

WAC 2026

The 13th Edition of the Web-as-Corpus Workshop

Proceedings of the Workshop

October 29, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-481-1

Introduction

The Web as Corpus (WaC) workshop series has a long history. The first WaC workshop took place at the Corpus Linguistics conference in Birmingham in 2005, at a time when treating the Web as a serious source of linguistic data was still far from standard practice. The attraction was obvious: the Web offered quantities and varieties of language data that traditional corpus construction could hardly match, particularly for less-resourced languages. At the same time, using it as a corpus meant confronting difficult questions of sampling, representativeness, language identification, cleaning, duplication, metadata, and the very meaning of a corpus when its source was something as large, dynamic, and heterogeneous as the Web.

WaC-13 marks a return of the workshop after a substantial hiatus. WaC-XI took place in 2017, while WaC-XII, planned for LREC 2020, had its physical meeting cancelled by the COVID-19 pandemic, although its proceedings were published. In the intervening years, the role of web data in language technology has changed dramatically. With the advent of large language models, web corpora have moved from being a sometimes unconventional source of language data to becoming a central component of the data infrastructure underlying modern NLP. Questions that occupied the WaC community from its beginnings now concern a much larger part of our field. What does a web crawl contain? What does it miss? How should it be filtered? Which languages, registers, communities, and kinds of content survive the successive stages from crawling to a released corpus? And what biases are introduced along the way?

The programme of WaC-13 reflects this new centrality of web data, while at the same time retaining the remarkable breadth that has always characterized Web-as-Corpus research. Several contributions examine the fundamental processing stages of contemporary web corpora: language identification, main-content extraction, quality filtering, register classification, and deduplication. Others ask what our pipelines systematically lose, with studies of endangered and low-resource languages, multilingual coverage, and the effects of filtering on geographical, sociolinguistic, thematic, and register diversity. Questions of representativeness span languages and regions from Sámi and Muskogean to Persian, South Slavic, Finnish, Sri Lankan, and cross-lingual resources.

At the same time, the programme shows how the object of study itself is changing. Contributions investigate AI-generated text on the Web, AI-mediated search, machine-translated web corpora, changing patterns in the questions found online, and the emerging infrastructure of crawler consent through `robots.txt`. Other papers extend the notion of Web as Corpus beyond conventional collections of HTML text, considering PDFs, web video and speech, online debates, Wiktionary, news, and even programming languages. Taken together, the papers remind us that there is no single “web corpus”: there are successive choices about what to crawl, extract, identify, retain, annotate, and ultimately make available for research and model training.

Given the changes the field has undergone in recent years, it is particularly fitting that our invited speaker is Sebastian Nagel of the Common Crawl Foundation. His talk, *The Web as a Corpus and Common Crawl: A Look Back, Forward and into Present Challenges*, connects the history of the field directly to its current infrastructure. Common Crawl occupies a special position in today’s web-data ecosystem. By maintaining an openly accessible, continuously updated archive of web crawl data, it provides infrastructure on which a substantial amount of academic and industrial research now depends. For researchers without the resources to operate a web-scale crawler themselves, such open infrastructure is especially important: access to the raw material of web-scale research should not be restricted to a small number of very large organizations. The prominence of Common Crawl in the WaC-13 programme itself, both as a source of data and as an object of critical investigation, also demonstrates that shared infrastructure is most valuable when a research community can study its strengths, limitations, coverage, and consequences openly.

Finally, WaC-13 would not have happened without SIGWaC, the ACL Special Interest Group on Web as Corpus. There is an appealing continuity here with the history of the workshop. Web-as-Corpus research grew to a large extent from the bottom up: researchers discovered that the Web offered extraordinary possibilities, developed methods and tools to make use of it, and formed a community around the

problems they encountered. SIGWaC, and the broader framework of the Association for Computational Linguistics, complement this bottom-up energy with something equally important: durable structures through which a research community can maintain contact, exchange knowledge, organize events, and renew itself across generations of researchers. Such structures should not determine where research goes: the most interesting developments will continue to emerge from researchers pursuing questions that were not anticipated by any committee or organization. That said, conferences, workshops, and SIGs give those developments places in which to meet. They help turn scattered individual efforts into conversations, and conversations into communities.

After a long break, we are particularly happy to bring the Web as Corpus community together again at WaC-13. We would like to thank all authors who submitted to the workshop, the reviewers who gave their time to provide valuable feedback on the submissions, and the EMNLP workshop chairs for their support. The technological context may have changed profoundly since 2005, but the central insight behind the workshop has proved remarkably durable: the Web is an extraordinary source of evidence about language, but turning the Web into data is never a neutral or trivial operation. In the age of large language models, understanding that process matters more than ever.

WaC-13 Organizing Committee

Organizing Committee

Program Chairs

Pedro Ortiz Suarez, Common Crawl Foundation, USA

Laurie Burchell, Common Crawl Foundation, USA

Nikola Ljubešić, Jožef Stefan Institute, Slovenia

Yves Scherrer, University of Oslo, Norway

Veronika Laippala, University of Turku, Finland

Vuk Dinić, Jožef Stefan Institute, Slovenia

Thom Vaughan, Common Crawl Foundation, USA

Program Committee

Program Chairs

Laurie Burchell, Common Crawl Foundation
Vuk Dinić, Jozef Stefan Institute and Jožef Stefan International Postgraduate School
Veronika Laippala, University of Turku
Nikola Ljubešić, Jožef Stefan Institute
Pedro Ortiz Suarez, Common Crawl Foundation
Yves Scherrer, University of Oslo
Thom Vaughan, Common Crawl Foundation

Reviewers

Mantaqa Abedin, Dominik Allen, Catherine Arnett

Pushkar Baranwal, Vladimír Benko, Arghya Bhattacharya, Magnus Breder Birkenes, Laurie Burchell, Albina Burlova

Hande Celikkanat, Aditya Channa, Antonis Charalampous, Pinzhen Chen, Austin Cheng, Oleg Chernikov, Steven Coats, Sophia Conrad

Ahrnish P Dahal, Constantinos Djouvas, Ona de Gibert

Valeria Efimova

Adriano Ferraresi, Marie Flesch, Luca Foppiano

Tomasz Galica, Susmita Gangopadhyay, Filip Ginter

Srinivas Harish, Jindřich Helcl, Erik Henriksson, Im Hyeonyeol

Maximilian Idahl

Sangwan Jeon, Robert Jäschke

Madhav Kapoor, Toprak Seda Karaosmanoğlu, Parul Khandelwal, Mateusz Klimaszewski, Jaroslav Kopcan, Olli Kuparinen, Andrey Kutuzov

Veronika Laippala, Yibo Li, Geng Liu, Nikola Ljubešić

Vukosi Marivate, Bruno Leonardo Santos Menezes

Kumarmanas Nethil

Stephan Oepen, Pedro Ortiz Suarez, Malte Ostendorff, Christian Oswald

Michael Paris, Charles Pert, Francesco Pierri, Paola Pietrandrea

Nishat Raihan, Tharindu Ranasinghe, Rishi Ravikumar, Georg Rehm, Jacob Rigal, Daniel Ruffinelli

Kengatharaiyer Sarveswaran, Nadav Shani, Serge Sharoff, Jastej Singh, Preety Singh, Rob Sniderman, Amr Sobhy, Egon Stemle, Vít Suchomel, Marek Šuppa

Sotaro Takeshita, Otto Tarkka, Lin Tian, Luisa Troncone

Sami Virpioja, Jelmer van der Linde

Jessica Wang, Sidney Wong

Marcos Zampieri, Gary Zhang, Ruiyang Zhang, Calvin Zhou

Keynote Talk

The Web as a Corpus and Common Crawl: A look back, forward and into present challenges

Sebastian Nagel
Common Crawl Foundation



2026-10-29 09:15:00 – Room: Room 1

Abstract: During the 1990s, the use of text corpora to study natural language became widespread. This shift away from intuition to data-driven methodologies based on representative collections of language in use became known as the “corpus revolution”. Corpora grew in size, reaching the landmark figure of 100 million words. Corpora also became digital during this decade.

During the same decade, the World Wide Web, invented in 1989, saw a swift breakthrough and began its exponential growth. It was therefore only natural that corpus linguists started to use the web as a resource. The “Web as a Corpus” emerged, and around 2005, the SIGWAC and WaCky initiatives started their work. Using web-harvested texts in the same way as established corpora came with many challenges, including selecting a suitable set of seed URLs, implementing and running a web crawler, extracting text from HTML, cleaning boilerplate content, detecting duplicates and near-duplicates, adding linguistic annotations and building user-friendly interfaces that enable linguists to explore the corpora.

Another initiative, Common Crawl, was founded in 2007. Its mission is to make web data more accessible to programmers and data scientists. Rather than operating a web crawler themselves, developers and researchers can focus on using web data to gain insights and build innovative products. Common Crawl published its first data set in 2008. Since then, more than 120 large-scale web crawls have been run, and the data has been used and cited in over 10,000 research papers. While a significant proportion of this research is about NLP and language modelling, Common Crawl’s scope extends to many other disciplines, including web science, information retrieval, the semantic web, internet security and digital forensics, and humanities and social sciences.

Around 2010, both the Web-as-Corpus initiative and Common Crawl saw a surge in interest and growing communities. A few years later, however, interest dropped. Although many tools for processing noisy web data were already available, scaling up and tuning these tools for a particular use case or language remained challenging. It was not until 2020, with the advent of the transformer era in language modelling coupled with the general availability of cloud computing, that it became possible to process terabytes and petabytes of web data. Since then, training language models has become the most important use case for web data. Nevertheless, many of the challenges described in the early Web-as-Corpus workshops remain unsolved and require continuous, collaborative research.

This talk will look back at the early days of corpus linguistics and the use of web data. It will discuss the differences and similarities between the Web-as-Corpus initiatives and Common Crawl, and provide an overview of recent challenges and future tasks.

Bio: Sebastian Nagel is a programmer and engineer at Common Crawl and the University of Konstanz. At Common Crawl, he runs and maintains the crawler and works to make web data more accessible. Prior to joining Common Crawl, Sebastian worked as a software developer at Exorbyte, where he implemented search and data quality solutions. Sebastian holds a PhD in computational linguistics from the University of Munich. As a longtime committer of Nutch and StormCrawler, and a member of the Apache Software Foundation, he strongly believes in open-source software development. He studied linguistics (Slavic languages) and cultural anthropology in Munich, Kazan and Prague. It was in Prague that he first attended a class in corpus linguistics, which sparked his interest in computational linguistics.

Table of Contents

<i>CHALIS: A Challenge Dataset for Language Identification in Difficult Scenarios</i> Michal Tichý and Jindřich Libovický	1
<i>The CreateDebate Arguments Dataset: A Large-Scale Naturalistic Corpus for Computational Argumentation</i> Antonis Charalampous and Constantinos Djouvas	8
<i>Who Published This Headline? Editorial Fingerprinting in French News Under Topic Control</i> Amr Sobhy	22
<i>A Taxonomy of Programming Languages for Code Generation</i> Nishat Raihan, Christian Newman and Marcos Zampieri	37
<i>What Persists across Web Crawls? Genre and News Topic Variation in South Slavic Web Corpora</i> Nikola Ljubešić, Vuk Dinić and Peter Rupnik	53
<i>HPLT3-PerRef: A Refined Large-Scale Persian Web Corpus via Supervised Document Quality Filtering</i> Alireza Razzaghi, Tomasz Galica, Erik Henriksson and Veronika Laippala	60
<i>Host-Aware Web Crawling for LLM Pretraining Data</i> Malte Ostendorff	70
<i>Auditing Source Exposure in Baidu and Google AI Search</i> Yibo Li, Enci Guan, Yuedan Cai, Geng Liu and Francesco Pierri	88
<i>Lost in Translation: Structured-Span Corruption in Machine-Translated Multilingual Web Corpora</i> Ruiyang Zhang	100
<i>Counting Documents Is Not Counting Text: Unit Bias in Web-PDF Corpus Statistics</i> Luca Foppiano	107
<i>Is Educational Content Enough? The Linguistic and Thematic Effects of Educational Filtering</i> Urho Kunnas, Veronika Laippala, Sampo Pyysalo and Tomasz Galica	115
<i>WiNCC: A Reproducible Pipeline for Constructing a Wikipedia-enriched Nepali Cross-lingual Corpus</i> Ahrnish P Dahal, Pushkar Baranwal, Amit Pundir, Geetika Jain Saxena and Sanjeev Singh ..	131
<i>Mapping Sámi Languages in a Norwegian National Web Crawl: Results from a Representativeness Study</i> Magnus Breder Birkenes	139
<i>Towards More Reliable Web Metadata: Do Register Classifiers Learn Register or Genre Features?</i> Antti Kanner, Tuomas Lundberg, Erik Henriksson and Veronika Laippala	147
<i>Same Pages, Cleaner Labels: Language Identification for Eight Low-Resource Languages in mC4 and FineWeb-2</i> Gary Zhang	161
<i>How Many Duplicates Are Left? Capture-Recapture and Sampling Inversion for Residual Duplication in FineWeb</i> Gary Zhang	170
<i>Using Search Engines to Trace Metalinguistic Discussions of the Noun “Female” in Digital Discourse</i> Marie Flesch	182

<i>Keeping the Ums: Conversation-Analysis-Informed Verbatim ASR for Web Speech</i>	
Steven Coats	192
<i>Segment-Level Register Analysis of the Multilingual Web</i>	
Erik Henriksson, Alireza Razzaghi, Tomasz Galica, Jiaqi Feng Guo, Anastasia Zhi-Ling Dong, Urho Kunnas, Erofilis Psaltaki, Tuomas Lundberg, Antti Kanner and Veronika Laippala	200
<i>Invisible on the Web: An Audit of Low-Resource Muskogean Languages in Web-Crawled Corpora</i>	
Sharvin Goyal	214
<i>The Crawl Gap: Where Endangered-Language Web Content Is Lost Before It Reaches a Corpus</i>	
Sharvin Goyal	220
<i>You Can Tell Who's Asking: What the Web's Questions Are Made Of, and Where They Come From</i>	
Calvin Zhou, Vincent McCloskey and Krishna Srinivasan	224
<i>Wiktionary as a Crowdsourced Lexicon for English Dialects</i>	
Sidney Wong	238
<i>Task-Agnostic Web Document Annotation with LLM-Generated Descriptors</i>	
Otto Tarkka, Erik Henriksson, Jenna Kanerva and Filip Ginter	252
<i>PAWN++ in the Wild: Evaluating AI-Generated Text Detection on the Web</i>	
Oleg Chernikov, Evgenii Ptitsyn and Albina Burlova	282
<i>Extraction Changes the Route: Main-Content Extraction and Language Identification in Web Corpora</i>	
Toprak Seda Karaosmanoğlu	297
<i>CeyNews: A Sri Lankan Trilingual News Corpus</i>	
Ajitha Sivakulasingam, Kengatharaiyer Sarveswaran and Tharindu Ranasinghe	308
<i>Probing Pretraining "Quality": Sociolinguistic Mobility as a Lens on Web Corpus Filters</i>	
Jessica Wang	320
<i>A Web-Derived Typology of Multimodal Infant Crying: Taxonomy, a Text-Only Web-Corpus Comparison, and Ethical Issues</i>	
Jacob Rigal and Jonathan Ginzburg	331
<i>Whose Robots.txt? Consent-Infrastructure Inequality Across the Languages of Web Corpora</i>	
Gary Zhang and Austin Cheng	347