Bulgaria.  
[\[link\]](#)
- 07.2013 **Stefan Evert, Egon Stemle, Paul Rayson, editors**, *Proceedings of the 8th Web as Corpus Workshop (WAC-8)*, Workshop at the seventh international Corpus Linguistics conference (CL2013), Jul 22, Lancaster, UK.  
[\[link\]](#)
- 06.2013 **Klara Kranebitter, Egon W Stemle**, *Constructing concept relation maps to support building concept systems in comparative legal terminology*, *Terminologie & Ontologie: Théories et Applications (TOTh'2013)*, Jun 6-7, Chamébray, France.  
(in Press)
- 05.2013 **Lionel Nicolas, Egon Stemle, Aivars Glaznieks, 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*, May 16-17, Padova, Italia.  
(in Press)
- 2012 **Aivars Glaznieks, Andrea Abel, Verena Lyding, Lionel Nicolas, Egon Stemle**, *Establishing a Standardised Procedure for Building Learner Corpora*, *Learner Language, Learner Corpora 2012 (LLLC2012)*, Oct 5-6, University of Oulu, Finland, in *Tarja Nikula, Sauli Takala, Sabine Ylönen, editors: Apples - Journal of Applied Language Studies*.  
(in Press)
- 09.2012 **Lionel Nicolas, Egon W Stemle, Klara Kranebitter**, *Towards high-accuracy bilingual phrase acquisition from parallel corpora*, in *Proceedings of KONVENS 2012*, pages 471–479, Vienna, Austria.  
ÖGAI, [\[link\]](#)
- 07.2012 **Francesca Bonin, Fabio Cavulli, Aronne Noriller, Massimo Poesio, Egon W Stemle**, *Annotating Archaeological Texts: An Example of Domain-Specific Annotation in the Humanities*, in *Proceedings of the Sixth Linguistic Annotation Workshop*, 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, Christian Girardi**, *The Humanities Research Portal: Human Language Technology Meets Humanities Publication Archives*, in *SDH2011 - Supporting Digital Humanities: Answering the unaskable*.  
[\[link\]](#)
- 11.2011 **Asif Ekbal, Francesca Bonin, Sriparna Saha, Egon Stemle, Eduard Barbu, Fabio Cavulli, Christian Girardi, Massimo Poesio**, *Rapid Adaptation of NE Resolvers for Humanities Domains using Active Annotation*, *Journal for Language Technology and Computational Linguistics (JLCL)*, 26(2):39–51.  
[\[link\]](#)

- 07.2011 **Brian Murphy, 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, 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, Massimo Poesio**, *Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus*, in *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.  
European Language Resources Association (ELRA), [\[link\]](#)
- 09.2009 **Johannes M Steger, Egon W Stemle**, *KrdWrd — The Architecture for Unified Processing of Web Content*, in *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*, Donostia-San Sebastian, Basque Country.  
[\[link\]](#)
- 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, Stefan Evert**, *FIASCO: Filtering the Internet by Automatic Subtree Classification*, Osnabrück, in *Proceedings of the Third Web as Corpus Workshop (WAC3)*, Louvain-la-Neuve.  
Presses universitaires de Louvain, [\[link\]](#)
- 07.2007 **Sebastian Blohm, Philipp Cimiano, Egon Stemle**, *Harvesting Relations from the Web - Quantifying 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, 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

- 09.2014 (Accepted) **Michel Genereux, Egon Stemle, Lionel Nicolas, Verena Lyding**, *Correcting OCR errors for German in Fraktur font*, CLICit - The first Italian Computational Linguistics Conference, Dec 9-10, Pisa, Italy.
- 08.04.2014 **Alexander Onysko\*, Egon W. Stemle\***, *Automated L1 identification in English learner essays and its implications for language transfer*, Work in Progress, Kompetenzzentrum Sprachen, Freie Universität Bozen, Bozen/Bolzano, Italy.
- 02.2014 **Aivars Glaznieks\*, Egon W. Stemle, Jennifer Frey**, *Nützlichkeit der Normalisierung von Dialekt für automatische Verarbeitungsschritte. Erfahrungen aus dem Projekt DiDi*, 7th workshop of the DFG scientific network Empirikom: Social Media Corpora for the eHumanities: Standards, Challenges, and Perspectives, Feb 19–21, TU Dortmund University, Germany.
- 22.11.2013 **Andrea Abel, Aivars Glazniek, Egon W. Stemle**, *Automatische Annotation von Schülertexten – Herausforderungen und Lösungsvorschläge am Beispiel des Projekts KoKo*, Arbeitsgruppe: Korpusbasierte Linguistik, Workshop at the 40. Österreichische Linguistiktagung, Nov 22-24, Universität Salzburg, Salzburg, Austria.
- 23.09.2013 **Aivars Glazniek, Egon W. Stemle**, *Herausforderungen bei der automatischen Verarbeitung von dialektalen IBK-Daten*, GSCL-Workshop Verarbeitung und Annotation von Sprachdaten aus Genres internet-basierter Kommunikation, Workshop at the International Conference of the German Society for Computational Linguistics and Language Technology, Sep 23, Darmstadt, Germany.

- 24.06.2013 **Egon W. Stemle, Verena Lyding**, *The future of BootCaT: A Creative Commons License filter*, BootCaTters of the world unite! (BOTWU), A workshop (and a survey) on the BootCaT toolkit, Jun 24, Department of Interpreting and Translation, University of Bologna Forlì, Italy.
- 03.2013 **Aivars Glaznieks\*, Egon Stemle, Andrea Abel, Verena Lyding**, *Herausforderungen bei der Erstellung eines L1-Lernerkorpus: Lösungsvorschläge aus dem Projekt "KoKo"*, DGfS 2013 Workshop on Modellierung nicht-standardisierter Schriftlichkeit, Workshop at the 35th Annual Conference of the German Linguistic Society (DGfS 2013), Mar 12-15, Potsdam, Germany.
- 03.2013 **Verena Lyding\*, Lionel Nicolas, Egon Stemle**, *interHist – an interactive visualization for statistically enhanced query structures*, DGfS 2013 Workshop on the Visualization of Linguistic Patterns, Workshop at the 35th Annual Conference of the German Linguistic Society (DGfS 2013), Mar 12-15, Potsdam, Germany.
- 13.02.2013 **Egon W. Stemle, Aivars Glaznieks**, *Technical Aspects in Harvesting Data from Social Network Sites*, International workshop: Building Corpora of Computer-Mediated Communication: Issues, Challenges, and Perspectives, Feb 13-15, Department of German Language and Literature, Faculty of Culture Studies, TU Dortmund University, Germany.
- 02.2013 **Aivars Glaznieks\*, Egon W. Stemle**, *The Project DIDI. Writing on Social Network Sites – A Corpus-based Observation of the Current Language Use in South Tyrol, with Particular Consideration of the Writers' Age*, International workshop: Building Corpora of Computer-Mediated Communication: Issues, Challenges, and Perspectives, Feb 13-15, Department of German Language and Literature, Faculty of Culture Studies, TU Dortmund University, Germany.
- 13.07.2012 **Egon W. Stemle**, *Web Corpus Creation and Cleaning*, Computer Applications in Linguistics: Student Research Workshop (CSRW2012), English Corpus Linguistics Group at the Institute of Linguistics and Literary Studies, Technische Universität Darmstadt, Germany.  
(Plenary Talk)
- 04.05.2012 **Egon W. Stemle, Verena Lyding, Lionel Nicolas**, *On visual Approaches towards Corpus Exploration*, 3rd Workshop of the Academic Network on "Internet Lexicography", EURAC research, Bozen/Bolzano, Italy.  
(Impulsreferat)

---

## 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.
- 09.2003 **Student Worker at EuroCogSci 2003**, *Conference*, First European Cognitive Science Meeting, jointly organized by the Cognitive Science Society and the German Cognitive Science Society.

---

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

- 12.2005 – **Member of the Search Committee (Besetzungskommission)**, Institute of Cognitive Science,  
– 02.2006 University of Osnabrück.  
BAT IIa research associate in Artificial Intelligence.
- 04.2005 – **Member of the Academic Studies Commission (Studienkommission)**, Cognitive Science Study  
– 03.2008 Programme, University of Osnabrück.
- 02.2005 – **Member of the Search Committee (Berufungskommission)**, Faculty of Humanities, University  
– 01.2006 of Osnabrück.  
W3 professor ship Artificial Intelligence and Cognitive Science.
- 10.2003 – **Member of the Steering Committee (Vorstand)**, Institute of Cognitive Science, University of  
– 03.2008 Osnabrück.
- 07.2003 – **Member of the Search Committee (Besetzungskommission)**, Institute of Cognitive Science,  
– 10.2003 University of Osnabrück.  
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**  
 French **Basic**  
 Italian **Basic**

*Main education language at university.*  
*5 years training during high-school.*

## Computer Skills

programming	Python, Java, JavaScript (ECMAScrip), Prolog, ML	scripting	Bash, Perl, PHP
scientific	R, CQP, Oracle/Sun Grid Engine, Apache Hadoop, Matlab, Om- niFind+UIMA, PDP++, Octave, Protégé	design	L <sup>A</sup> T <sub>E</sub> X, gnuplot, Inkscape, Scribus, InDe- sign, Illustrator, Photoshop, GIMP
OS	Linux, OS X, Windows 2000/XP/Vista	services	LAMP Stack, PostgreSQL, Apache Tomcat, Exim, Mailman, SpamAssas- sin, Trac, ISC Bind, Cyrus, Dovecot, OpenVPN, Shorewall, Darwin Calendar Server, FreeIPA, OpenLDAP, NIS/YP, Kerberos+NFSv4