浙江大学 ZHEJIANG UNIVERSITY

THE UNIVERSITY OF CHICAGO

“语言演化与计算”国际工作坊

International Workshop on Language Evolution and Computation

2026年9月22-23日

浙江大学紫金港校区 求是大讲堂 潘方仁求是馆201室

22–23 September 2026

Zijingang Campus, Zhejiang University, Hangzhou, China

主办 / Host

浙江大学外国语学院丨美国芝加哥大学语言学系

School of International Studies, Zhejiang University丨Department of Linguistics, The University of Chicago

会议语言：英语 / Working Language: English

About the Workshop

值此美国艺术与科学院院士、芝加哥大学 John Matthews Manly 杰出讲席教授、古根海姆学者、美国语言学会前秘书兼司库 Lenore A. Grenoble 教授携团来访之际，浙江大学外国语学院与美国芝加哥大学语言学系联合举办“语言演化与计算”国际工作坊。

工作坊汇集来自浙江大学、芝加哥大学、中央民族大学、M. K. 阿莫索夫东北联邦大学以及首尔国立大学的学者，围绕语言演化这一主题，分别从计算、计量、类型学与语言记录等角度展开讨论，并以圆桌交流收束，商议后续学术合作之可能。

The workshop is held on the occasion of Professor Grenoble's visit to Zhejiang University with her colleagues from Chicago. It brings together scholars from Zhejiang University, the University of Chicago, Minzu University of China, M. K. Ammosov North-Eastern Federal University and Seoul National University, and takes up questions of language evolution from computational, quantitative, typological, and documentary points of view. A roundtable on the final morning considers what forms of international collaboration might follow.

Program

Day 1 — Tuesday, 22 September	
Time	**Session**
10:00–10:05	Opening — Li Wenchao
10:05–10:20	Group Photo
10:20–11:00	Keynote I — Prof. Lenore A. Grenoble（The University of Chicago） （40 min）
11:00–11:15	Break
11:15–11:55	Keynote II — Li Wenchao（Zhejiang University） （40 min）
12:00–14:30	Lunch
14:30–15:00	Talk — Yan Jianwei（Zhejiang University）
15:00–15:30	Talk — Jaehong Shim & Sangah Lee（The University of Chicago & Seoul National University）
15:30–15:45	Break
15:45–16:15	Talk — Aitalina Timofeeva（M. K. Ammosov North-Eastern Federal University）
16:15–16:45	Talk — Stepan Pavlov（M. K. Ammosov North-Eastern Federal University）
16:45–18:00	Free Time
18:00–	Dinner

Day 2 — Wednesday, 23 September	
Time	**Session**
09:00–09:40	Keynote III — Prof. Ding Shiqing（Minzu University of China） （40 min）
09:40–10:10	Talk — Vitalii Ivanov（M. K. Ammosov North-Eastern Federal University）
10:10–10:40	Talk — Ekaterina Podorozhnaia（M. K. Ammosov North-Eastern Federal University）
10:40–11:00	Break
11:00–12:00	Roundtable Discussion: Prospects for Collaboration — Li Wenchao（60 min）
12:10–14:00	Lunch

Keynote Speakers

Keynote I

Lenore A. Grenoble | The University of Chicago

美国艺术与科学院院士　Member, American Academy of Arts and Sciences

芝加哥大学 John Matthews Manly 杰出讲席教授　John Matthews Manly Distinguished Service Professor, The University of Chicago

古根海姆学者　Guggenheim Fellow

美国语言学学会前秘书兼司库　Former Secretary-Treasurer, Linguistic Society of America

Professor Grenoble is a world-leading scholar in language documentation, endangered languages, and language contact. Her research covers Slavic linguistics, Arctic and Siberian languages, and the theory and practice of language revitalization. She has participated in several major international collaborations on the documentation of endangered languages.

Keynote II

Li Wenchao

浙江大学外国语学院教授、博士生导师

Professor and Doctoral Supervisor, School of International Studies, Zhejiang University

李文超的研究领域为计量语言学与计算语言学，主要关注人类的语言如何构造、如何演化，以及计算方法能否帮助那些正走向沉寂的濒危语言。具体聚焦语言结构的类型与演化，以及濒危民族语言的可计算资源建设，并探索深度学习在语言分析和濒危语言保护中的应用。

Li Wenchao works at the meeting point of quantitative and computational linguistics, asking how the world's languages are structured, how they change over time, and how computational methods can help document and sustain those now at risk of falling silent. Her research moves across the typology and evolution of linguistic structure and the building of computable resources for endangered minority languages—increasingly drawing on deep learning both to analyse language and to keep it alive.

Keynote Speakers

Keynote III

丁石庆 教授 Prof. Ding Shiqing

中央民族大学语言文学院 教授、博士生导师

Professor and Doctoral Supervisor, School of Chinese Ethnic Minority Languages and Literatures, Minzu University of China

丁石庆教授长期致力于中国少数民族语言、阿尔泰语系语言的记录、保护与研究，在阿尔泰语系研究领域享有重要学术影响。现任国家语委语言文字规范标准审定委员会委员、中国语言资源保护工程核心专家组成员、中国民族语言学会副会长、中国少数民族语言资源保护研究中心常务副主任、中国阿尔泰学研究中心副主任、江苏师范大学语言科学与艺术学院特聘教授。

Ding Shiqing is a leading scholar in the documentation, preservation, and study of China's ethnic minority languages and the Altaic language family. He serves as a member of the Language Standards Review Committee of the State Language Commission, a core expert of the Program for the Protection of China's Language Resources, Vice President of the Chinese Association of Ethnic Languages, Executive Deputy Director of the Research Center for the Protection of Chinese Ethnic Minority Language Resources, and Deputy Director of the Chinese Center for Altaic Studies, as well as Distinguished Professor at the School of Language Sciences and Arts, Jiangsu Normal University.

团队核心成员 / Core Team Members

阎建玮 Yan Jianwei

浙江大学“百人计划”研究员，博士

ZJU Hundred Talents Program Research Professor, PhD

阎建玮，研究方向为计量语言学与语言类型学，主持国家社科基金青年项目，围绕形态复杂度、语序类型与依存方向已有系列成果，发表于《外语教学与研究》《语言文字应用》及 Zeitschrift für romanische Philologie、Linguistics Vanguard 等国内外刊物。

Yan Jianwei works in quantitative linguistics and linguistic typology. As principal investigator of a Youth Project of the National Social Science Fund of China, he has produced a body of work on morphological complexity, word-order typology, and dependency direction, published in journals including Foreign Language Teaching and Research (Waiyu Jiaoxue yu Yanjiu), Applied Linguistics (Yuyan Wenzi Yingyong), Zeitschrift für romanische Philologie, and Linguistics Vanguard.

Participating Scholars

Jaehong Shim

PhD Student, Department of Linguistics, The University of Chicago
博士研究生，美国芝加哥大学语言学系

Ekaterina Podorozhnaia

Research Assistant, Laboratory of Computational Technologies and Artificial Intelligence, M. K. Ammosov North-Eastern Federal University
研究助理，计算技术与人工智能实验室，M. K. 阿莫索夫东北联邦大学

Aitalina Timofeeva

Junior Researcher, Arctic Linguistic Ecology Lab, M. K. Ammosov North-Eastern Federal University
初级研究员，北极语言生态实验室，M. K. 阿莫索夫东北联邦大学

Stepan Pavlov

Junior Researcher, Arctic Linguistic Ecology Lab, M. K. Ammosov North-Eastern Federal University
初级研究员，北极语言生态实验室，M. K. 阿莫索夫东北联邦大学

Vitalii Ivanov

Junior Researcher, Arctic Linguistic Ecology Lab, M. K. Ammosov North-Eastern Federal University
初级研究员，北极语言生态实验室，M. K. 阿莫索夫东北联邦大学

Abstracts

Keynote I — Lenore A. Grenoble

Title: The Arctic Linguistic Ecology Lab: Research and Perspectives

Abstract:

The Arctic Linguistic Ecology Lab was founded in 2021 at the M. K. Ammosov North-Eastern Federal University in Yakutsk under the auspices of a grant to study the Preservation of Linguistic and Cultural Diversity and Sustainable Development of the Arctic and Subarctic of the Russian Federation. The laboratory studies the vitality of the official languages of the Republic of Sakha (Yakutia): Sakha, Dolgan (Turkic); Even and Evenki (Tungusic); Chukchi (Chukotko-Kamchatkan); and Forest and Tundra Yukaghir (Yukaghiric). Since its inception, the lab has been dedicated to the collection and analysis of data from speakers across the territory of Yakutia and in neighboring Chukotka. In this talk I introduce the work of the Arctic Linguistic Ecology Lab, our over-arching goals and findings to date.

Our foundational work involved:

1. The development of methods for investigating language contact and shift through experimentally-oriented documentation. We developed, and continue to develop, short experiments that can be run in field conditions and can be replicated across different language communities to obtain comparable data on the linguistic change in process, capturing shift as it is ongoing. A key goal is to understand if the changes across the typologically and genealogically diverse languages of Yakutia are similar or different as speakers shift to Russian, and to determine whether it is possible to identify stages in shift.

2. The first stage of the project involved extensive documentation of the languages of Yakutia, with the collection of sociolinguistic interviews (conducted in the target language to provide both socio- and linguistic data), narratives and targeted elicitations, through traditional and new experimental methods.

3. The team in Chicago has been actively engaged in the development of tools for conducting research, including the use of Whisper for ASR in Sakha. Since 2023, the Chicago team has been actively engaged in the processing and analysis of Dolgan and Even recordings, preparing them for archiving according to international (DELAMAN) standards, and in early preparations for the Evenki language corpus.

Building on this foundation, our current focus brings together an interdisciplinary team to study changes in the Arctic and northern communities of the Republic of Sakha (Yakutia) and the Chukotka Autonomous Okrug, as well as the evolution of the linguistic and cultural landscape over time and space.

We pursue several interconnected goals:

1. to study linguistic change and variation from a historical perspective using the methods of historical

linguistics, ethnography, archival and legacy materials, and genetics;

2. to understand transformations in the population structure, in population movements, in language, and in culture;

3. to analyze contemporary change and variation using large-scale data collection methods, digital technologies and statistical modeling, as well as the latest innovations in theoretical and methodological approaches to study the causes and consequences of these changes for society; and

4. to apply mathematical modeling to predict future changes.

Keynote II —Li Wenchao

Title: From Grammars to Treebanks to Genealogy: The Deep History of the Transeurasian Languages

Abstract:

Most of the world's endangered languages have no digital corpus, which places them beyond the reach of the computational methods now standard in linguistics. This talk concerns a group of such languages — the Tungusic and Mongolic families — and uses them to re-examine a long-standing question from a new angle: the Transeurasian hypothesis, which holds that Turkic, Mongolic, Tungusic, Japonic, and Koreanic descend from a common ancestor some 9,000 years ago. Debated for over a century and argued almost entirely on lexical grounds, the hypothesis is approached here from the structural side.

The first challenge is one of data. Most of these languages have no running text, and thus lie outside the scope of standard treebank methods. I describe how we constructed Universal Dependencies treebanks for fifteen of them from the example sentences of published reference grammars, under a shared annotation scheme that keeps the languages comparable. Because the languages are annotated alike, the treebanks also reinforce one another — evidence from the family as a whole can help build and parse the next language in it. These appear to be the first computable treebanks for most of the languages concerned, and are being prepared for release through the Universal Dependencies repository.

Drawing on these resources, I test whether structural features can recover the family tree, using a Bayesian framework first calibrated on Indo-European, where the history is independently known — a benchmark that shows the method finds structure where structure exists. At Transeurasian time depths, it does not. The features distinguish head-final from head-initial languages clearly enough, but resolve nothing below that level: none of the major branches is recovered as a group. The clearest indication is Hindi — Indo-European, and unrelated to any of these languages, yet SOV — which falls squarely within the Transeurasian cluster. What the features track, in other words, is word-order type rather than common descent.

This is best read as a productive negative result. The study contributes new treebanks for languages that previously had none, and a clearer sense of where structural data cease to be informative about the deep past.

Abstracts

Keynote III — Ding Shiqing

Title: The Multidimensional Value of Low-Resource Languages in Northern China and a Differentiated, Precision-Based Approach to Their Protection

Abstract:

Northern China is home to the three major branches of the Altaic language family and several Indo-European languages, together with a number of mixed languages, forming a low-resource language community marked by typological diversity and pronounced stratification in its patterns of evolution. Taking the full range of low-resource languages across northern China as its object of study, this research draws on quantitative data concerning linguistic structure, current usage, and vitality to construct a tripartite framework of driving forces—natural evolution, ecological evolution, and superimposed evolution—and to clarify how three dimensions of ecological variables—social contact, productive culture, and community transmission—differentially shape language structure. On this basis, the study systematically develops a five-dimensional value system for the low-resource languages of northern China, distinguishes the hierarchical differences in value and the risks to survival among languages of different evolutionary types, and examines the practical difficulties facing current language-protection efforts. Grounded in these differences in value and divergences in evolution, it constructs a stratified, categorized, and precisely adapted system and practical pathway for differentiated protection, thereby providing both theoretical support and practical reference for the dynamic protection, living transmission, and sustainable use of low-resource languages in northern China.

Abstracts

Talk —Yan Jianwei

Title: Diachronic Changes in Morphological Richness and Word Order Freedom from Latin to Romance: A Treebank-Based Quantitative Study

Abstract:

The evolution from Latin to the Romance languages provides a particularly revealing case for investigating how morphological and syntactic systems change over time. This study examines the diachronic relationship between morphological richness and word order freedom from Latin to Romance using large-scale morphologically and syntactically annotated corpora from Universal Dependencies (UD). The analysis covers Latin and seven Romance languages, Spanish, Portuguese, Galician, Catalan, French, Italian, and Romanian, represented by 23 treebanks. Four Chinese treebanks, including Classical and Modern Chinese, are additionally used as reference points for a highly analytic language type. Morphological richness is quantified using the moving-average mean size of paradigm, which captures the average number of word forms associated with lemmas, while word order freedom is measured using cosine similarity of constituent-order distributions across main and subordinate clauses. The results reveal a clear diachronic shift from the highly inflectional system of Latin toward morphologically less rich Romance languages. This morphological reduction is accompanied by increasing regularization of constituent order, with Romance languages showing a stronger tendency toward SVO ordering than Latin. More importantly, the cross-corpus comparison identifies a significant association between morphological richness and word order flexibility: languages and historical stages with richer morphology tend to permit greater variation in constituent order, whereas morphological simplification is associated with increasing syntactic rigidity. These findings provide quantitative evidence for a diachronic trade-off between morphological complexity and word order freedom in the transition from synthetic Latin to more analytic Romance systems. By integrating diachronic corpora, dependency annotation, and quantitative typological measures, this study illustrates how computational approaches can help uncover systematic patterns of structural change and contribute to a more general understanding of language evolution.

Abstracts

Talk — Jaehong Shim & Sangah Lee[1]

Title: From Middle Korean to Altaic-type Languages: A BiLSTM-based Toolkit for Low-Resource Morphological Analysis

Abstract:

Computational processing of morphologically rich, low-resource languages faces two fundamental constraints: the absence of large-scale training corpora, and the difficulty of leveraging contextual information to resolve morphological syncretism. This study proposes a morphological analysis model-building and search tool designed to operate under these constraints.

The tool adopts a hybrid architecture that trains a BiLSTM (Bidirectional Long Short-Term Memory) model on gold annotations constructed directly by the researcher, combined with rule-based candidate generation. This design allows the tool to produce meaningful output even at early stages when training data is extremely limited, while performance improves incrementally as annotation progresses. The BiLSTM's structural integration of forward and backward context provides a principled basis for addressing morphological syncretism without relying on large-scale pretrained language models—a choice motivated by the observation that, given the limited data available for the target languages, the advantages of Transformer-based large language models are correspondingly limited.

The tool is currently being developed with a primary focus on languages typologically similar to Korean—agglutinative, head-final languages—with Middle Korean serving as the first case study for validation. While the pipeline itself has reached a functional stage, refinement of its internal algorithms and construction of training data are proceeding in parallel. This process has surfaced not only the challenge of data scarcity, but also a more fundamental difficulty in annotation: present-day researchers cannot fully substitute for the intuitions of historical native speakers.

Building on this development experience, the presentation discusses the potential for extending the tool to other Altaic-type head-final languages, including those of the Manchu-Tungusic, Turkic, and Mongolic families. It further outlines plans to support Universal Dependencies format export to ensure interoperability with existing language resources and articulates a longer-term goal of extending the tool's applicability beyond typologically similar languages to low-resource languages more broadly.

[1] Professor, Department of Linguistics, Seoul National University.

Abstracts

Talk — Aitalina Timofeeva

Title: Digital Language Tools as a Critical Asset for the Indigenous Languages of Yakutia

Abstract:

This paper addresses the precarious state of six indigenous languages of Yakutia – Sakha (Yakut), Dolgan, Even, Evenki, Yukaghir, and Chukchi. With the exception of Sakha (vulnerable), all are endangered or severely endangered according to UNESCO. In the digital age, a language absent from digital communication risks being excluded from modern life and disappearing faster. Therefore, digital tools are not merely technological aids but a prerequisite for cultural survival.

The current level of digitalization is highly uneven. Sakha has the most developed digital ecosystem: the National Corpus of the Sakha Language (over 15-million-word usages), generative AI (Gemini, GigaChat, Telegram bots), “Sakhalyy” language learning app, TTS/STT projects (“Saŋa-Suruk”, “Istiŋ”), virtual assistants (“Ayta”, “AiUol”) etc. The other five languages have more limited resources. Key actors include the National Library of Yakutia, Ammosov North-Eastern Federal University, scientific institutes, international research projects, grant-funded projects, enthusiasts and technology companies such as Yandex and Sberbank. State support has grown, but remains insufficient for large-scale digitalization.

Technologically and methodologically, data quality is problematic: corpora are often compiled and annotated by non-speakers, lack domain representativeness, and are not adapted for everyday or educational use. For example, annotating Sakha morphology is challenging due to a lack of standards and universal terminology, polysemy and homonymy, and compatibility issues with annotated corpora of other languages. Under-resourced languages lack data due to their small number of speakers and/or their geographical dispersion and the labor-intensive nature of corpus creation.

Abstracts

Talk — Stepan Pavlov

Title: Annotating Spoken Language in ELAN: Practical Experience

Abstract:

This paper presents practical experience in annotating spoken Sakha (Yakut) language data in ELAN. The material includes two types of recordings: interviews and a linguistic experiment. In the experiment, participants watched a short silent cartoon and then retold it in their native language. There were no special rules for the retelling. The interviews usually last about thirty minutes, while the cartoon retellings are usually three to five minutes long.

The annotation in ELAN includes several tiers. The first tier contains a detailed transcription of the original speech. The transcription keeps the main features of real spoken language, such as pauses, hesitations, reduced forms, and other sounds that can be heard in the recording. The second tier contains a Russian translation. The third tier gives a standard literary version of the Sakha text without reductions or non-standard forms. The material is also divided into separate lexical units. Each lexical unit can have its own Russian translation. Another tier is planned for morphological glossing, but glossing has not yet been completed. A separate tier is used for researchers' comments.

The practical work includes audio segmentation, transcription, translation, and the creation of annotation tiers. Several difficulties can appear during this work. Some parts of the recordings are difficult to understand because of unclear pronunciation. Dialectal forms are especially important because they may differ from standard literary Sakha. Code-switching between Sakha and Russian is another common challenge.

The experience shows that ELAN is useful for working with spoken Sakha data. It allows researchers to connect the audio recording with several levels of linguistic annotation. The annotated material can also be used for further linguistic analysis and for preparing data for a future language corpus.

Abstracts

Talk — Vitalii Ivanov

Title: ArcticLing-SVFU: A Digital Platform for Documenting and Studying Arctic Languages

Abstract:

Several languages of the indigenous peoples of the North and the Arctic are currently in a highly vulnerable state for a variety of reasons. Language documentation serves as a critical approach to preserving and supporting these languages. However, a major obstacle to their documentation and preservation is the acute shortage of structured digital tools, resources, and corpora. To address this issue, the M. K. Ammosov North-Eastern Federal University is developing the digital analytical and information platform ArcticLing-SVFU, which serves as a comprehensive scientific and educational infrastructure. This presentation provides an overview of the platform's architecture and functional capabilities, and outlines the primary directions for its future development.

The platform aggregates linguistic data, including literary texts, field audio recordings, transcriptions, and dictionaries, providing researchers with analytical tools for frequency analysis, concordance generation, and the geographical visualization of language distribution.

Looking ahead, the platform's development trajectory is directed toward integrating artificial intelligence. Future developments will focus on applying machine learning (ML) and natural language processing (NLP) to the low-resource languages of the Arctic, particularly through the implementation of automatic speech recognition (ASR) and machine translation (MT) systems. Ultimately, the ArcticLing-SVFU platform aims to consolidate modern digital tools to facilitate the documentation, preservation, and development of the indigenous languages of the North and the Arctic.

Abstracts

Talk — Ekaterina Podorozhnaia

Title: NLP for the Sakha Language

Abstract:

The present study addresses the development of speech and natural language processing technologies for Sakha, a low-resource Turkic language spoken primarily in the Republic of Sakha (Yakutia). Despite its active use, Sakha remains underrepresented in modern digital technologies due to the limited availability of annotated speech corpora and language resources. Its agglutinative morphology, vowel harmony, and language-specific phonemes and graphemes create additional challenges for automatic language processing.

Research conducted at the Laboratory of Computational Technologies and Artificial Intelligence at M. K. Ammosov North-Eastern Federal University focuses on adapting modern neural architectures to Sakha. Current work includes automatic speech recognition (STT/ASR) and text-to-speech synthesis (TTS). For speech recognition, multilingual pretrained models such as Whisper-large-v3 and Wav2Vec2-BERT were adapted using Sakha speech data from Common Voice and institutional recordings. Full fine-tuning of Whisper-large-v3 achieved a word error rate of 8%, while Wav2Vec2-BERT combined with an n-gram language model achieved 13% WER.

For speech synthesis, XTTS-v2 was adapted using approximately 90 minutes of validated single-speaker Sakha speech, including tokenizer modification for Sakha-specific graphemes. The resulting system was implemented as a Telegram bot, demonstrating the practical applicability of speech technologies for a low-resource language. These results show that transfer learning can substantially reduce the data barrier for Sakha. Further development requires larger and higher-quality corpora, improved language modeling, morphological analysis, and closer integration of speech technologies with broader NLP tools.

Roundtable Discussion: Prospects for Collaboration

Time: 23 September 2026, 11:00–12:00

Moderator: Li Wenchao (Zhejiang University)

Topics:

The sharing of corpora and methods, and the training of young scholars, exploring possible directions for future academic collaboration.

致谢 / Acknowledgements

本次工作坊由浙江大学外国语学院与美国芝加哥大学语言学系联合主办，并得到中央民族大学中国少数民族语言资源保护研究中心的支持。谨向所有参会学者及支持单位致以诚挚谢意。

The workshop is jointly hosted by the School of International Studies at Zhejiang University and the Department of Linguistics at The University of Chicago, with the support of the Research Center for the Protection of Chinese Ethnic Minority Language Resources at Minzu University of China. We extend our sincere thanks to all participating scholars and supporting institutions.