Curriculum Vitae

Prof Serge Sharoff

September 6, 2025

1 Current position

Title Professor of Language Technology and Digital Humanities,

Address Centre for Translation, Interpreting and Localisation Studies
School of Languages, Cultures and Societies
Faculty of Arts, Humanities and Cultures
University of Leeds
Leeds, LS2 9JT, UK

Tel. +44(0)113 343 7287

e-mail s.sharoff@leeds.ac.uk

2 My research interests

Artificial Intelligence and more specifically Large Language Models, such as ChatGPT, have recently made a profound impact on how we interact with the computer by providing the ability to produce new texts in response to prompts. Fundamental research in this area is at the core of my expertise, I've been doing this since my own PhD in the 1990s (the topic was on building language models for information extraction). This pre-dated LLMs, but the idea of linking language to meanings remains the same.

My current research interests focus on the relationship between linguistic structures and language technology, such as the capability of language technology to understand stylistic properties of texts beyond the keywords, how to use language technology to help language learners and professional translators, how to interpret the inner blackboxes of large language models via linguistic structures.

3 Research awards

2024-2026 EU Horizon "iDem: Innovative and Inclusive Democratic Spaces for Deliberation and Participation" (PI for Leeds, €539,954) `https://idemproject.eu/`

2020-2022 EPSRC "AI tracing for detecting COVID19 misinformation" (PI, £39,954)

2019 EPSRC/Alan Turing Institute "Translation Workflow Automation by Determining Text Difficulty" (PI, £33,337), with UN and WTO.

2014-2016 InnovateUK grant "Palodiem: Process Automation for Localisation of Dialogue in Entertainment Media" (PI, £216,411)

2011-2014 EU Marie-Curie grant "HyghTra: High-quality Hybrid Machine Translation" (CoI, €541,575)

2012 JISC grant "#IICT: Increasing Interoperability between Corpus Tools" (PI for Leeds, in collaboration with Coventry and Lancaster)

2010-2011 AHRC grant "Intellitext: Intelligent Tools for Creating and Analysing Electronic Text Corpora" (CoI, £159,293)

2010-2012 EU FP7 grant, "TTC: Terminology Extraction, Translation Tools and Comparable Corpora" (PI for Leeds, €369,786)

2010-2012 EU FP7 grant, "Accurat: Machine Translation for Under-resourced languages" (CoI for Leeds, €340,174)

2010-2011 EU LLP grant, "Kelly: Keywords for Language Learners" (PI for Leeds, €124,478)

2009 Google Research Award for Automatic Web documents classification ($75,000)

2005-2007 EPSRC grant "ASSIST: Automated assistance for translators" (CoI, £124,000, in collaboration with Lancaster)

2000-2002 Fellowship from the Alexander von Humboldt Foundation, Germany

1997-2000 EU INCO-Copernicus grant, "AGILE: Adaptive Multilingual Generation" (CoI)

4 Timeline of my academic history

2003–now Research Fellow, Lecturer, Senior Lecturer, Professor, University of Leeds

2000-2002 Alexander von Humboldt Fellow, University of Bielefeld

1990-2000 Research Assistant – Senior Research Fellow, Russian Research Institute for Artificial Intelligence.

5 Education

May, 1997 PhD in Computer Science, Moscow Lomonosov State University (the thesis title was "An instrumental tool for development of linguistic processors").

1985-1990 MSc in mathematics, Urals Gorky State University.

6 Academic publications

Over 150 academic publications with 4675 citations in total, see a separate list below.

`https://scholar.google.co.uk/citations?user=qcnf4QsAAAAJ`

See also: publications.pdf

I'm on the editorial board of the journals Language Resources and Register Studies, as well as on the Programme Committte for a number of conferences.

7 PhD Supervision

27 PhD students supervised to successful completion. My past PhD sudents work in a range of jobs in academia and in industry over the entire world.

Serge Sharoff, Publications

[1] Nouran Khallaf, Stefan Bott, Carlo Eugeni, John O'Flaherty, Serge Sharoff, and Horacio Saggion. Democracy made easy: Simplifying complex topics to enable democratic participation. In *Proc Artificial Intelligence and Easy and Plain Language in Institutional Context at Machine Translation Summit*, Geneva, June 2025.

[2] Lama Alqurashi, Serge Sharoff, Janet Watson, and Jacob Blakesley. BERT-based classical Arabic poetry authorship attribution. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6105–6119, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[3] Nouran Khallaf, Carlo Eugeni, and Serge Sharoff. Reading between the lines: A dataset and a study on why some texts are tougher than others. In Michael Zock, Kentaro Inui, and Zheng Yuan, editors, *Proceedings of the First Workshop on Writing Aids at the Crossroads of AI, Cognitive Science and NLP (WRAICOGS 2025)*, pages 24–34, Abu Dhabi, UAE, January 2025. International Committee on Computational Linguistics.

[4] Dmitri Roussinov, Serge Sharoff, and Nadezhda Puchnina. Controlling out-of-domain gaps in LLMs for genre classification and generated text detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3329–3344, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[5] Yuqian Dai, Serge Sharoff, and Marc De Kamps. Graphic improvements: Adding explicit syntactic graphs to neural machine translation. *Neurosymbolic Artificial Intelligence*, 1:29498732251340044, 2025.

[6] Serge Sharoff. Form and function: automatic methods for prediction of functions. In Rebekah Wegener, Anne McCabe, Akila Sellami-Baklouti, and Lise Fontaine, editors, *Transdisciplinary Systemic Functional Linguistics*. Routledge, 2025.

[7] Nurbanu Aksoy, Nishant Ravikumar, and Serge Sharoff. Enhancing image-to-text generation in radiology reports through cross-modal multi-task learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5977–5985, Torino, Italia, May 2024. ELRA and ICCL.

[8] Horacio Saggion, John O'Flaherty, Thomas Blanchet, Serge Sharoff, Silvia Sanfilippo, Lian Muñoz, Martin Gollegger, Almudena Rascón, José Luis Martí, and Sandra Szasz. Making democratic deliberation and participation more accessible: the iDEM project. In *SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations*, A Coruña, Spain, May 2024.

[9] Nurbanu Aksoy, Serge Sharoff, Selcuk Baser, Nishant Ravikumar, and Alejandro Frangi. Beyond images: an integrative multi-modal approach to chest x-ray report generation. *Frontiers in Radiology*, 4, 2024.

[10] Souad Boumechaal and Serge Sharoff. Attitudes, communicative functions, and lexicogrammatical features of anti-vaccine discourse on Telegram. *Applied Corpus Linguistics*, 4, 2024.

[11] Sarah Miller, Serge Sharoff, Geoffrey Hall, and Prabhu Arumugam. Evaluating text pre-processing strategies for clinical document classification with BERT. In *Proc AIHealth 2024 - The First International Conference on AI-Health*, 2024.

[12] Andreas van Cranenburgh, Laura Allen, Serge Sharoff, and Karina van Dalen-Oskam. Computational methods for the analysis of fiction genres. In Ninke Stukker, John A Bateman, Danielle McNamara, and Wilbert Spooren, editors, *Multidisciplinary Views on Discourse Genre*, pages 135–167. Routledge, 2024.

[13] Dmitri Roussinov and Serge Sharoff. BERT goes off-topic: Investigating the domain transfer challenge using genre classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, December 2023.

[14] Mikhail Lepekhin and Serge Sharoff. FTD at SemEval-2023 task 3: News genre and propaganda detection by comparing mono-and multilingual models with fine-tuning on additional data. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 549–555, 2023.

[15] Dmitri Roussinov, Serge Sharoff, and Nadezhda Puchnina. Fine-tuning language models to recognize semantic relations. *Language Resources and Evaluation*, pages 1–24, 2023.

[16] Serge Sharoff, Reinhard Rapp, and Pierre Zweigenbaum. *Building and Using Comparable Corpora for Multilingual Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Springer Nature, 2023.

[17] Nouran Khallaf, Serge Sharoff, and Rasha Soliman. Towards Arabic sentence simplification via classification and generative approaches. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 43–52, Abu Dhabi, United Arab Emirates, December 2022.

[18] Valeriy Lobov, Alexandra Ivoylova, and Serge Sharoff. Applying natural annotation and curriculum learning to named entity recognition for under-resourced languages. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4468–4480, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[19] Yuqian Dai, Marc de Kamps, and Serge Sharoff. BERTology for machine translation: What BERT knows about linguistic difficulties for translation. In *Proc LREC*, Marseilles, June 2022.

[20] Mikhail Lepekhin and Serge Sharoff. Estimating confidence of predictions of individual classifiers and theirensembles for the genre classification task. In *Proceedings of the Language Resources and Evaluation Conference*, pages 5974–5982, Marseille, France, June 2022. European Language Resources Association.

[21] Jose Sosa and Serge Sharoff. Multimodal pipeline for collection of misinformation data from telegram. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1480–1489, Marseille, France, June 2022. European Language Resources Association.

[22] Serge Sharoff. What neural networks know about linguistic complexity. *Russian Journal of Linguistics*, 26(2):371–390, 2022.

[23] Nouran Khallaf and Serge Sharoff. Automatic difficulty classification of Arabic sentences. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 105–114, Kyiv, Ukraine (Virtual), April 2021. Association for Computational Linguistics.

[24] Serge Sharoff. Genre annotation for the web: text-external and text-internal perspectives. *Register studies*, 3:1–32, 2021.

[25] Serge Sharoff. Topography of internet corpora. In *Proc Dialogue, Russian International Conference on Computational Linguistics*, Moscow, June 2020.

[26] Dmitri Roussinov, Serge Sharoff, and Nadezhda Puchnina. Recognizing semantic relations: Error analysis of the use of transformers vs. recurrent path models. In *Proc LREC*, Online, May 2020.

[27] Serge Sharoff. Know thy corpus! Robust methods for digital curation of Web corpora. In *Proc LREC*, Online, May 2020.

[28] Yu Yuan and Serge Sharoff. Sentence level human translation quality estimation with attention-based neural networks. In *Proc LREC*, Online, May 2020.

[29] Dmitri Roussinov, Serge Sharoff, and Nadezhda Puchnina. Recognizing semantic relations: Attention-based transformers vs. recurrent models. In *Proc European Conference on Information Retrieval (ECIR)*, Lisbon, April 2020.

[30] Serge Sharoff. Finding next of kin: Cross-lingual embedding spaces for related languages. *Journal of Natural Language Engineering*, 26:163–182, 2020.

[31] Maria Kunilovskaya and Serge Sharoff. Building functionally similar corpus resources for translation studies. In *Proc RANLP*, pages 583–592, Varna, September 2019.

[32] Nouran Khallaf, Serge Sharoff, and Michael Ingleby. Towards an Arabic text simplification system. In *Proc Corpus Linguistics*, Cardiff, Jul 2019.

[33] Tagir Gadzhiev and Serge Sharoff. Genre-shift detection using functional text dimensions. In *Proc Student Research Workshop at Dialogue, Russian Computational Linguistics Conference*, Moscow, June 2019.

[34] Mikhail Bulygin and Serge Sharoff. Applying an automatic FTD classifier to the annotation of the GICR corpus. In *Proc Dialogue, Russian International Conference on Computational Linguistics*, 2019.

[35] Serge Sharoff. Language adaptation experiments via cross-lingual embeddings for related languages. In *Proc LREC*, Miyazaki, Japan, May 2018.

[36] Yu Yuan, Yuze Gao, Yue Zhang, and Serge Sharoff. Cross-lingual terminology extraction for translation quality estimation. In *Proc LREC*, Miyazaki, Japan, May 2018.

[37] Yu Yuan and Serge Sharoff. Investigating the influence of bilingual phraseology on trainee translation quality. In *Proc LREC*, Miyazaki, Japan, May 2018.

[38] Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. A multilingual dataset for evaluating parallel sentence extraction from comparable corpora. In *Proc LREC*, Miyazaki, Japan, May 2018.

[39] Mikhail Bulygin and Serge Sharoff. Using machine translation for automatic genre classification in Arabic. In *Proc Dialogue, Russian International Conference on Computational Linguistics*, 2018.

[40] Serge Sharoff. Functional text dimensions for the annotation of Web corpora. *Corpora*, 13(1):65–95, 2018.

[41] Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proc Tenth Workshop on Building and Using Comparable Corpora*, Vancouver, August 2017.

[42] Serge Sharoff. Toward Pan-Slavic NLP: Some experiments with Language Adaptation. In *Proc Workshop on Balto-Slavic Natural Language Processing at EACL*, Valencia, April 2017.

[43] Dagmar Divjak, Serge Sharoff, and Tomaž Erjavec. Slavic corpus and computational linguistics. *Journal of Slavic Linguistics*, 25(2):171–198, 2017.

[44] Reinhard Rapp, Michael Zock, and Serge Sharroff. New areas of application of comparable corpora. In Inguna Skadina, Rob Gaizauskas, Bogdan Babych, Nikola Ljubešić, Dan Tufis, and Andreis Vasiljevs, editors, *Using Comparable Corpora for Under-Resourced Areas of Machine Translation*. Springer Verlag, 2017.

[45] Serge Sharoff. Corpus and systemic functional linguistics. In Tom Bartlett and Gerard O'Grady, editors, *The Routledge Handbook of Systemic Functional Linguistics*, pages 533–546. Routledge, 2017.

[46] Serge Sharoff, Dirk Goldhahn, and Uwe Quasthoff. *Frequency Dictionary: Russian*, volume 9 of *Frequency Dictinaries*. Leipziger Universitätsverlag, 2017. Uwe Quasthoff, Sabine Fiedler, Erla Hallsteindóttir (editors).

[47] Yu Yuan, Bogdan Babych, and Serge Sharoff. Reference-free system for automated human translation quality estimation. In *Proc Information Systems and Technologies (CISTI)*, Lisbon, 2017.

[48] Yu Yuan and Serge Sharoff. Review of 'new directions in empirical translation process research: Exploring the CRITT TPR-DB' edited by Michael Carl et al. *Babel*, 63(3):450–456, 2017.

[49] Yu Yuan, Serge Sharoff, and Bogdan Babych. MoBiL: A hybrid feature set for automatic human translation quality assessment. In *Proc Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, May 2016.

[50] Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. Towards preparation of the second BUCC shared task: Detecting parallel sentences in comparable corpora. In *Proc Ninth Workshop on Building and Using Comparable Corpora*, Portoroz, Slovenia, May 2016.

[51] Noushin Rezapour Asheghi, Serge Sharoff, and Katja Markert. Crowdsourcing for web genre annotation. *Language Resources and Evaluation*, pages 1–39, 2016.

[52] Bogdan Babych and Serge Sharoff. Ukrainian part-of-speech tagger for hybrid MT: Rapid induction of morphological disambiguation resources from a closely related language. In *Proc Fifth Workshop on Hybrid Approaches to Translation (HyTra)*, 2016.

[53] Erika Dalan and Serge Sharoff. Genre classification for a corpus of academic webpages. In *Proc SIGWAC-X*, pages 90–98, Berlin, 2016.

[54] Sophiko Daraselia and Serge Sharoff. Enriching Georgian dictionary entries with frequency information. In *Proc EURALEX*, Tbilisi, 2016.

[55] Roger Evans, Alexander Gelbukh, Gregory Grefenstette, Patrick Hanks, Miloš Jakubíček, Diana McCarthy, Martha Palmer, Ted Pedersen, Michael Rundell, Pavel Rychlỳ, Serge Sharoff, and David Tugwell. Adam kilgarriff's legacy to computational linguistics and beyond. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 3–25. Springer, 2016.

[56] Reinhard Rapp, Serge Sharoff, and Pierre Zweigenbaum. Recent advances in machine translation using comparable corpora. *Natural Language Engineering*, 22:501–516, 2016.

[57] Miguel Rios and Serge Sharoff. Language adaptation for extending post-editing estimates for closely related languages. *The Prague Bulletin of Mathematical Linguistics*, 106:181–192, 2016.

[58] Danil Selegey, Tatiana Shavrina, Vladimir Selegey, and Serge Sharoff. Automatic morphological tagging of Russian social media corpora: training and testing. In *Proc Dialogue, Russian International Conference on Computational Linguistics*, 2016.

[59] Miguel Rios and Serge Sharoff. Large scale translation quality estimation. In *Proc Deep Machine Translation Workshop*, Prague, September 2015.

[60] Miguel Rios and Serge Sharoff. Obtaining SMT dictionaries for related languages. In *Proc the Eighth Workshop on Building and Using Comparable Corpora*, pages 68–73, Beijing, China, July 2015.

[61] Serge Sharoff, Pierre Zweigenbaum, and Reinhard Rapp. BUCC shared task: Cross-language document similarity. In *Proc Eighth Workshop on Building and Using Comparable Corpora*, pages 74–78, Beijing, China, July 2015. Association for Computational Linguistics.

[62] Anisya Katinskaya and Serge Sharoff. Applying multi-dimensional analysis to a Russian webcorpus: Searching for evidence of genres. In *Proc BSNLP*, Sofia, 2015.

[63] M.B. Lagutin, A.Y. Katinskaya, V.P. Selegey, S. Sharoff, and A.A. Sorokin. Automatic classification of web texts using functional text dimensions. In *Proc Dialogue, Russian International Conference on Computational Linguistics*, volume 1, pages 398–413, 2015.

[64] Serge Sharoff. Approaching genre classification via syndromes. In *Proc Corpus Linguistics*, Lancaster, 2015.

[65] Serge Sharoff. Review of 'Web Corpus Construction' by Roland Schäfer and Felix Bildhauer. *Computational Linguistics*, 41, 2015.

[66] Serge Sharoff, Vladimir Belikov, Nikolay Kopylov, Alexey Sorokin, and Tatyana Shavrina. Corpus with automatically resolved morphological ambiguity: towargs methodology of linguistic research. In *Proc Dialogue, Russian International Conference on Computational Linguistics*, 2015.

[67] Noushin Rezapour Asheghi, Katja Markert, and Serge Sharoff. Semi-supervised graph-based genre classification for web pages. In *Proc TextGraphs-9: the workshop on Graph-based Methods for Natural Language Processing*, pages 39–47, Doha, Qatar, October 2014.

[68] Marilena Di Bari, Serge Sharoff, and Martin Thomas. Multiple views as aid to linguistic annotation error analysis. In *Proc LAW VIII - The 8th Linguistic Annotation Workshop*, Dublin, Ireland, August 2014.

[69] Ran Xu and Serge Sharoff. Evaluating term extraction methods for interpreters. In *Proc 4th International Workshop on Computational Terminology (Computerm)*, pages 86–93, Dublin, Ireland, August 2014.

[70] Noushin Rezapour Asheghi, Serge Sharoff, and Katja Markert. Designing and evaluating a reliable corpus of web genres via crowd-sourcing. In *Proc Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, may 2014.

[71] Reinhard Rapp and Serge Sharoff. Extracting multiword translations from aligned comparable documents. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra)*, pages 87–95, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

[72] Nikolay Belikov, Vladimir Kopylov, Vladimir Selegey, and Serge Sharoff. Variational corpus statistics using author profiles. In *Proc Dialogue, Russian International Conference on Computational Linguistics*, Bekasovo, 2014.

[73] Sophiko Daraselia and Serge Sharoff. Towards creating a large corpus for Georgian. In *Proc* 7^{th} *Biennial IVACS Conference*, Newcastle, 2014.

[74] Richard Forsyth and Serge Sharoff. Document dissimilarity within and across languages: a benchmarking study. *Literary and Linguistic Computing*, 29:6–22, 2014.

[75] Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation*, 48(1):121–163, 2014.

[76] Serge Sharoff, Stefania Spina, and Sofie Johansson-Kokkinakis. Introduction to the special issue on resources and tools for language learners. *Language Resources and Evaluation*, 48(1):1–3, 2014.

[77] Alexey Sorokin, Anisya Katinskaya, and Serge Sharoff. Associating symptoms with syndromes: Reliable genre annotation for a large Russian webcorpus. In *Proc Dialogue, Russian International Conference on Computational Linguistics*, Bekasovo, 2014.

[78] James Wilson, Serge Sharoff, Paul Stephenson, and Anthony Hartley. Innovative methods for LSP-teaching: How we use corpora to teach business Russian. In *Languages for Specific Purposes in the Digital Era*, pages 197–222. Springer, 2014.

[79] Pavel Braslavski, Alexander Beloborodov, Maxim Khalilov, and Serge Sharoff. English-to-Russian MT evaluation campaign. In *Proc* 51^{st} *Annual Meeting of the ACL*, pages 262–267, Sofia, Bulgaria, August 2013.

[80] Serge Sharoff, Reinhard Rapp, Pierre Zweigenbaum, and Pascale Fung, editors. *BUCC: Building and Using Comparable Corpora*. Springer, 2013.

[81] Vladimir Belikov, Nikolay Kopylov, Alexander Piperski, Vladimir Selegey, and Serge Sharoff. Corpus as language: from scalability to register variation. In *Proc Dialogue, Russian International Conference on Computational Linguistics*, Bekasovo, 2013.

[82] Pavel Braslavski, Alexander Beloborodov, Maxim Khalilov, and Serge Sharoff. ROMIP MT evaluation track 2013: Organizers' report. In *Proc Dialogue, Russian International Conference on Computational Linguistics*, Bekasovo, 2013.

[83] Marilena Di Bari, Serge Sharoff, and Martin Thomas. Sentiml: functional annotation for multilingual sentiment analysis. In *Proc 1st International Workshop on Collaborative Annotations in Shared Environment: metadata, vocabularies and techniques in the Digital Humanities*, pages 15–22, 2013.

[84] Alexander Piperski, Vladimir Belikov, Nikolay Kopylov, Vladimir Selegey, and Serge Sharoff. Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In *Proc* 8^{th} *Web as Corpus Workshop (WAC-8)*, 2013.

[85] Serge Sharoff. Measuring the distance between comparable corpora between languages. In Serge Sharoff, Reinhard Rapp, Pierre Zweigenbaum, and Pascale Fung, editors, *BUCC: Building and Using Comparable Corpora*, pages 113–129. Springer, Berlin/New York, 2013.

[86] Serge Sharoff, Reinhard Rapp, and Pierre Zweigenbaum. Overviewing important aspects of the last twenty years of research in comparable corpora. In Serge Sharoff, Reinhard Rapp, Pierre Zweigenbaum, and Pascale Fung, editors, *BUCC: Building and Using Comparable Corpora*, pages 1–17. Springer, 2013.

[87] Serge Sharoff, Elena Umanskaya, and James Wilson. *A frequency dictionary of Russian: core vocabulary for learners*. Routledge, London, 2013.

[88] Bogdan Babych, Kurt Eberle, Johanna Geiß, Mireia Ginestí-Rosell, Anthony Hartley, Reinhard Rapp, Serge Sharoff, and Martin Thomas. Design of a hybrid high quality machine translation system. In *Proc Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 101–112, Avignon, France, 2012.

[89] Richard Forsyth and Serge Sharoff. Quantifying document dissimilarity within and across languages: a benchmarking trial. In *Proc 6th Inter-Varietal Applied Corpus Studies (IVACS) group International Conference on Corpora across Linguistics*, Leeds, UK, 2012.

[90] Tatiana Gornostay, Anita Gojun, Marion Weller, Ulrich Heid, Emmanuel Morin, Béatrice Daille, Helena Blancafort, Serge Sharoff, and Claude Méchoulam. Terminology extraction, translation tools and comparable corpora: Ttc concept, midterm progress and achieved results. In *Workshop on Creating Cross-language Resources for Disconnected Languages and Styles at LREC*, Istanbul, 2012.

[91] Reinhard Rapp, Serge Sharoff, and Bogdan Babych. Identifying word translations from comparable documents without a seed lexicon. In *Proc Eighth Language Resources and Evaluation Conference, LREC*, Istanbul, 2012.

[92] Serge Sharoff and Anthony Hartley. Lexicography, terminology and ontologies. In Alexander Mehler, Laurent Romary, and Dafydd Gibbon, editors, *Handbook of Technical Communication (HAL 8)*, pages 317–346. Mouton de Gruyter, Berlin, 2012.

[93] Siva Reddy and Serge Sharoff. Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources. In *Proc IJCNLP'11*, Chiang Mai, Thailand, November 2011.

[94] Helena Blancafort, Ulrich Heid, Tatiana Gornostay, Claude Méchoulam, Béatrice Daille, and Serge Sharoff. User-centred views on terminology extraction tools: Usage scenarios and integration into MT and CAT tools. In *Proc TRALOGY Conference "Translation Careers and Technologies: Convergence Points for the Future"*, 2011.

[95] Richard Forsyth and Serge Sharoff. From crawled collections to comparable corpora: An approach based on automatic archetype identification. In *Proc Corpus Linguistics Conference*, Birmingham, 2011.

[96] Serge Sharoff and Joakim Nivre. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In *Proc Dialogue, Russian International Conference on Computational Linguistics*, Bekasovo, 2011.

[97] James Wilson, Serge Sharoff, and Paul Stephenson. "Value added teaching": corpus-based methods for LSP teaching. In N. Talaván, E. Martín Monje, and F. Palazón, editors, *Technological Innovation in the Teaching and Processing of LSPs*. Madrid, 2011.

[98] James Wilson, Anthony Hartley, Serge Sharoff, and Paul Stephenson. Advanced corpus solutions for humanities researchers. In *Proc Advanced Corpus Solutions, PACLIC 24*, pages 36–43, Tohoku University, November 2010.

[99] Serge Sharoff. Balancing form and function in corpus research. *International Journal of Corpus Linguistics*, 15(2):412–417, September 2010.

[100] Paul Rayson, Scott Piao, Serge Sharoff, Stefan Evert, and Begonia Villada Moirón. Multiword expressions: hard going or plain sailing. *Language Resources and Evaluation Journal*, 44:1–7, January 2010.

[101] Alexander Mehler, Serge Sharoff, and Marina Santini, editors. *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York, 2010.

[102] Helena Blancafort, Béatrice Daille, Tatiana Gornostay, Ulrich Heid, Claude Méchoulam, and Serge Sharoff. TTC: Terminology extraction, translation tools and comparable corpora. In *Proc EURALEX2010*, Leeuwarden, 5-6 July 2010.

[103] Yunhua Qu, Nan Zhang, Serge Sharoff, Yuting Yang, Ruoyuan Gao, and Chengzhi Xu. Using an integrated feature set to generalize and justify the chinese-to-english transferring rule of "zhe" aspect. *Journal of Computer Science and Technology (PRC)*, 2010.

[104] Marina Santini, Alexander Mehler, and Serge Sharoff. Riding the rough waves of genre on the web. In Alexander Mehler, Serge Sharoff, and Marina Santini, editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York, 2010.

[105] Serge Sharoff. Analysing similarities and differences between corpora. In *Proc 7th Conference of Language Technologies (Jezikovne Tehnologije)*, Ljubljana, 2010.

[106] Serge Sharoff. In the garden and in the jungle: Comparing genres in the BNC and Internet. In Alexander Mehler, Serge Sharoff, and Marina Santini, editors, *Genres on the Web: Computational Models and Empirical Studies*, pages 149–166. Springer, Berlin/New York, 2010.

[107] Serge Sharoff, Zhili Wu, and Katja Markert. The Web library of Babel: evaluating genre collections. In *Proc Seventh Language Resources and Evaluation Conference, LREC*, Malta, 2010.

[108] Zhili Wu, Katja Markert, and Serge Sharoff. Fine-grained genre classification using structural learning algorithms. In *Proc ACL 2010*, Uppsala, 2010.

[109] Bogdan Babych, Anthony Hartley, and Serge Sharoff. Evaluation-guided pre-editing of source text: improving MT-tractability of light verb constructions. In *Proc 13th European Association for Machine Translation*, pages 36–43, Barcelona, May 2009.

[110] Eric Atwell, Latifa Al-Sulaiti, and Serge Sharoff. Arabic and arab english in the arab world. In *Proc CL2009 International Conference on Corpus Linguistics*, 2009.

[111] Olga Lyashevskaya and Serge Sharoff. *Chastotny slovar sovremennogo russkogo yazyka*. Azbukovnik, Moscow, 2009.

[112] Marina Santini, Georg Rehm, Serge Sharoff, and Alexander Mehler. Automatic genre identification: Issues and prospects. *Journal for Language Technology and Computational Linguistics*, 24(1), 2009.

[113] Marina Santini and Serge Sharoff. Web genre benchmark under construction. *Journal for Language Technology and Computational Linguistics*, 25(1):125–141, 2009.

[114] Serge Sharoff, Bogdan Babych, and Anthony Hartley. 'irrefragable answers': using comparable corpora to retrieve translation equivalents. *Language Resources and Evaluation Journal*, 43:15–25, 2009.

[115] Bogdan Babych, Anthony Hartley, and Serge Sharoff. Generalising lexical translation strategies for MT using comparable corpora. In *Proc Sixth Language Resources and Evaluation Conference, LREC*, Marrakech, 2008.

[116] Marco Baroni, Francis Chantree, Adam Kilgarriff, and Serge Sharoff. Cleaneval: a competition for cleaning web pages. In *Proc Sixth Language Resources and Evaluation Conference, LREC*, pages 638–643, Marrakech, 2008.

[117] Svitlana Kurella, Anthony Hartley, and Serge Sharoff. Corpus-based tools for computer-assisted acquisition of reading abilities in cognate languages. In *Proc of the Sixth Language Resources and Evaluation Conference, LREC*, Marrakech, 2008.

[118] Svitlana Kurella, Serge Sharoff, and Anthony Hartley. Rhetorical text structure in acquiring reading skills in L3. In *Proc of Teaching and Language Corpora Conference, TaLC 2008*, Lisbon, 2008.

[119] Olga Mudraya, Scott S. L. Piao, Paul Rayson, Serge Sharoff, Bogdan Babych, and Laura Löfberg. Automatic extraction of translation equivalents of phrasal and light verbs in English and Russian. In S. Granger and F. Meunier, editors, *Phraseology: an interdisciplinary perspective*, pages 293–309. John Benjamins, 2008.

[120] Serge Sharoff, Mikhail Kopotev, Tomaž Erjavec, Anna Feldman, and Dagmar Divjak. Designing and evaluating a Russian tagset. In *Proc Sixth Language Resources and Evaluation Conference, LREC*, Marrakech, 2008.

[121] Serge Sharoff, Svitlana Kurella, and Anthony Hartley. Seeking needles in the Web haystack: finding texts suitable for language learners. In *Proc Teaching and Language Corpora Conference, TaLC 2008*, Lisbon, 2008.

[122] Bogdan Babych, Anthony Hartley, and Serge Sharoff. A dynamic dictionary for discovering indirect translation equivalents. In *Translating and the Computer 29. Proceedings of the Twenty-ninth International Conference on Translating and the Computer, 29-30 November 2007, Londres*, pages 1–21, 2007.

[123] Bogdan Babych, Anthony Hartley, and Serge Sharoff. Translating from under-resourced languages: comparing direct transfer against pivot translation. In *Proc MT Summit XI*, pages 412–418, Copenhagen, 2007.

[124] Bogdan Babych, Anthony Hartley, Serge Sharoff, and Olga Mudraya. Assisting translators in indirect lexical transfer. In *Proc 45th Annual Meeting of the ACL*, pages 739–746, Prague, 2007.

[125] Judy Delin, Catherine Barnes, Stephen Lillford, and Serge Sharoff. Linguistic support for concept selection decisions. *AIEDAM: Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 21:123–135, 2007.

[126] Nick Pendar and Serge Sharoff. Supervised lexical acquisition for persian from a web corpus. In *Proc Computational Approaches to Arabic Script-based Languages*, pages 106–113, Stanford, 2007.

[127] Serge Sharoff. Central planning vs. free market: comparing the distribution of topics and genres in the Russian National Corpus and Internet. In *Proc Dialogue, Russian International Conference on Computational Linguistics*, pages 573–581, Bekasovo, 2007.

[128] Serge Sharoff. Classifying web corpora into domain and genre using automatic feature identification. In *Proc Web as Corpus Workshop*, Louvain-la-Neuve, 2007.

[129] Serge Sharoff and Alina Secara. Who is talking to whom? Linguistic cues for engaging with the audience. *CoDesign, International Journal of CoCreation in Design and the Arts*, 3(2):175–182, 2007.

[130] Dragoş Ciobanu, Anthony Hartley, and Serge Sharoff. Using richly annotated trilingual language resources for acquiring reading skills in a foreign language. In *Proc Fifth Language Resources and Evaluation Conference, LREC*, Genoa, 2006.

[131] Serge Sharoff. Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna, 2006.

[132] Serge Sharoff. How to handle lexical semantics in SFL: a corpus study of purposes for using size adjectives. In Geoff Thompson and Susan Hunston, editors, *System and Corpus: exploring connections*, pages 184–205. Equinox, 2006.

[133] Serge Sharoff. Open-source corpora: using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462, 2006.

[134] Serge Sharoff. Translation as problem solving: uses of comparable corpora. In *Proc Third International Workshop on Language Resources for Translation Work, Research and Training at LREC*, pages 24–28, Genoa, 2006.

[135] Serge Sharoff. A uniform interface to large-scale linguistic resources. In *Proc Fifth Language Resources and Evaluation Conference, LREC*, pages 539–542, Genoa, 2006.

[136] Serge Sharoff, Bogdan Babych, and Anthony Hartley. Using collocations from comparable corpora to find translation equivalents. In *Proc Fifth Language Resources and Evaluation Conference, LREC*, pages 465–470, Genoa, 2006.

[137] Serge Sharoff, Bogdan Babych, and Anthony Hartley. Using comparable corpora to solve problems difficult for human translators. In *Proc International Confenrence on Computational Linguistics and Association of Computational Linguistics, COLING-ACL 2006*, pages 739–746, Sydney, 2006.

[138] Serge Sharoff, Bogdan Babych, Paul Rayson, Olga Mudraya, and Scott Piao. ASSIST: Automated semantic assistance for translators. In *Companion Volume to Proc European Association of Computational Linguistics*, pages 139–142, Trento, 2006.

[139] Dragoş Ciobanu, Anthony Hartley, and Serge Sharoff. Acquiring reading skills in a foreign language in a multilingual, corpus-based environment. In *Proc Use of New Technologies in Foreign Language Teaching (UNTELE 2005)*, Compiegne, 2005.

[140] D. Dobrovol'skij, A. Kretov, and S. Sharoff. A corpus of parallel texts: architecture and usage parameters. In *Russian National Corpus: 2003-2005. Results and perspectives*. Indrik, 2005. (in Russian).

[141] Serge Sharoff. The communicative potential of verbs of "away-from" motion in English, German and Russian. *Functions of language*, 12(2):205–240, 2005.

[142] Serge Sharoff. Methods and tools for development of the Russian Reference Corpus. In Dawn Archer, Andrew Wilson, and Paul Rayson, editors, *Corpus Linguistics Around the World*, pages 167–180. Rodopi, Amsterdam, 2005.

[143] Serge Sharoff. Phenomenology and cognitive science. In S. Franchi and G. Guzeldere, editors, *Mechanical Bodies, Computational Minds*, pages 471–487. MIT Press, Cambridge, 2005.

[144] Serge Sharoff. Harnessing the lawless: using comparable corpora to find translation equivalents. *Journal of Applied Linguistics*, 1(3):333–350, 2004.

[145] Serge Sharoff. Towards basic categories for describing properties of texts in a corpus. In *Proc Forth Language Resources and Evaluation Conference, LREC*, Lisbon, 2004.

[146] Serge Sharoff. What is at stake: a case study of Russian expressions starting with a preposition. In *Proc ACL04 Workshop Multiword Expressions: Integrating Processing*, pages 17–23, Barcelona, 2004.

[147] Serge Sharoff, Anthony Hartley, and Peter Llewellyn-Jones. Sentence generation in British Sign Language. In *Extended abstracts of posters presented at the Third International Conference on Natural Language Generation*, pages 40–43, Brighton, 2004.

[148] Serge Sharoff. K sozdaniyu predstavitelnogo corpus sovremennogo russkogo yazyka. In *Proc Dialogue, Russian International Conference on Computational Linguistics*, pages 1–5, Bekasovo, 2003. (in Russian).

[149] Serge Sharoff. Predstavitel'nyj corpus russkogo jazyka v kontekste mirovogo opyta. *NTI, series 2*, 5:8–19, 2003.

[150] Serge Sharoff. 'when i use a word, it means just what i choose it to mean'. In C. Inchaurralde and C. Florén, editors, *Interaction and cognition in linguistics*, pages 277–293. Peter Lang, 2003.

[151] Serge Sharoff. Meaning as use: exploitation of aligned corpora for the contrastive study of lexical semantics. In *Proc Language Resources and Evaluation Conference (LREC)*, pages 447–452, May 2002. Las Palmas, Spain.

[152] Dafydd Gibbon, Thorsten Trippel, and Serge Sharoff. Concordancing for parallel spoken language corpora. In *Proc Eurospeech 2001*, pages 2059–2062, 2001.

[153] Serge Sharoff. Lexis: between the grammar and the domain model. In A. Melby and A. Lommel, editors, *Selected papers presented at the 26th LACUS Forum*, pages 369–380. LACUS, Chapel Hill, NC, 2000.

[154] Serge Sharoff and Vlad Zhigalov. Register-domain separation as a methodology for development of natural language interfaces to databases. In Angela Sasse and Chris Johnson, editors, *Proc Human–Computer Interaction — INTERACT'99, IFIP TC.13*, 1999.

[155] John A. Bateman and Serge Sharoff. Multilingual grammars and multilingual lexicons for multilingual text generation. In *Multilinguality in the lexicon II*, ECAI'98 Workshop, pages 1–8, Brighton, UK, 1998. European Conference on Artificial Intelligence.

[156] Irina Kononenko and Serge Sharoff. Understanding short texts with integration of knowledge representation methods. In D. Bjorner, M. Broy, and I.V. Pottosin, editors, *Perspectives of System Informatics*, volume Vol. 1181 of *Springer Lecture Notes in Computer Science*, pages 111–121. Springer Verlag, 1996.

[157] Serge Sharoff and Lena Sokolova. Contrastive analysis of software manuals. In *Proc MULSAIC'96, workshop at ECAI'96*, pages 57–60, 1996.

[158] Serge Sharoff. Phenomenology and cognitive science. *The Stanford Humanities Review*, 4(2):190–206, 1995.

[159] Serge Sharoff and Lena Sokolova. Analysis of rhetorical structures in technical manuals and their multilingual generation. In *Proceedings of the Workshop on Multilingual Generation (IJCAI'95)*, pages 119–128, Montréal, 1995.

[160] Serge Sharoff. System for development of linguistic processors: Snoop. In *Proc East-West Conference on Artificial Intelligence*, pages 184–188, Moscow, September 1993.