# Studies in East Asian Comparative Linguistics

von

**Dr. Andreas Hölzl**

Kumulative Habilitationsschrift

im Fach

Allgemeine Vergleichende Sprachwissenschaft

Eingereicht an der Humanwissenschaftlichen Fakultät

der Universität Potsdam

zur Erlangung des akademischen Grades

Doctor habilitatus (Dr. habil.)

Department Linguistik
Humanwissenschaftliche Fakultät
Universität Potsdam
originally submitted in 2025
minimally revised after acceptance in 2026

# Contents

# Preface

This habilitation thesis contains a representative selection of my work on the languages of East Asia over the last decade, excluding work published in the context of my doctoral dissertation (Hölzl 2016). Also excluded were studies where I am not the first author, edited volumes, studies unrelated to East Asia, reviews, lemmata in dictionaries, unpublished manuscripts, studies that are currently under review. Unlike my dissertation (defended in 2017), which focused on a typology of questions in Northeast Asia (Hölzl 2018a), this habilitation has a much broader focus that also includes aspects of historical linguistics and addresses various features of the languages of East Asia, including Southeast Asia and especially Southwest China. Two of the altogether 14 studies address aspects of interrogativity similar to my dissertation but are independent publications that go much beyond my previous work and have been published many years after the completion of the PhD (Hölzl 2022a; Hölzl 2024p). One of the studies was written and published during the work on my dissertation but addresses a different topic (Hölzl 2015).

This habilitation thesis is dedicated to the memory of **Prof. Wolfgang Schulze (1953–2020)**, who not only first introduced me to linguistic typology and contact linguistics, but also supervised my doctoral dissertation from 2013 to 2017, gave me an opportunity to teach, helped me obtain a scholarship from the German Academic Scholarship Foundation (*Studienstiftung des deutschen Volkes*), and supported me in finding my way into the field of Tungusic studies (Schulze & Hölzl 2013). One of the studies in this habilitation — Hölzl (2018c) — was published as a contribution to the festschrift that I dedicated to him together with PD Peter-Arnold Mumm two years before Wolfgang sadly passed away (Hölzl & Mumm 2018). This habilitation is also dedicated to **Prof. Gisbert Fanselow (1959–2022)**, who offered me my current postdoc position (funded by the German Research Foundation, DFG) and with whom I originally planned to complete my habilitation before he became ill and sadly also passed away. Recently, a memorial volume was edited by Alexiadou et al. (2024). Since I am not the first author of my contribution, it is not included in this habilitation (Adam & Hölzl 2024).

I would like to thank Prof. Martin Salzmann for his constant help and support during the completion of this habilitation and in continuing Gisbert's project in Potsdam. Many thanks to Prof. Doreen Georgi and Prof. Malte Zimmermann for comments on an earlier version of this habilitation. Some of the studies included here were written or prepared when I was employed as a research assistant and lecturer at the Institute of Sinology in Munich and as a postdoctoral researcher at the former Institute of Comparative Language Science in Zurich (funded by the Swiss National Science Foundation, SNF). I would like to thank Prof. Hans van Ess, Prof. Balthasar Bickel, and Prof. Paul Widmer for offering me these positions.

I also would like to thank Prof. Hilary Chappell for discussing Sinitic languages and for a research visit to Paris to present my work on Longjia. Many thanks to Dr. Veronika Zikmundová for many discussions on Manchuric languages and several kind invitations to present my research in Prague.

Thanks to Prof. Ilja Seržant, Dr. Sergey Say, Dr. Cem Keskin, Dr. Christian Döhler, Dr. Andreas Pregla, Dr. Sebastian Nordhoff, Dr. Benedikt Peschl, Dr. Gao Xinyi, Dr. Raúl Bendezú Araujo, and Wakweya Olani Gobena for interesting discussions on linguistic typology and historical linguistics. Thanks to Jeanne Lecavelier des Etangs-Levallois for checking a French abstract. Many thanks to all of my language consultants who provided data used in the individual studies as well as in the introduction to this habilitation, including Saffanah Fathin (Indonesian), Lahari Chatterjee (Bengali), and Prof. Dubi Nanda Dhakal (Nepali).

I would like to thank several family members of Xie Yongxiu 谢永秀 and Zhao Defu 赵德富, two of the last speakers of Longjia, including Zhao Weihong 赵伟虹, Zhao Shan 赵珊, Zhao Chaomin 赵朝敏, Zhao Jingyi 赵静宜, and Zhao Chaoqiong 赵朝琼 for their help during fieldwork and for providing information and materials.

Many thanks to Prof. Guido Seiler and Prof. Martin Haspelmath for allowing me to join the editorial team of an edited volume (Enke et al. 2024), so that my study could be published in time to be included here (Hölzl 2024p). I am indebted to Dr. Thomas E. Payne for coauthoring one of the studies in this volume (Hölzl & Payne 2022a) and for coediting the volume it appeared in (Hölzl & Payne 2022b). Many thanks also go to all of the anonymous reviewers, proofreaders, editors, and other people involved in the publishing process of the studies assembled here.

My deepest gratitude to my *lǎoshī* Erika Schödel for first introducing me to the Chinese language while I was still attending high school and for making all of this possible.

I want to thank my wife Yadi Hölzl, also a coauthor of one of the studies (Hölzl & Hölzl 2025), for her unwavering support in daily life and research, for her help with questions concerning Chinese, and for her invaluable help with finding and discussing data on the Ta-Li languages. Finally, many thanks to her parents Liu Mingqiao 刘明巧 and Wu Jun 吴军 for their support and their help in conducting fieldwork in China (2014 on Sibe, 2019 on Lujia, 2025 on Longjia).

This habilitation was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 317633480-SFB 1287.

The habilitation process was concluded with my talk entitled "Eynu: The history of a secret language in China" and the following colloquium (15.04.2026). Many thanks to the habiliation committee, Prof. Isabell Wartenburger, and the three reviewers Prof. Martin Haspelmath, Prof. Wolfgang Behr, and Prof. Alena Witzlack-Makarevich.

# 1 Overview of this habilitation

## 1.1 Outline

This cumulative habilitation thesis is primarily concerned with Comparative Linguistics of East Asian languages. It includes a selection of **14 studies**, 12 of which I wrote as sole author and two as first author. These studies include 10 journal papers and four book chapters (Table 1). Together, the selection constitutes a comprehensive and representative overview of my work on the languages of East Asia. All studies are written in **English**. In addition to English abstracts, one paper contains a French abstract and two papers an abstract in Chinese. Taken together, the studies comprise approximately **463** pages, including bibliographies, appendices, photos, etc. Based on their overall topic, the studies are grouped into **four parts**:

- **part I** (1-4): comparative linguistics and language typology (see Section 2.3)
- **part II** (5-7): comparative linguistics and language contact (see Section 2.4)
- **part III** (8-12): comparative linguistics of Tungusic languages (see Section 2.5)
- **part IV** (13-14): comparative linguistics of the Ta-Li languages (see Section 2.6)

Together, the four parts include typological (**part I**), areal (**part II**), and genealogical aspects (**parts III**, **IV**), which allows for a comprehensive understanding of the linguistic diversity of East Asia. What all studies have in common is a panchronic comparative perspective as well as a focus on languages found in East Asia (EA). Especially, the typological studies in **part I** also include comparisons with languages from other parts of the world. The last **part IV** on the Ta-Li languages in Southwest China is the shortest because some results of my research in this area will be published separately in a forthcoming monograph on the Longjia language.

| Part | No. | Paper | Published | Open access | Type | Length |
|---|---|---|---|---|---|---|
| I | 1 | Hölzl 2015 | + | - | journal paper | 41 |
| | 2 | Hölzl 2025a | + | + | journal paper | 29 |
| | 3 | Hölzl 2024p | + | + | book chapter | 31 |
| | 4 | Hölzl 2022a | + | + | book chapter | 60 |
| | | | | | | **161** |
| II | 5 | Hölzl 2024j | online first | - | journal paper | 45 |
| | 6 | Hölzl 2023a | + | - | journal paper | 27 |
| | 7 | Hölzl 2018b | + | - | book chapter | 36 |
| | | | | | | **108** |
| III | 8 | Hölzl & Payne 2022a | + | + | book chapter | 20 |
| | 9 | Hölzl 2022b | + | - | journal paper | 49 |
| | 10 | Hölzl 2020a | + | + | journal paper | 9 |
| | 11 | Hölzl 2021e | + | + | journal paper | 19 |
| | 12 | Hölzl 2018c | + | - | journal paper | 36 |
| | | | | | | **133** |
| IV | 13 | Hölzl 2021a | + | + | journal paper | 22 |
| | 14 | Hölzl & Hölzl 2025 | online first | - | journal paper | 39 |
| | | | | | | **61** |
| Total | | 14 studies | | | | **463** |

Table 1: Overview of the studies included in this habilitation thesis

East Asia as understood here includes languages from at least 18 language families (Section 2.2), many of which are addressed in the individual studies.[1] Naturally, some families or subfamilies figure more prominently than others. For instance, Mongolic (a branch of Khitano-Mongolic) and Sinitic (a branch of Trans-Himalayan/Sino-Tibetan) are referred to in many of the studies. Baic (study 5) and the Ta-li languages (studies 13-14) might constitute peripheral Sinitic groups as well. The family most widely represented is Tungusic. This can be seen not only from the separate **part III** on Tungusic (studies 8-12) but also from the fact that most of the papers from **part I** (i.e., studies 1-4) and **part II** (i.e., studies 6-7) also contain aspects relating to typological and areal aspects of Tungusic languages. At first glance, **part IV** on the Ta-Li languages is unrelated to Tungusic. However, since the 1980s a part of the former Ta-Li speakers — the Luren — are officially classified as Manchus, one of the officially recognized Tungusic minorities of China. While this classification has been shown to be mistaken (study 14), it led to the presence of symbols of the Manchus and inscriptions in written Manchu in the villages of the former Luren (see study 13).

[1] See here for an external index of all languages, areas, and families addressed: https://andreashoelzl.de/index-habilitation/

This frontmatter of the habilitation thesis consists of a comprehensive introduction (Section 2), a short conclusion (Section 3), and a few remarks on the division of labor and corrections. The following Introduction (Section 2) will first discuss the concept of a General or Panchronic Comparative Linguistics (Section 2.1) and will briefly introduce the languages of East Asia, focusing on those included here (Section 2.2). The following four Sections 2.3, 2.4, 2.5, and 2.6 summarize and discuss selected topics that are found throughout the individual studies of **parts I**, **II**, **III**, and **IV**, respectively.

## 1.2 Overview of the studies

The following is a complete list of all studies included in this cumulative habilitation thesis:

**Comparative linguistics and language typology**

1. Andreas Hölzl. 2015. A typology of negation in Tungusic. *Studies in Language* 39(1). 118–159. DOI: https://doi.org/10.1075/sl.39.1.05hoe
2. Andreas Hölzl. 2025a. A typology of the deictic day name system in Manchuric. *Eurasiatica: Quaderni di studi su Balcani, Anatolia, Iran, Caucaso e Asia Centrale* 22. 189–217. (**Open access**.) DOI: http://doi.org/10.30687/978-88-6969-913-9/008
3. Andreas Hölzl. 2024p. The complexification of Tungusic interrogative systems. In Dankmar Enke et al. (eds.), *Language change for the worse* (Studies in Diversity Linguistics 33), 211–241. Berlin: Language Science Press. (**Open access**.) DOI: https://doi.org/10.5281/zenodo.13347670
4. Andreas Hölzl. 2022a. 'What's your name?' in Tungusic and beyond. In Andreas Hölzl & Thomas E. Payne (eds.), *Tungusic languages: Past and present* (Studies in Diversity Linguistics 32), 89–148. Berlin: Language Science Press. (**Open access**.) DOI: https://doi.org/10.5281/zenodo.7053365

**Comparative linguistics and language contact**

5. Andreas Hölzl. 2024j. Postnominal flagging and OV in Sinitic: Areal and typological perspectives. *Studies in Language* 49. 1–45. DOI: https://doi.org/10.1075/sl.23055.hol
6. Andreas Hölzl. 2023a. Diachronic changes in constructional networks: Evidence from Manchuric nominal morphology. *Constructions and Frames* 15(2). 160–186. DOI: https://doi.org/10.1075/cf.00071.hol
7. Andreas Hölzl. 2018b. Constructionalization areas: The case of negation in Manchu. In Evie Coussé, Peter Andersson & Joel Olofsson (eds.), *Grammaticalization meets Construction Grammar* (Constructional Approaches to Language 21), 241–276. Amsterdam: Benjamins. DOI: https://doi.org/10.1075/cal.21.c9

**Comparative linguistics of Tungusic languages**

8. Andreas Hölzl & Thomas E. Payne. 2022a. Introduction. In Andreas Hölzl & Thomas E. Payne (eds.), *Tungusic languages: Past and present* (Studies in Diversity Linguistics 32), 1–20. Berlin: Language Science Press. (**Open access**.) DOI: https://doi.org/10.5281/zenodo.7053359
9. Andreas Hölzl. 2022b. The etymology of "Manchu": A critical evaluation of the riverside hypothesis. *International Journal of Eurasian Linguistics* 4(2). 160–208. DOI: https://doi.org/10.1163/25898833-00420028
10. Andreas Hölzl. 2020a. Bala (China): Language snapshot. *Language Documentation and Description* 19. 162–170. (**Open access**.) DOI: https://doi.org/10.25894/ldd76
11. Andreas Hölzl. 2021e. The only known text from Bala, an extinct Tungusic language. *Studia Orientalia Electronica* 9(1). 173–191. (**Open access**.) DOI: https://doi.org/10.23993/store.101600
12. Andreas Hölzl. 2018c. Udi, Udihe, and the language(s) of the Kyakala. *International Journal of Diachronic Linguistics and Linguistic Reconstruction* 15. 111–146. DOI: https://doi.org/10.5281/zenodo.15657731

**Comparative linguistics of the Ta-Li languages**

13. Andreas Hölzl. 2021a. Longjia (China): Language contexts. *Language Documentation and Description* 20. 13–34. (**Open access**.) DOI: https://doi.org/10.25894/ldd29
14. Andreas Hölzl & Yadi Hölzl. 2025. The lost voices of Guizhou: Rediscovering the languages of the Luren. *Cahiers de Linguistique Asie Orientale*. 1–39. DOI: https://doi.org/10.1163/19606028-bja10054

# 2 Introduction

## 2.1 General or Panchronic Comparative Linguistics

This habilitation thesis is concerned with different aspects of *Comparative Linguistics* (German *Vergleichende Sprachwissenschaft*) as applied mainly to East Asian languages. Here, Comparative Linguistics is understood as a general, panchronic approach to linguistic diversity making use of data from more than one language (or doculect) to address specific linguistic phenomena.

The comparison of languages or *Sprachvergleichung*, to use an older German expression (e.g., von Humboldt 1946 [1820]: 9; von Klaproth 1831: 35), is an essential undertaking for an understanding of the linguistic diversity of the world. Arguably, a comparative aspect is beneficial or even necessary for many different lines of research (e.g., W. P. Lehmann 1992), including *contrastive* approaches addressing two or more languages, historical *comparative* reconstruction that usually involves dozens of languages, or *cross-linguistic* typological studies that can include up to several hundred or more languages. Even *descriptive* work on a single language is often assumed to require a comparative background that provides cross-linguistically valid (or universal) categories (cf. C. Lehmann 1989; Haspelmath 2010; Croft 2016 for a discussion of the problem).

A comparison usually involves putting two or more things in close (spatial) proximity to examine them together in detail, either simultaneously or in sequence. But *to compare* (ultimately from Latin *con* 'with' + *pār* 'equal') more generally also means to put them *on a par with* each other, at least to some degree. To be comparable, these two things necessarily share some common property, function, or shape (traditionally known as the *tertium comparationis*, cf. Krzeszowski 1984), i.e. they need to be similar or alike in some respects. Such a comparison thus offers a reference point that is based on (perceived) similarities and is useful for the identification of differences (see the idea of the *comparative concepts* below). These similarities and differences in turn allow a categorization or classification into types, the establishment of scales, inferences on diachronic changes, etc.

However, over at least the last two and a half centuries, the word *comparative* has been used within linguistics in a variety of different senses (e.g., Collinge 1995; Morpurgo Davies 1992; Fox 2006; McElvenny 2024: 6-18). Before attempting a more concrete definition of Comparative Linguistics, a brief overview of previous uses of the term is therefore in order. Note that I will concentrate on purely linguistic approaches and will neglect related fields like (Indo-European) *comparative mythology* (e.g., Mallory & Adams 2006: 423-441), *comparative religion*, or *comparative literature* (German *Komparatistik*).

**Comparative Dictionary/Vocabulary/Word List** The term *comparative vocabulary* has been used very early, such as in the 'Comparative vocabularies of all the languages of the world' (Latin *Linguarum totius orbis vocabularia comparativa*) by Pallas (1786, 1789). Often, such attempts consist of a synchronic comparative list (of genetically related languages) that is based on translating the metalanguage (e.g., English, Latin). However, they may contain some diachronic information or reconstructions, such as in the 'comparative dictionary' (Russian *sravnitel'nyj slovar'*) of Tungusic by Cincius (1975/77). In this case, the comparison is necessarily based on cognates.

**Comparative Linguistics** This field of study is often — implicitly or explicitly — understood as *(Historical) Comparative Linguistics* (German *(Historisch-)Vergleichende Sprachwissenschaft*), i.e. exclusively focusing on **diachronic** aspects (e.g., Szemerényi 1990; McMahon 1994: 2; Trask 2000: 64; Mallory & Adams 2006: passim). This reading of the term is made explicit in the following representative quotation (my boldface):

> We say that languages are related to each other when they are all derived from one common ancestor. 'Comparative linguistics' is the term we use to describe the study of the relationships which exist between such cognate languages. It is thus **a form of historical linguistics**, though historical linguistics is not always necessarily comparative in nature. (Beekes 2011: 4)

Consequently, this field is also often referred to as *Comparative-Historical Linguistics* (German *Vergleichende Historische Sprachwissenschaft*). The name is furthermore frequently used in the context of Indo-European languages and was sometimes understood as almost synonymous with Indo-European Linguistics (Brugmann 1904: 28; Krahe 1970; cf. Ross & Durie 1996: 4), a field that is also occasionally referred to as *Comparative Indo-European Linguistics* (German *Vergleichende Indogermanische Sprachwissenschaft*) (e.g., Beekes 2011).

Anttila (1989: 19) has a slightly more narrow understanding of *Comparative Linguistics* as part of a more general field of study (my boldface):

> **Genetic linguistics** is a cover term for **both historical and comparative linguistics** because both deal with languages showing genetic affinity: historical linguistics treats linear relationships, and comparative linguistics treats collateral relationships. Historical linguistics treats change of various kinds; comparative linguistics sees through change as best it can and establishes earlier stages when much of the change had not yet taken place. (Anttila 1989: 19)

The rather restricted use of the term is connected with the so-called comparative method, "the central method in comparative linguistics" (Anttila 1989: 229) (on which see right below).

As opposed to these views, there is also a more synchronic interpretation of Comparative Linguistics. For instance, Haspelmath (2010: 664) explicitly equates the term with typology when stating that the two terms can be used "interchangeably"; but he also indicates that it is slightly "broader" in scope.

**Comparative Method** The emphasis on diachronic aspects is perhaps most obvious in the term *comparative method* (German *(Historisch-)Vergleichende Methode*) that is often explicitly contrasted with synchronic approaches: "The comparative method is **central to historical linguistics**. It is the method by which we demonstrate linguistic relatedness and reconstruct proto-languages." (Bowern 2017: 1, my boldface) The comparative method usually starts with *comparative word lists* of potential cognates. From these, sound correspondences are extracted that, in turn, allow the inference of regular phonological developments (or sound laws). In a sense, it is thus fundamentally based on *comparative phonology* (e.g., Jakobson 1955: 13-17). The comparative method is usually understood as including the reconstruction of earlier stages of a language family with the help of the established sound laws. But this is also sometimes referred to explicitly as *comparative reconstruction* (e.g. Hock & Joseph 2009: 16). Despite certain limitations, the comparative method to this day is considered "the gold standard" by many representatives of historical linguistics (Kiparsky 2015: 65). It is further exemplified in Section 2.5.4 (see also Anttila 1989: 229-263; Ross & Durie 1996; Rankin 2003; Mallory & Adams 2006: 39-45; Hock & Joseph 2009: 427-454; Fortson 2010: 1-4; Campbell 2013: 107-158; Bybee 2015: 209-236; Trask 2015: 191-233; Weiss 2015; Hölzl 2024b).

**Comparative Philology** An older synonym for Comparative Linguistics emphasizing the diachronic dimension — *Comparative Philology* (German *Vergleichende Philologie*) — is no longer widely used as the term *philology* in the meantime has changed its meaning (e.g., Latham 1862; Lord 1966: 7; Trask 2000: 64). Nevertheless, philology as the study of written texts of a language remains important, especially for historical linguistics that usually focuses on the oldest available data from languages.

**Comparative Grammar** Another term that is frequently understood as referring to the diachronic study of more than one language is *Comparative Grammar* (German *Vergleichende Grammatik*) (von Schlegel 1808: 28; Schleicher 1848: 27; Bopp 1857; Brugmann 1904; Meillet 1908; Jakobson 1955: 17-21). The exact meaning of the term is often left unexplained but is usually understood as involving the comparison of diachronic changes: "Accordingly, a 'comparative grammar' of Greek and Latin means starting from the very beginning and recounting perhaps 3,000 years of virtually independent **lines of development**." (Sihler 1995: 2, my boldface) In similar vein, the classical work on the Tungusic languages by Benzing (1956) is referred to as an attempt at a *comparative grammar* of Tungusic languages (German *Versuch einer vergleichenden Grammatik*) and is primarily concerned with diachronic aspects, the comparative method, and the reconstruction of Proto-Tungusic. Note that the term *grammar* here is often understood as referring to the whole language (including phonology and the lexicon) rather than the grammatical structure (as opposed to the lexicon). Although relatively infrequent, the term *Comparative Grammar* in a mostly diachronic sense remains in use in contemporary publications (e.g., Sihler 1995; Schulze 2005: 64; Fulk 2018).

However, the term *Comparative Grammar* more recently is also employed for what appears to be mostly synchronic typological research (e.g., Syea 2017, 2024; Sidwell 2020). Following this usage, a series with Language Science Press edited by Martin Haspelmath is called *Research on Comparative Grammar* and focuses on "the area of broadly **comparative typological linguistics** that takes into account the world-wide diversity of human languages."[2] (my boldface) Martin Haspelmath prefers the term *Comparative Grammar* (or the more general *Comparative Linguistics*) because it is more easily understood outside of the field than the name *typology* (p.c. 2025). C. Lehmann (1989) employs the term *General Comparative Grammar* as a subfield of *General Comparative Linguistics*.

**Comparative Syntax** A subfield of Comparative Grammar in the first sense was referred to as *(Historical) Comparative Syntax* (German *(Historisch-)Vergleichende Syntax*) (e.g., Brugmann 1904: VIII-IX). Delbrück (1893: 10) argued "that it cannot be the task of a **historical comparative syntax** to discuss all the sentence-connecting devices of the individual languages, but that it must be content with **uncovering the oldest layer"**.[3] (my boldface) Here, once again, we see a usage very similar to the one found in the comparative method above.

However, as early as by von der Gabelentz (1869, 1875), the term was also used with a basically typological meaning. Like Comparative Grammar, the term *Comparative Syntax* is still in use in a synchronic sense today, but has several different interpretations, including typological (e.g., Holm & Patrick 2007; Haspelmath 2014) and theoretical ones (e.g., Kayne 2013). Both interpretations can also be found in the forthcoming *Cambridge Handbook of Comparative Syntax* (published in 2026) that gives "a comprehensive overview of comparative syntax, the study of universal and variable properties of the structure of building blocks in natural language."[4]

[2]For the definition, see the series' homepage: https://langsci-press.org/catalog/series/rcg (2025-03-30)

[3]In the original German: "dass es nicht die Aufgabe einer historisch-vergleichenden Syntax sein kann, alle Satzverbindungsmittel der Einzelsprachen zu erörtern, sondern dass sie sich begnügen muss, die älteste Schicht aufzudecken."

[4]For the explanation, see the volume's homepage: https://tinyurl.com/43czpc52 (2025-04-03) The volume was published after the completion

**(Comparative) Typology** The field known as linguistic typology, language typology, or simply typology (German *(Sprach)typologie*) — a term first coined by von der Gabelentz (1894) — usually makes use of the terms *cross(-)linguistic* (German *sprachübergreifend*) or *cross(-)linguistic comparison*, which in some uses come very close to being synonyms to *comparative* but usually imply the inclusion of a larger sample of languages or refer to properties that several or all languages have in common (e.g., Comrie 1989; Croft 2003; Moravcsik 2013; Song 2018 among others). Typology, defined as "the systematic study and comparison of language structures" (Veluppilai 2012: 1), obviously involves a comparative component at its very core. Therefore, it can be understood as a branch of Comparative Linguistics (Hölzl 2024n). Occasionally, the term *Comparative Typology* is employed, most notably by Hawkins (1986) for his comparison of English and German. Since typology already presupposes a comparison of sorts, the term is perhaps tautological; unless comparative-*typological* is understood as contrasting with comparative-*historical* (see right below). A similar problem exists for the expression *cross-linguistic typology* that is sometimes used. Typology is further discussed in Section 2.3.

The classification of languages (or constructions) into *types* (Hölzl 2024m) presupposes differences between these languages, i.e. a certain amount of (structural) linguistic diversity (see Section 2.2.3). An alternative but less common term for typology that emphasizes this variability of language structure is, therefore, *Diversity Linguistics* (German *Vielfaltslinguistik*, also the name of a conference series). For instance, a former series with Language Science Press edited by Martin Haspelmath is entitled *Studies in Diversity Linguistics* and included, next to reference grammars, work "in broadly **comparative typological linguistics** that takes into account the world-wide diversity of human languages."[5] (my boldface) Three of the studies within this thesis were published in this series (Section 1.2).

**Comparative Concepts (and Descriptive Categories)** Within the field of typology, there is a common problem concerning the comparability of different languages. Haspelmath (2010) and Croft (2016) draw a distinction between *comparative concepts* and language-specific *descriptive categories* (my boldface):

> **Comparative linguists** create **comparative concepts** against which the descriptive categories of particular languages can be matched. These comparative concepts must be universally applicable; that is, they must be based exclusively on more primitive universally applicable concepts: universal conceptual-semantic concepts, general formal concepts, and other comparative concepts (or on extralinguistic situations). (Haspelmath 2010: 681)

This difference is perhaps most obvious in the distinction between a universal conceptual space against which language-specific semantic maps can be compared (see Section 2.3). Here, *comparative* is again used in a mostly synchronic sense. More recently, Haspelmath (2025) has also used the term *general-comparative* in this meaning.

**Contrastive Linguistics** A field of research closely related to typology is *Contrastive Linguistics* (German *Kontrastive Linguistik*). Like typology, it includes an obvious comparative aspect. The main difference, however, appears to be a restriction of the comparison to only a few or even two languages (e.g., Theisen 2016). As such "contrastive linguistics can be viewed as a pilot study in typology" (van der Auwera 2012: 69). While important insights can be gleaned from contrastive approaches (e.g., for language teaching), the restriction to a few languages is merely a difference in quantity, not quality.

**Comparative Language Science** A very general approach can be found at the Institute for the Interdisciplinary Study of Language Evolution in Zurich, where *Comparative Linguistics* is broadly defined as "the study of human language as a species-specific phenomenon in all facets of its occurrences."[6] More recently, the term was replaced by the newly coined concept of *Comparative Language Science*, a literal translation from the German *Vergleichende Sprachwissenschaft* that seems to highlight the connection with the natural sciences (especially biology) and a quantitative approach. It also increasingly references animal communication systems and aspects of evolution.

**Towards a General or Panchronic Comparative Linguistics** All studies in this habilitation thesis include some *comparative* component, though they do not fit neatly into the categories discussed above. In fact, most of the aforementioned aspects are represented in one way or another. For instance, studies 1 (Hölzl 2015), 2 (Hölzl 2025a), 3 (Hölzl 2024p), 4 (Hölzl 2022a), and 5 (Hölzl 2024j) have a clear typological, i.e. *cross-linguistic* background. Some of these include a discussion of conceptual spaces and thus of *comparative concepts* (i.e., studies 1, 2, 5). Given the focus on Tungusic languages in several of the studies, some of them could well be referred to as *Comparative Grammar* or *Syntax* of these languages, either synchronic or diachronic (e.g., studies 1 Hölzl 2015, 4 Hölzl 2022a, 8 Hölzl & Payne 2022a). In a way, study 6 (Hölzl 2023a), comparing Manchu with Udihe, or study 7 (Hölzl 2018b), focusing on a comparison of Manchu and Mongolian, could be understood as *contrastive* in nature, though both include several other languages as well. The *comparative method* is applied to Tungusic languages in several studies of this

of this thesis.

[5] For the definition, see the series' homepage: https://langsci-press.org/catalog/series/sidl (2025-05-15)

[6] For the definition, see the following manuscript: https://tinyurl.com/2u8nvcr5 (2025-03-30)

habilitation thesis, most prominently study 9 on the *comparative phonology* of Tungusic and the reconstruction of Proto-Tungusic *mg (Hölzl 2022b). Studies 10 (Hölzl 2020a) and 13 (Hölzl 2021a) are *descriptive* studies of Bala and Longjia, respectively, but crucially also contain comparisons with other languages such as Manchu and Caijia, to put the data into perspective. Some studies include a superficial *comparative philological* component. For example, study 11 (Hölzl 2021e) analyzes the only known Bala text with comparative data from other Manchuric varieties. Several studies, including 9 (Hölzl 2022b: 195-200), 12 (Hölzl 2018c), or 14 (Hölzl & Hölzl 2025), contain lexical comparisons and *comparative word lists*.

A label such as *typology* is therefore too narrow for this habilitation thesis. A broad and general name is called for as a cover term for all of the studies assembled here. Since the individual studies do not only include aspects of grammar, such as syntax, but also of the lexicon and phonology, I avoid the terms *Comparative Grammar* and *Comparative Syntax* as labels, too. Given the emphasis on quantity and evolution in *Comparative Language Science*, I do not use this term either. The name *Contrastive Linguistics* is inadequate, since a large number of approximately 250 doculets is included in the studies. Instead, I make use of the broader and more traditional term *Comparative Linguistics*, but propose a more *general* understanding of that term from a *panchronic* perspective. While the diachronic dimension remains crucial, Comparative Linguistics necessarily also includes a variety of other aspects that go beyond the traditional comparative approach, including typological research.

Similar views are, of course, already present in many earlier approaches, such as P. W. Schmidt (1926: 28), for whom Comparative Linguistics is "to investigate how the comparison of languages with one another reveals both their uniqueness and their historical connections with other languages."[7] As Nichols (1997: 360) puts it, such a general approach entails "a broader field of comparative linguistics whose concerns are not only genealogical" in nature. Ellis (1966) and C. Lehmann (1989) use the name *General Comparative Linguistics* for such a field of study.

The term *panchronic* is not very common in the field but has previously been used in a similar way for the characterization of Greenberg's "recognition of the inseparability of diachronic and synchronic considerations in linguistic research" (Croft, Denning & Kemmer 1990: ix).[8] That many properties of language structure cannot exclusively be explained synchronically but require an additional diachronic dimension today comes close to being a truism in linguistic typology (e.g., Dryer 2019 on certain word order properties). As such, the approach proposed here could be referred to as *amphi*chronic, i.e. as *both syn*chronic and *dia*chronic (cf. Kiparsky 2006). For instance, study 4 on the "personal name question" — *What's your name?* in English — contains a cross-linguistic and largely synchronic background but also addresses the diachrony of the question and its component parts within the history of the Tungusic language family (Hölzl 2022a). However, linguistic diversity is the result of a multitude of different factors. Being part of a specific question-answer sequence, the personal name question also requires certain *enchronic* (i.e., interactional) aspects. Arguably, a proper description of its meaning furthermore requires *microgenetic* (e.g., cognitive) aspects, such as the reference to a larger frame of shared background knowledge (referred to as the "personal name frame"). The classification of Manchuric as a semi-creole (Section 2.4) involves modifications in language structure by means of the uncontrolled second-language acquisition by speakers of other languages (i.e., an *ongotenetic* factor). Therefore, a more general and comprehensive Comparative Linguistics as advocated here should probably include all "causal frames" described by Enfield (2014: 65):

> *microgenetic*, invoking cognitive and motoric processes for producing and comprehending language and other goal-directed behavior; *ontogenetic*, invoking lifespan processes by which people, usually as children, acquire linguistic and cultural knowledge and skills; *phylogenetic*, invoking ways in which the requisite cognitive capacities have evolved in our species; *enchronic*, invoking the sequential interlocking of social actions in linguistic clothing; *diachronic*, invoking historical change, conducted socially in human populations; and *synchronic*, any approach, such as linguistic or ethnographic description, that does not explicitly invoke notions of process.

The term *pan*chronic is here defined as a cover term for *all* of Enfield's causal frames, although the main focus in this habilitation thesis still lies on *syn*chronic and *dia*chronic aspects.

In sum, I proposes the name *Comparative Linguistics* (German *Vergleichende Sprachwissenschaft*) as a cover term for a broad field of study that includes both historical comparative linguistics and linguistic typology alongside other components. For contrastive purposes, it can also be referred to as ***General*** or ***Panchronic*** *Comparative Linguistics* (German *Allgemeine/Panchron(isch)e Vergleichende Sprachwissenschaft*) to distinguish it from purely synchronic or diachronic approaches. Research in Comparative Linguistics can be conducted either quantitatively or qualitatively. While both approaches offer distinct advantages, this habilitation thesis is mostly concerned with the latter. Most of the studies assembled here do not assume any specific framework or theory, although three of the studies have a more or less specific constructionist background (studies 4, 6, 7) and some aspects of Cognitive Linguistics or Cognitive Grammar are used for the analysis or description of semantic aspects (e.g., studies 4, 9).

[7] In the original German: "den Versuchen nachzugehen, wie aus der Vergleichung der Sprachen untereinander sowohl ihre Eigenart als auch ihre geschichtlichen Zusammenhänge mit anderen Sprachen zutage gefördert werden."

[8] Note that the term is also used in a slightly different meaning in phonology (e.g., Jacques 2011).

## 2.2 The languages of East Asia

### 2.2.1 Loose Definition

The studies in this habilitation thesis focus on languages in *East Asia* (EA). However, EA is variously defined by different scholars (e.g., Goddard 2005; Heggarty & Renfrew 2014; Hölzl 2018a). Drawing clear-cut boundaries is neither feasible nor helpful for the analysis of linguistic diversity in the region. Hence, it is here loosely defined as stretching from Chukotka in the North to Mainland Southeast Asia (MSEA) in the South. The eastern border is formed by the Pacific Ocean, Taiwan, and the Japanese Archipelago. In the West, the area under investigation finds its approximate border at the Yenisey, the Pamir mountains, and the Himalayas. This definition roughly includes the following political entities: Cambodia, China, Japan, Korea, Laos, Mongolia, Myanmar, Russia east of the Yenisey river, Taiwan, Thailand, and Vietnam. It may also include some adjacent parts of Kazakhstan, Nepal, Northeast India, and Maritime or Insular Southeast Asia (ISEA). This habilitation thesis includes languages from the entirety of this vast area, but naturally the focus is on a few of them. Given the typological aspect, the individual studies — especially in **part I** — also address languages from other parts of the world. East Asia as defined here does not constitute a coherent linguistic area and exhibits strong differences in terms of climate, geography, population density, and linguistic diversity. It can roughly be divided into *Northeast Asia* (NEA) (e.g., Hölzl 2018a) and *(Mainland) Southeast Asia* (SEA) (e.g., Enfield 2018, 2021, Enfield & Comrie 2015a). *Sinitic* languages are located in between these two areas geographically and typologically (e.g., Szeto & Yurayong 2021). The northern part of East Asia is represented here by Tungusic languages in **parts I-III** (Section 2.5), Sinitic languages by study 5 (Hölzl 2024j), and the southern part by the Ta-Li languages in **part IV** (Section 2.6), although the latter could be archaic or peripheral Sinitic languages as well.

### 2.2.2 Relevance

East Asia in many ways played a crucial role in the development of the linguistic diversity of the world we observe today. Many languages of Europe, including Turkic (e.g., Chuvash, Balkan Turkic, Bulgar, Tatar, Turkish), Mongolic (e.g., Kalmyk), and possibly even Uralic (e.g., Estonian, Finnish, Hungarian, Mari, Mordvin, Saami) trace their origin to EA as defined above. Recent genetic evidence has shown that parts of the European Avars (Gnecchi-Ruscone et al. 2022) and Huns (Gnecchi-Ruscone et al. 2025), especially the elites, originate from or have ties to EA. Although the classification of their languages remains disputed, these may have belonged to language families still in existence in EA. Genetic evidence has also confirmed that Indo-European languages were most likely spoken by the representatives of the Yamnaya culture and the preceding genetic Caucasus–lower Volga cline (e.g., Lazaridis et al. 2025; Mallory 2025) that trace parts of their genetic legacy to the Ancient North Eurasians in southern Siberia (e.g., Allentoft et al. 2024). Moreover, NEA is the origin of all Native American populations and, therefore, of all Native American languages. The predecessors of all approximately 170 Native American language families at one point were thus all spoken in NEA (e.g., Reich 2018; Raff 2022; Campbell 2024), possibly associated with archaeological complexes such as the Djuktaj culture in the Trans-Baikal region (e.g., Potter 2023, p.c. 2024). Presumably, later migrations brought Na-Dene languages, the unknown language or languages of the so-called Paleo-Eskimos, and the Eskaleut languages from NEA into North America and Greenland. Speakers of Austronesian languages spread from Taiwan and ultimately reached Easter Island, Hawaii, Indonesia, Madagascar, New Zealand, the Philippines, and, as confirmed recently through genetic evidence, even South America (e.g., Ioannidis et al. 2020). The Munda languages in India (a branch of Austroasiatic) also likely originate in Southeast Asia (Anderson 2015). While this list is not exhaustive, more importantly EA until today remains a treasure trove of linguistic diversity itself.

### 2.2.3 Linguistic diversity

Nettle (1999: 10) differentiates between three types of linguistic diversity: (1) phylogenetic diversity (the number of language families or "isolates", here collectively referred to as stocks), (2) language diversity (the number of languages), and (3) structural diversity (the number of different language types). This section briefly addresses all three types of diversity within East Asia.[9]

**Phylogenetic (genealogical) diversity** Excluding recent migrations, East Asia as defined above includes languages from some 18 different stocks of vastly different size, age, and geographical distribution: Ainuic, Amuric (Nighvng and Nivkh, earlier Gilyak), Austroasiatic, Austronesian, Chukotko-Kamchatkan, Eskaleut (Eskimo-Aleut), Hmong-Mien (Miao-Yao), Indo-European, Japonic (Japanese-Ryūkyūan), Khitano-Mongolic, Koreanic, Kra-Dai (Tai-Kadai), Trans-Himalayan (Sino-Tibetan), Tungusic, Turkic, Uralic, Yeniseian, and Yukaghiric. If East Asia is defined even more broadly, stocks like Burushaski in Northern Pakistan, Kusunda in Nepal, or Ongan and Great Andamanese on the Andaman islands could be included in the list.

[9] The external index contains a list of all languages and stocks addressed here: https://andreashoelzl.de/index-habilitation/

The phylogenetic diversity is distributed very unevenly across the area. Whereas NEA contains 14 different stocks, many of which are located at or close to the Pacific Rim, SEA only contains 5 stocks. Trans-Himalayan (mainly Sinitic) is found in both NEA and SEA. According to Campbell (2018), there are approximately 400 stocks worldwide, 406 to be more precise, though this figure remains uncertain. According to this estimate, EA harbors some 4.5% of the worldwide phylogenetic diversity. Including the four peripheral stocks mentioned above, this percentage rises to 5.5%.

However, the number of stocks in EA is not entirely certain given the possible existence of previously overlooked language "isolates", i.e. languages with no proven genealogical affiliation to any other family, and macro-families. Apart from the case of Jiamao on Hainan (Norquest 2016), one potential language isolate discussed in study 14 (Hölzl & Hölzl 2025) is referred to as "Luren B" (in contrast with the "Luren A" or Lujia language). While some of the proposed macro-families, such as "Altaic" (even if relabeled as "Trans-Eurasian"), remain problematic (see Janhunen 2023 for a discussion), some other proposals appear more promising. These include, for example, "Austro-Tai" (i.e., Austronesian and Kra-Dai) as formulated by Ostapirat (2005, 2018) or "Dene-Yeniseian" (i.e., Yeniseian and Na-Dene languages in North America) as developed by Vajda (2010, 2022). If any of these or other proposed families is accurate, the estimate of phylogenetic diversity in East Asia would need to be revised accordingly. The studies in this habilitation thesis mention most of the stocks found in East Asia (with the exceptions of Ainuic, Chukotko-Kamchatkan, Ongan, and Yeniseian). But, as specified in Section 1, a focus lies on some of them, including Sinitic (perhaps encompassing Baic and the Ta-Li languages), a major branch of Trans-Himalayan, (Khitano-)Mongolic, and especially the Tungusic languages as all four parts of this thesis address them in one form or another.

**Language diversity** The exact number of languages in East Asia is more difficult to determine than the number of stocks. Well-known problems include the uncertain boundary between language and dialect, the existence of extinct languages, etc. NEA harbors only relatively few languages, some of which are distributed over vast areas (e.g., Russian, Evenki, Mongolian, Sakha, Japanese, Korean), whereas SEA has a much larger number of individual languages, many of which are distributed over comparatively small areas. Whereas the educated guess for NEA by Hölzl (2018a: 44) is around 120-150 languages, Enfield & Comrie (2015b: 6) list 280 languages for MSEA and 583 for what they call "Greater MSEA" that includes ISEA. If the approximate number of up to 700 languages is roughly accurate, this represents about 10% of the languages of the world (assuming a total of approximately 7,000 languages).[10] This thesis addresses over 250 doculects, including languages from outside of East Asia.

**Structural diversity** The structural diversity is still more difficult to ascertain. The outcome will depend on the level of description and the selection of features. In a phonological typology, one could draw a distinction between tonal languages in SEA and Sinitic, and mostly atonal languages in NEA. In a morphological typology, the languages of East Asia range from mainly "isolating" in the south (e.g., Vietnamese), through largely "agglutinating" (or concatenative) in most of the northern part (e.g., Mongolian), to "polysynthetic" in the far north (e.g., Kerek) (e.g., Hölzl 2024c,i). Syntactically, there is a clear contrast concerning some features, such as dominant OV in NEA as opposed to VO in SEA (Dryer 2013c). This contrast is even visible in features not directly correlating with basic word order, such as dominant AdjN in NEA versus NAdj order in SEA (Dryer 2013a). In this thesis, structural diversity is mainly addressed in **parts I** (Section 2.3) and **II** (Section 2.4) that discuss aspects of negation (studies 1 Hölzl 2015, 7 Hölzl 2018b), the lexical field of deictic day names (studies 2 Hölzl 2025a), questions and interrogative systems (studies 3 Hölzl 2024p, 4 Hölzl 2022a), and argument marking (studies 5 Hölzl 2024j, 6 Hölzl 2023a).

**Sinitic** Sinitic languages are found over a vast area from Northeast China to Hainan in the South and from Shandong in the East to Xinjiang in the West. As observed by Dryer (2017) and others, Sinitic has several cross-linguistically unusual word order properties that appear intermediate, e.g. mostly VO as in SEA but RelN as in NEA. Furthermore, the internal division of Sinitic approximates the distinction between NEA (fewer and larger languages like Jin or Mandarin) and SEA (more and smaller languages like Gan, Hakka, Xiang, Yue, Min, Wu) (e.g., Ramsey 1987). This distinction also has some typological correlates, such as a cline from fewer tones in the North to more tones in the South. After investigating 30 different features, Szeto & Yurayong (2021: 594) conclude that

> the typological variation across Sinitic is largely consistent with the Altaic-MSEA contrast; to wit, many of the features which mark the typological contrast between Altaic and MSEA are precisely those which also highlight the north-south divide of Sinitic

"Altaic" here refers to a typological grouping of languages (i.e., Khitano-Mongolic, Turkic, Tungusic) that all had contact with (Northern) Sinitic and are typologically representative of NEA. The Sinitic languages (including Baic) as well as areal interactions with languages to the north and west (e.g., Mongolic, Tibetic, Turkic) are discussed in Study 5, which focuses on word order and the emergence of postnominal flagging (Hölzl 2024j).

The Ta-Li languages in Guizhou may also represent peripheral Sinitic languages. While they generally exhibit many features representative of SEA (e.g., tones, classifiers, isolating morphology), they also have typical Sinitic features, such as RelN & VO. Additionally, they show further unusual feature combinations, such as prenominal adjectives as in NEA in combination with postnominal numerals as in many languages in SEA (Section 2.6).

[10] For instance, *Ethnologue* currently lists 7,159 languages: https://www.ethnologue.com/insights/how-many-languages/ (2025-05-18)

## 2.3 Comparative linguistics and language typology

### 2.3.1 Classification of languages

Comparative Linguistics usually proceeds in several steps. A first step is mostly concerned with the identification of similarities (and differences) between languages or parts of languages. A second step attempts to classify and explain these similarities. In a classic overview, Aikhenvald & Dixon (2001: 1-4) differentiate between (a) universal properties or tendencies, (b) chance resemblances, (c) borrowing or diffusion, (d) genetic retention, and (e) parallel development. Arguably, the last point is the byproduct of a more fundamental reason that should be added to the list, viz. (f) limited structural (or logical) possibilities. For instance, the relative position of numeral, numeral classifier, and noun only has six logical possibilities. Only four of these are cross-linguistically attested ([Num-Clf]-**N**, [Clf-Num]-**N**, **N**-[Num-Clf], **N**-[Clf-Num]) while two others are not, probably because they violate the constituent order (*Clf-N-Num, *Num-N-Clf) (Her 2017). Given the extremely limited range of possibilities, many languages from around the world will therefore have the same word order (i.e., will be similar) without being in areal contact or genetically related. Note that this differs from chance resemblances such as similar sounding words, because the range of possible sound combinations is nearly endless.

In a third step, languages or parts of languages are classified based on these similarities and differences. Obviously, universal properties and chance resemblances are irrelevant for this endeavor (a + b). Traditionally, there are three ways of classifying languages, viz. from an areal (c), genealogical (d), or typological perspective (e + f). For instance, Sibe is in contact with several languages with numeral classifiers, including Mandarin and Uyghur, and has developed a small classifier system itself, which is rather unique among Tungusic languages (Section 2.5). As opposed to this, the Ta-Li languages all have numeral classifiers which indicates a common inheritance from an earlier proto-language (Section 2.6). From a purely typological perspective the word order in Sibe and the Ta-Li languages can be classified as [Num-Clf]-N and N-[Num-Clf] order, respectively.

These three different types of classification are addressed in **part I** on typology (this section), **part II** on language contact (Section 2.4), and **parts III** on Tungusic (Section 2.5) as well as **IV** on Ta-Li languages (Section 2.6), respectively. While every part focuses on one of these perspectives, the others are nevertheless relevant factors to be taken into account. All three are necessary for a comprehensive classification of languages.

For instance, although study 1 is a typology of negation (Hölzl 2015), it is heavily concerned with genealogical and areal explanations of the types of negation found among Tungusic languages. For instance, for standard negation many Tungusic languages employ an inflected negative verb while the lexical verb stands in a connegative form here referred to as 'fixed form of the negated lexical verb' (FFNLV) following Nedjalkov (1994: 6).

(1) Arong Ewenke

a. saa-ra-n
know-PRS-3SG
'(he/she) knows'

b. **ə**-ʃi-n saa-**ra**
NEG-PRS-3SG know-FFNLV
'does not know' (Chaoke & Kalina 2017: 74)

This type of negation can be directly traced to Proto-Tungusic, i.e. the type of negation can be explained by the retention from earlier stages of the language. However, some Tungusic languages employ other types of negation that are either a development from the Proto-Tungusic pattern or an innovation. For instance, Manchu, probably due to Mongolic influence, has replaced the negative verb with the negative existential *akū* 'there is no', e.g. *sa-r(a)kū* (< *sa-ra akū*) 'do/does not know' (on which see study 7, Hölzl 2018b).

The classification in terms of similarity as sketched above needs to be differentiated from other classifications, for example based on geography. This can be illustrated with the term "Paleo-Asiatic" (or "Paleo-Siberian") languages (Hölzl 2024g). It is widely acknowledged that the term does not designate a valid language family. The languages involved are furthermore not all in mutual contact and do not share the same typological profile (e.g., Hölzl 2018a: 15). Therefore, they also do not form a valid linguistic area, i.e. a region with strong language contact phenomena and the widespread diffusion of similar properties. The term "Paleo-Asiatic" is thus best understood as a mere geographical label for a number of mostly endemic languages or language families in Russian NEA (mostly Amuric, Chukotko-Kamchatkan, Yeniseian, Yukaghiric) that excludes a number of language families of the area that are also distributed outside of the region (e.g., Khitano-Mongolic, Tungusic, Turkic). As such it is comparable to the term "Papuan", which is primarily understood geographically but also excludes all Austronesian and European languages as well as one Pama-Nyungan language found in the area (e.g., Palmer 2018: 4-6).

### 2.3.2 A systemic approach

Languages contain several subsystems or domains that exhibit specific properties and require a system-specific or domain-specific typology. This habilitation thesis contains case studies for several different examples, three of which will be briefly discussed here: (1) the domain of negation, (2) deictic day name systems, and (3) interrogative (word) systems. Of these, the second is conceptually related to the numeral system and relatively simple in structure.

The other two are much more complex, multidimensional, and exhibit connections to many different parts of the language.

One common typological method of mapping the inner structure of a linguistic subsystem or domain is the conceptual space model mentioned in Section 2. Following the terminology in Croft (2003: 134), the universal *conceptual space* can be differentiated from the language-specific *semantic map*. Such a conceptual space is often shown with separate comparative concepts connected by (straight) lines whereas the semantic map is a closed line that includes one or more of these concepts in a contiguous space. The method combines onomasiological and semasiological approaches into one coherent format and allows the visual comparison of languages. For instance, study 5 (Hölzl 2024j) uses the idea of a conceptual space for the comparison of several flags (case markers or adpositions) for categories such as instruments or companions.

**Negation** Studies 1 (Hölzl 2015), 7 (Hölzl 2018b)

Based on Tungusic and some cross-linguistic evidence, Study 1 (Hölzl 2015) attempts a rudimentary conceptual space for the universal domain of negation. Languages differ in the number of negators they possess and in which functions these express. A description thus requires several different comparative concepts, such as standard negation (NEG) or negative existence (NEX). Negators in the languages of the world exhibit language-specific conflation patterns of these concepts. For instance, Italian uses the same negator *non* for the expression of standard negation (***non*** *so* 'I don't know') and negative existence (***non*** *c'è* 'there is no') (NEG=NEX). As opposed to this, the Tungusic language shown in (1b) uses the negative verb *ə-* for the former category but has the specialized negator *aaʃin* for the latter (NEG≠NEX). A somewhat more elaborate conceptual space of negation that includes more categories was developed independently by Veselinova (2015). Both approaches are in need of further research concerning the exact connections between the individual categories and the cross-linguistic frequency of certain patterns. Study 7 (Hölzl 2018b) focuses on the development from negative existence to standard negation in Manchuric.

**Deictic day name systems** Study 2 (Hölzl 2025a)

The deictic day name system is probably universal or near-universal, although the number of distinctions drawn differs from language to language. This system, similar to the numeral system, has the topology of a line. But unlike the numeral system, it has a deictic center ('today'). Both sides of the deictic center are conceptually related to and perhaps based on the lower numeral system, 1 corresponding to 'tomorrow', 2 to 'the day after tomorrow' and so on (e.g., English ***two*** *days ago*). Although the system is theoretically unbounded on both sides of 'today', the system is usually rather limited and does not extend beyond a few days on either side. All individual categories are based on the concept of DAY, which can be seen from many languages that have analyzable forms (e.g., Turkish *bu-gün* 'this-day'). There are typically no gaps in the system and usually relatively few forms per language. Despite being present in most or all languages, there is only very little research on the architecture of this system. Study 2 discusses the limited previous research, sketches a typology of the system, develops a preliminary conceptual space, and describes the properties and development of the deictic day name system in Manchuric (Hölzl 2025a).

**Interrogative (word) systems** Studies 3 (Hölzl 2024p), 4 (Hölzl 2022a)

In comparison, the system formed by the interrogatives, also variously known as *question words*, *interrogative words*, *wh-words* and so forth, is much more complex. Unlike the previous system, it is multidimensional and has no clear topology, although it can also be shown with a conceptual space (on which see Hölzl 2018a; Ding & Dong 2024). Unlike the deictic day name system, the interrogative system contains various semantic categories, such as PERSON ('who'), THING ('what'), PLACE ('where'), etc. Another complicating factor is its relation to other systems, such as the demonstrative and the relative system (e.g., Diessel 2003; Hölzl 2024o).

In addition, unlike the deictic day name system, the interrogative system exhibits a complex relation to the rest of the linguistic system (language) as a whole that can be described as *self-similarity* (Hölzl 2024o). It usually mirrors the word-class distinction of the language in a miniature format, which can be understood as a fractal property. The word class can be observed in distributional (e.g., word order patterns) or morphological properties (e.g., inflection).

A well-known feature of the interrogatives in some but not all languages is a shared initial element, such as *w-* in German (e.g., ***w****er* 'who', ***w****ie* 'how', ***w****as* 'what'). Given that it is shared by all or most interrogatives but not longer fully analyzable, it can also be referred to as a *submorpheme* similar to the shared element of English ***sn****ail*, ***sn****uff*, ***sn****ort*, etc. This initial element is a type of "phonetic-semantic reverberation" (Hockett 1987: 71) also referred to as a *resonance* (e.g., Bickel & Nichols 2007: 209-210; Mackenzie 2009). Following the convention established in Hölzl (2018a), this partly analyzable resonance can be indicated with a tilde, e.g. German *w~*, to differentiate it from fully analyzable morphemes indicated with hyphens. Similar phenomena can also be observed in the demonstrative and relative systems. Consider the selected examples from Nepali in Table 2 that, like many other Indo-Aryan languages, exhibits parallels between the three different classes. Interrogatives start with *k(ə)~*, relatives with *dz(ə)~*, and

demonstratives with either *e~* for proximal or *te~* for distal meaning. Also note the shared final resonances, such as *~səri* for the category of MANNER ('how', 'so', 'like this', 'like that'). In this case, the initial resonance can be considered a *class marker* within the system (columns) while the final resonance functions as a *category marker* across the interrogative, relative, and demonstrative systems (rows) (Hölzl 2024o).

| | **Int** | **Rel** | **Dem (prox)** | **Dem (dist)** | **Category** |
|---|---|---|---|---|---|
| MANNER | **kə**səri | **dzə**səri | **e**səri | **te**səri | ~səri |
| KIND | **kə**sto | **dzə**sto | **e**sto | **te**sto | ~sto |
| LOCATION, DIRECTION | **kə**ta | **dzə**ta | **e**ta | **te**ta | ~ta |
| QUANTITY, EXTENT | **kə**ti | **dzə**ti | **e**ti | **te**ti | ~ti |
| DEGREE | **kə**ttiko | **dzə**ttiko | **e**ttiko | **te**ttiko | ~ttiko |
| **Class** | k(ə)~ | dz(ə)~ | e~ | t(e)~ | |

Table 2: Selected Nepali interrogatives, relatives, and demonstratives (Dubi Nanda Dhakal p.c.)

In some cases, the class and category markers are still analyzable to some degree, such as in German adverbial interrogatives like *wohin* 'where to' (Table 3). But note that German also exhibits many fused interrogatives that are no longer fully analyzable.

| | where (> what) | there/that | here/this | **Category** |
|---|---|---|---|---|
| towards (there) | wo-hin | da-hin | hier-hin | -hin |
| towards (here) | wo-her | da-her | hier-her | -her |
| at, by, with | wo-bei | da-bei | hier-bei | -bei |
| with | wo-mit | da-mit | hier-mit | -mit |
| to, after | wo-nach | da-nach | hier-nach | -nach |
| from, of | wo-von | da-von | hier-von | -von |
| through | wo-durch | da-durch | hier-durch | -durch |
| about, around | wor-um | dar-um | hier-um | -um |
| at, on | wor-an | dar-an | hier-an | -an |
| **Class** | wo(r)- | da(r)- | hier- | |

Table 3: Selected German interrogatives (> relatives) and demonstratives

The function of the (initial) resonance as a class marker (in fused systems) is also confirmed by a quantitative study by Slonimska & Roberts (2017) that investigated 226 languages from 66 families. This study found "that initial segments of wh-words were more similar than initial segments of randomly selected word sets and conceptually related word sets (e.g., body parts, actions, pronouns)" (1) and that "214 out of 226 languages (95%) had interrogative words with initial segments that were more detectable [i.e., salient] than randomly selected words" (13). The study thus proved the cross-linguistic validity of the concept of an initial resonance within the interrogative system.

Although there is limited typological research on the phenomenon, the resonance as class marker crosslinguistically appears to be always word-initial as in German or Nepali. This fundamentally asymmetric property of languages like Nepali is the result of diachronic developments, i.e. the fusion of formerly analyzable forms. It therefore can only be found in so-called *fused* interrogative systems that were formerly analyzable (see Muysken & Smith 1990; Hölzl 2018a). Often, these are based on interrogatives like 'what' or 'which' that combine and subsequently fuse with other elements (lexical or grammatical). For instance, Cameroon Pidgin English contains the fused interrogatives ***w**eiti(n)* (< *what thing*), ***w**usai* 'where' (< *which side*), and ***w**ishtaim* 'when' (< which time') that all can be traced to analyzable expressions in English (Ayafor & Green 2017: 120). In other words, the synchronic properties cannot be explained without reference to both diachronic developments and earlier synchronic stages of the language. As I put it in study 3: "today's resonance is yesterday's morphosyntax" (Hölzl 2024p: 213).

Note that such a class-marking resonance can only develop in languages with certain typological properties. These include dependent-head word order (e.g., *which book*, *how many books*, *what kind of book*, *how far*), postpositions (NAdp) rather than prepositions (AdpN), and suffixing morphology for both derivation and inflection.

(2) German

a. **welch**-es Buch
which-N.SG.NOM/ACC book
'which book'

b. [**was** für ein] Buch
what for a.N.SG.NOM/ACC book
'what kind of book'

On the contrary, prepositions and prefixes prevent a resonance from emerging, e.g. Cameroon Pidgin English

***foseka*** *weiti* ‘why’ (< *for (the) sake of what thing*). In the following example from Caijia, the interrogative follows the head noun (head-dependent order), which will likewise prevent the emergence of a resonance.

(3) Caijia
sɣ$^{33}$ [**lɔ$^{24}$** peŋ$^{21}$]
book which CLF
‘which book’ (Lü 2022: 272)

In the long run, the emergence of such properties will also lead to the gradual but sometimes slow disappearance of resonances. For instance, Italian *perché* ‘why’ is a contraction of the preposition *per* and interrogative *che* into one form (cf. English *what for*). Due to the preposition, the interrogative no longer shares the old resonance in [k].

Note that the relevant properties are, at least to a degree, correlated with basic word order (OV/VO). For instance, there is a well-known bidirectional correlation between prepositions and VO order on the one hand and between postpositions and OV on the other, although SVO languages exhibit a mixed pattern (Dryer 2007, 2011, 2019; Hölzl 2024n). As a result, the resonance as class marker is more commonly encountered in Northeast Asian and other Asian languages with OV order and postpositions (e.g., Indo-Aryan) than in Southeast Asian languages with VO order and prepositions. Northeast Asian languages like Mongolic, Tungusic, or Turkic in addition have a general tendency for suffixing morphology and head-finality, even in cases that probably do not correlate with OV order directly (e.g., AdjN, NumN, DemN, IntN).

Table 4 illustrates this contrast with the help of examples from Bengali (OV) and Indonesian (VO), both of which have at least some fused interrogatives that are not synchronically analyzable. In the former, modifiers (including interrogatives) often precede the head noun (*kʌn boi-ta* ‘which book-CLF’), whereas they follow in the latter (*buku yang mana* ‘book that which’, *buku apa* ‘book what’). Indonesian furthermore has several still analyzable elements preceding the interrogatives, e.g., *di* ‘in’, *ke* ‘to’, or *bagai* ‘like, as’. Over the course of time, these differences lead to very different interrogative systems. Whether the correlation with basic word order is globally valid remains to be investigated with a larger cross-linguistic sample of languages from around the world.

| **Meaning** | **Bengali** | **Indonesian** |
|---|---|---|
| who | **k**e | siapa |
| which one | **k**ʌn-ta | yang mana |
| why | **k**eno | kenapa, mengapa |
| when | **k**ɔk$^{h}$on | kapan |
| to do what | **k**ɔra | untuk apa |
| where | **k**o̪t̪$^{h}$aj | di mana |
| whence | **k**o̪t̪$^{h}$aj | dari mana |
| whither | **k**o̪t̪$^{h}$a t̪$^{h}$eke | ke mana |
| how much | **k**ɔ̪to-ta | berapa (banyak) |
| how many | **k**ɔ̪to-gulo | berapa (banyak) |
| what | **ki** | apa |
| for what reason, what for | **ki** dzonjo | untuk apa |
| how (manner) | **ki** b$^{h}$abe | bagai-mana |
| what kind of | **ki** dhɔroner | yang mana |

Table 4: Parts of the interrogative systems in Bengali (Lahari Chatterjee p.c.) and Indonesian (Saffanah Fathin p.c.)

Study 3 (Hölzl 2024p: 213) is a discussion of the Tungusic interrogative system, focusing on the detrimental consequences of the complete loss in several languages of the initial resonance due to phonological changes (e.g., Manchu *ai*, Nanai ***x**aɪ* ‘what’, Manchu *udu*, Nanai ***x**ado* ‘how many’). The consequences of the change are described as a form of complexification. Note that this regular phonological change needs to be differentiated from reanalysis or cases of aphaeresis in some Indo-European languages that only occurred under specific conditions (e.g., Hackstein 2017) and so-called *atrophied* interrogative systems that are characterized by the sporadic drop of the original interrogative as in Sranan (e.g., Muysken & Smith 1990).

Study 4 (Hölzl 2024p) addresses a few issues relating to the interrogative system, especially the usage of different types of interrogatives like ‘who’ (PERSON), ‘what’ (THING), or ‘how’ (MANNER) within the “personal name question”. For instance, Tungusic originally used a construction that can be translated as ‘Who is your name?’. Due to contact with Chinese, Mongolic, and Russian, some of the languages also started to use other interrogatives, such as ‘what’ and ‘how’. The study proposes that the variable interpretation of the concept NAME as such could be one factor that is responsible for the cross-linguistic use of different interrogatives within the personal name question.

## 2.4 Comparative linguistics and language contact

Several studies in this habilitation thesis have language contact or areal phenomena as their main focus or at least refer to language contact as one factor for the explanation of the patterns observed in the languages of East Asia. The main issues to be addressed in this section are the behavior of flagging (case, adpositions) under language contact (Section 2.4.1), shared developments (Section 2.4.2), and contact languages (Section 2.4.3).

### 2.4.1 Flagging in language contact

Studies 5 (Hölzl 2024j), 6 (Hölzl 2023a)

Flagging is a cover term used in typological research that lumps case marking and adpositions together into one category (e.g., Haspelmath 2019; Croft 2022). While the label has clear limitations (e.g., obscuring language-internal distinctions between different layers of flagging), it can function as a useful concept for certain purposes. Two rather different kinds of developments of flagging are addressed here. Study 6 (Hölzl 2023a) investigates flagging (and other nominal categories, such as possession) in the Manchuric branch of Tungusic located in Northeast China that was in strong contact with several languages of the area, including Sinitic, Khitano-Mongolic, or Koreanic. Manchuric is contrasted with other Tungusic languages, such as Udihe, to illustrate the massive reduction in complexity as well as the restructuring of nominal morphology of Manchuric languages. Manchuric not only lost several case markers, although a few are retained in complex postpositions, but also developed a few new ones from adpositions. In addition, the Manchuric case markers underwent debonding (a form of degrammaticalization) from nominal suffixes to phrasal enclitics (similar to the English genitive *'s*). These changes are part of a larger reduction in complexity and restructuring of the Manchuric branch that is further addressed in Section 2.4.3.

As opposed to this rather rare development, study 5 (Hölzl 2024j) addresses about a dozen Sinitic languages in contact with Tibetic, Khitano-Mongolic, and Turkic that all developed OV order as well as postnominal flags. Most of these languages are located around a fault line between two Eurasian macro areas that exhibit OV and VO word order, respectively (see Section 2.2). As opposed to the simplification (i.e., loss of flags) in Manchuric, these Sinitic languages gained in complexity in either developing flags internally or in borrowing them from the neighboring languages. Some of these languages are further addressed in Section 2.4.3 as well.

These two developments suggest that very different types of language contact must have been at play in the history of these languages (study 5 Hölzl 2024j: 32). In the first case, the restructuring and reduction in complexity appear to be the result of uncontrolled second language acquisition and perhaps partial creolization. In the second case, where no simplification is observed, the driving factor appears to be longstanding bilingualism instead.

### 2.4.2 Shared developments

Study 7 (Hölzl 2018b)

Language contact can not only lead to the simplification of languages or to the replication of matter or pattern (e.g., Matras 2020: 256), examples of which are found throughout several studies in this thesis, but also to shared developments. An example is the development of standard negation from a negative existential mentioned in Section 2.3.1. Whereas study 1 (Hölzl 2015) discusses negation in all Tungusic languages, study 7 (Hölzl 2018b) focuses on the aberrant pattern in Manchuric, which is influenced by Mongolic languages. It shows that Mongolic and Manchuric languages not only share a common constructional network for negation, but also have analogous diachronic changes and developments of new constructions by means of the Croft cycle (aka negative existential cycle). Based on the terms *shared grammaticalization* and *grammaticalization areas* used in the literature, the names *shared constructionalization* and *constructionalization areas* are proposed as cover terms for these phenomena.

### 2.4.3 Contact languages

In more extreme cases, language contact can lead to the emergence of new languages. Such *contact languages* as "speech forms arising out of language contact" (Thomason 1997: 71) are usually classified into three different types: pidgin languages, creole languages, and so-called (bilingual) mixed languages (e.g., Matras 2020: 297-332; Hölzl 2024d). These can roughly be defined as follows:

- pidgin: "structurally limited language that is the result of strong and uncontrolled contact between speakers of mutually unintelligible languages and which itself has no native speakers" (Hölzl 2024h)
- creole: "structurally simplified and partly mixed language that is the result of strong and uncontrolled contact between speakers of mutually unintelligible languages" (Hölzl 2024e); can develop out of pidgins
- mixed language: "language that in its essential components consists of languages that are not closely related and therefore cannot be clearly assigned to any single language family" (Hölzl 2024f)

Within EA, there are only few pidgins, including Chinese Pidgin English and Chinese Pidgin Russian. Creole languages are by and large restricted to Spanish-based languages in the Philippines in ISEA, such as Cavite Chabacano. Mixed languages in EA include Mednyj or Copper Island Aleut, a mixture of Aleut and Russian (see the references in Hölzl 2018a), as well as Eynu in Xinjiang, which is a (mixed) secret language or cryptolect of Uyghur that consists of an Uyghur grammar with a lexicon that derives from Persian and some other languages (see Hölzl 2021c and the references therein). However, this traditional classification of contact languages exhibits some general problems, such as graded categories (McWhorter 2007), some of which are briefly addressed in the following.

**Manchuric as a partly creolized language** studies 1 (Hölzl 2015: 151), 6 (Hölzl 2023a: 183) and various others

First, there is a disputed category of **semi-creoles** including languages like Afrikaans that "exhibit both structural reduction and mixture to a degree intermediate between full transmission and creolization" (McWhorter 2007: 257). Additionally, they may exhibit some restructuring but less so than true creole languages. As such, a semi-creole can be characterized as "a transitional language located on a continuum somewhere between creole and non-creole" (Markey 1982: 204). In East Asia, an example could be Yilan Creole on Taiwan, a Japanese-based language that emerged in contact with Atayal, the classification of which as true creole is disputed (e.g., Tan 2023; Chien 2024). Among other things, creole languages often exhibit VO word order (e.g., Huber & the APiCS Consortium 2013) while Yilan Creole preserves Japanese OV order (Hölzl 2024e). Another candidate for a semi-creole in EA, is the Manchuric branch of Tungusic, which differs strongly from the rest of the language family but does not qualify as a true creole language. For instance, it also preserves Tungusic OV order among many other features.

However, it has often been noted that Manchu—here referring to Manchuric in general—deviates strongly from the other Tungusic languages, hence the alternative designation of the language family as "Manchu-Tungusic". Previously, two competing scenarios have been proposed as an explanation for these aberrant features (e.g., study 8 Hölzl & Payne 2022a: 6-7). Either Manchuric represents the first branch of Tungusic and therefore retains archaic features not found elsewhere (e.g., Kazama 2003) or Manchuric underwent massive changes due to language contact (e.g., Vovin 2006). While there are some unique sound changes, such as PT *mg > Proto-Manchuric *ŋg (study 9 Hölzl 2022b), these are not necessarily indicative for an early branching. Since Hölzl (2012) I have argued in favor of the second hypothesis (see Hölzl 2023b for the most recent overview). In this, I follow Benzing (1953, 1956) , who compared Manchu to English and its position within Germanic. I have extended this idea and have proposed that Manchu could be a semi-creole comparable to Afrikaans (studies 1 Hölzl 2015: 151, 6 Hölzl 2023a: 183).

There are countless differences between Manchuric and the rest of Tungusic in both lexicon and grammar that might justify such a classification (e.g., Tsumagari 1997). For instance, Manchuric has an entirely different system of question marking (Hölzl 2018a). Some examples discussed in this habilitation thesis include drastic differences in negation (studies 1 Hölzl 2015, 7 Hölzl 2018b), a different deictic day name system (study 2 Hölzl 2025a), an innovative interrogative (word) system (study 3 Hölzl 2024p), a unique multifunctional speech act verb *se-* 'to say' used as quotative and complementizer (study 4 Hölzl 2022a: 128-129), major differences in the nominal morphology (study 6 Hölzl 2023a), and otherwise unknown lexical items or pronouns (study 11 Hölzl 2021e: passim). While the list could easily be extended, it is noteworthy that Manchuric often preserves traces of Proto-Tungusic features that were only lost, simplified, or restructured at a later date. For instance, study 6 shows that the nominal morphology exhibits several unusual developments, such as a reduction in complexity (although some forms are preserved as relics) and a true restructuring of the system, including the debonding of case markers and the shift from head to dependent marking in adnominal possession (Hölzl 2023a).

It would be tempting to lump all these and other differences together, but the classification as a semi-creole is only justified if the changes involved happened during a relatively short period and did not accumulate over time. Furthermore, if certain differences are restricted to a subset of Manchuric, they cannot be taken as evidence for a partial creolization that must have occurred before the split into different languages. An example is the labiodentalization addressed in Section 2.5 that was claimed to be defining for Manchuric but in actuality is restricted to certain varieties. Another example is the existence of numeral classifiers in some varieties (Hölzl & Cathcart 2019).

(4) a. spoken Sibe
ilan **gian** [am bo]
three CLF big room
'three big rooms' (Jin 1993: 73)

b. written Sibe
emu **da** [toro hailan]
one CLF peach tree
'a peach tree' (SUISJ 1984: 3)

The first is a borrowing from Chinese *jiān* 间 whereas the second is derived from the noun *da* 'root'. However, these are a relatively recent innovation mostly found among the Sibe. Future studies, therefore, should pay more attention to the absolute and relative chronology of these changes (as argued in study 6 Hölzl 2023a). The research also has to take into account formerly neglected and recently extinct varieties that sometimes preserve archaisms or lack certain innovative features (Section 2.5). For instance, one obvious innovation in Manchuric that is found

in all varieties is the plain finite verb form that also serves as the citation form (*-mbi* in written Manchu, study 11 Hölzl 2021e: 181). This indicates an early simplification and restructuring of the verbal morphology.

The changes may have taken place due to large-scale adult second language acquisition (e.g., McWhorter 2007) by speakers of Khitano-Mongolic, Koreanic, Sinitic, other Tungusic, and perhaps unknown languages. However, the changes clearly were already in place before the advent of the Manchu Qing dynasty (1636-1912) and perhaps even the Jurchen Jin dynasty (1115-1234). The major obstacle for the hypothesis, therefore, is the lack of linguistic evidence for the period when the partial creolization may have happened. The exact dating and explanation thus remains a major task for future studies that will also have to include historical, archaeological, and genetic evidence.

**Kilen as a monophyletic mixed language** Study 8 (Hölzl & Payne 2022a: 3-4) and various others

Second, there are languages that exhibit a mixture but do not fall under the traditional label of mixed language. An example in East Asia discussed by Sandman (2016, 2021) is Wutun that, by and large, consists of elements from Sinitic or Amdo Tibetan and less so from the Mongolic language Bonan. Wutun is therefore not a true mixed language but clearly part of Trans-Himalayan that both Sinitic and Tibetic belong to (cf. Mazzoli & Sippola 2021: 21). Similar uncertainties exist for Daohua (Sinitic + Tibetic) or Baic (study 5 Hölzl 2024j) and perhaps an unrecorded language called *Yiyu Daozhuang*, allegedly a mixture of Yi and Lujia (study 14 Hölzl & Hölzl 2025).

Because of problems such as this, in Hölzl (2024f) I tentatively propose a bifurcation between two basic types of mixed languages that, in analogy to biology, could be termed *polyphyletic* and *monophyletic*. The former are truly mixed languages based on contact between unrelated languages that exhibit a relatively clear-cut distinction between the two (or more) sources as defined above. There exist different subtypes but their main characteristic is that they cannot clearly be assigned to one single language family due to the mixture (see Bakker 2017). The latter are not usually recognized as mixed languages because — to different degrees — they consist of components of more or less closely related languages and usually exhibit a less clear-cut distinction as in the case of Wutun. Note that here, too, the mixture must go beyond isolated cases of borrowing but must affect large chunks of the linguistic system so that they cannot clearly be assigned to a single branch of the language family. What both types of mixed languages have in common is an **unresolved genetic classification**, in terms of families in the case of the true (polyphyletic) mixed languages and in terms of branches in the case of the monophyletic ones. Both types of languages are, therefore, a major problem for the traditional tree model and classification of languages.

Two Tungusic languages widely recognized to exhibit some sort of mixed character are Kilen and Kili (or Kur-Urmi Nanai) (e.g., Janhunen 2024b: 10). While the mixture of Kili has been discussed many times, especially by Gerhard Doerfer (e.g., Doerfer 1978 among others), there are relatively few studies on the classification of Kilen (e.g., Kazama 1998). Both languages are addressed in several studies in this habilitation thesis, but a focus generally lies on Kilen (see also Hölzl 2017a,b). Apart from some Mongolic elements, the sources of both Kili and Kilen are all Tungusic languages, i.e. there is no doubt about their classification as Tungusic. However, as summarized in study 8 (Hölzl & Payne 2022a: 3-4), previous studies strongly disagree about the classification of Kilen and Kili within Tungusic. They are, therefore, prime candidates for my concept of monophyletic mixed languages.

The mixture appears to be the result of longstanding multilingualism and the central location of Kilen at the confluence of the Amur, the Sungari, and the Ussuri where all four branches of Tungusic — Manchuric, Nanaic, Udegheic, and Ewenic — come together. In the case of Kilen, the mixture thus involves elements from all four branches. These must have entered the language at different times but the absolute and relative chronology is difficult to determine due to a lack of sources before ca. 1850. The following are selected examples discussed within this habilitation thesis. Many lexical items and grammatical properties, such as standard negation, can only be of (southern) Nanaic origin (studies 1 Hölzl 2015: 124-127; 5 Hölzl 2018b: 251). Based on sound laws, some words like *gərbi* for 'name' can only have an Ewenic source instead (study 4 Hölzl 2022a: 114). A large number of lexical and some grammatical elements, including the negative existential *antɕi*, can only have come from Udegheic (study 12 Hölzl 2018c: 114-115). Many lexical items (especially nouns) appear to be from a Manchuric source close to the written Manchu language. This also includes the entire numeral system (Hölzl 2017b: 109). However, there are also previously overlooked archaic Manchuric elements similar to Alchuka, such as the word *tiɔrɔ* 'day after tomorrow' (study 2 Hölzl 2025a: 206f.). One of the central mechanism for the mixing might be *relabeling*: "a process that consists in assigning a lexical entry a new label derived from a phonetic string drawn from another language" (Lefebvre 2015: 1). Given the massive amount of relabeled forms from various sources, applying the comparative method leads to incoherent results (studies 3 Hölzl 2024p: 227, 9 Hölzl 2022b: 199, Section 2.5.4). It is especially the extreme intensity of the mixture that makes a classification of the language problematic and thus warrants classifying Kilen as a monophyletic mixed language. For lack of a better alternative, both Kilen and Kili are nevertheless tentatively classified as "basically" (southern) Nanaic in the phylogeny of Tungusic in Section 2.5, but this mainly illustrates the fact that a rigid tree model might not be adequate for mixed languages in the first place.

## 2.5 Comparative linguistics of Tungusic languages

This section gives a very brief introduction to the Tungusic languages (Section 2.5.1), addresses the etymology of the name *Manchu* (Section 2.5.2), the study of recently extinct Tungusic (and other) languages (Section 2.5.3), and their relevance for the classification of Tungusic (Section 2.5.4). Typological approaches to Tungusic languages are included in Section 2.3. The role of language contact in the classification of several Tungusic languages is discussed in Section 2.4. The claimed existence of Manchu speakers in Southwest China is addressed in Section 2.6.

### 2.5.1 Tungusic languages: Introduction

Study 8 (Hölzl & Payne 2022a)

Tungusic languages constitute a small language family of approximately 20 languages and a large number of dialects distributed over vast areas in Northern Asia (e.g., Vovin, Alonso de la Fuente & Janhunen 2024; Hölzl 2024l). Tungusic languages can be classified into four main branches (viz Ewenic, Udegheic/Orochic, Nanaic, Jurchenic/Manchuric), the mutual relation of which is still unresolved and being discussed. Altogether 11 of the 14 studies included in this thesis have Tungusic languages as their main focus. The remaining three studies mention Tungusic languages as well (see Section 1). The studies included here include data from all Tungusic languages, including many dialects, but a focus usually was on the Manchuric branch. Study 8 (Hölzl & Payne 2022a) is an introduction to the language family, to its classification, and to the volume *Tungusic languages: Past and present* (Hölzl & Payne 2022b).

### 2.5.2 The name *Manchu*

Study 9 (Hölzl 2022b)

The Manchus were the most influential of all Tungusic speakers. They founded the Qing dynasty (1636-1912), conquered large parts of East Asia, and ruled for almost 300 years. Despite the historical importance of the speakers and their language, the name *manju* 'Manchu' that was officially introduced in 1635 remained unclear. In fact, the etymology of the name may be among the most controversial topics that has vexed the field for several hundred years. There are dozens of proposals, most of which are untenable on linguistic or other grounds (see the literature in study 9 Hölzl 2022b). Given the vastness of the proposals, even prominent scholars simply gave up on finding a solution. For instance, Crossley (1997: 211) simply claimed that "all is flimsy speculation". Elliott (2001: 71) concluded: "We may never have any answers as to what 'Manchu' really meant before 1635 and it may not matter much if we ever do". Nothing could be further from the truth, however. Not only is the meaning of the name of paramount importance for a proper understanding of the Manchus, but the etymology is also not as unclear as many have assumed. While there are countless proposals, the majority can be ruled out for being implausible on various grounds (e.g., phonology, semantics, word formation, etc.).

In fact, the only plausible solution from a linguistic and cultural perspective may be what I have termed the "riverside hypothesis" in study 9 (Hölzl 2022b). According to this hypothesis the name Manchu may have originated as a general term for 'stream' or the name of a river (such as the lower Amur or Sungari) on which the ancestors of the Manchus were settling. The etymology was advocated for by some of the foremost experts in the field (i.e., Cincius, Benzing, Ikegami, Norman). Nevertheless, there remained several open questions and problems, especially concerning the semantic and phonological details. While Stary, one of the most prominent Manchu scholars, was opposed to the hypothesis, his counterarguments are weak and can be shown to pose no problem for the hypothesis. Study 9 (Hölzl 2022b) shows that the etymology is plausible from multiple perspectives, including the (complex) phonological evolution, the semantic development (i.e., a metonymy), the existence of many parallels for such a naming practice among other Tungusic speakers, and the areal distribution of cognates among other Tungusic languages. Most likely, the semantic shift from denoting a river to the adjacent people took place in the collocation *manju gurun* 'Manchu people/state', which is not only very frequent but also attested very early.

### 2.5.3 Recently extinct languages

Studies 10 (Hölzl 2020a), 11 (Hölzl 2021e), 12 (Hölzl 2018c), 13 (Hölzl 2021a), 14 (Hölzl & Hölzl 2025)

Currently, many of the languages spoken on this planet — including most of Tungusic — are endangered and consequently there is intense research on language endangerment and language revitalization. However, scholars usually agree that countless languages will disappear within this century and that only very few, if any, can be saved from extinction. What seems to be missing but will become ever more important within this century is what can be referred to as *the study of recently extinct languages*. This proposed field of study addresses the following research questions, among others:

- When is a language truly extinct?

- Is the former speech community still around? Did they survive the decline of the language and culture? Is fieldwork in the area feasible?
- Are there any elderly semi-speakers left? How much information can still be gained from working with them?
- Are there any toponyms or loanwords in neighboring languages?
- Are there any recordings of the language? What format are they in (e.g., handwritten, audio recordings)?
- Are they accessible or digitalized? How old are they? What is the language of description?
- How reliable are the available records (e.g., typos, typographic issues, internal consistency)?
- How to differentiate between typos, unknown dialectal features, synchronic variation, and differences in transliteration?
- Is the author known? Did he or she publish anything on these languages that is considered valuable?
- Are there any notes by the same author that can help clarify the handwriting and identify typos?
- Is there independent confirmation of the records? Are they considered fakes by some?
- What do the records contain (word lists, sample sentences, texts)? How extensive are they?
- Is there other material available on this, on neighboring, or on related language?
- What is the format of notation (e.g., IPA, Chinese characters, Pinyin, Cyrillic)? Is a transliteration necessary?
- Were the notes edited before? Is there a research tradition that needs to be taken into account?
- How can they be published or made available to other researchers?

Many of these questions are already well known from work on fragmentary languages. What sets this field of study apart, however, is the recency of the decline and the possibility of finding semi-speakers or records that are in danger of being lost. Unlike for languages like Tocharian, Etruscan, or Gothic that have been extinct for hundreds or even thousands of years, this leads to a certain urgency to acquire as much information as possible before it is too late.

Most Tungusic languages are endangered or moribund and some are already extinct (e.g., Janhunen 2005; Vovin, Alonso de la Fuente & Janhunen 2024: passim). Consequently, Tungusic contains several recently extinct languages, such as the language Arman that was made accessible in Doerfer & Knüppel (2013). Several languages that went extinct in the latter half of the 20th century include Alchuka, Bala, Chinese Kyakala, or Hezhen. Available references for the former three have been summarized in study 8 (Hölzl & Payne 2022a: 10) and Zikmundová & Gao Wa (2024).

The potential relevance of these languages (and similar cases) cannot be overestimated. For instance, the traditional historical phonology of Tungusic assumes that Nanaic preserves Proto-Tungusic *p- as *p-* while it developed to *f-* in Manchuric. However, the three Manchuric varieties Alchuka, Bala, and Chinese Kyakala all have cases where the *p-* is preserved. In the Nanaic language Hezhen, on the contrary, the *p- developed to *f-* just like in Manchu. The significance of these properties are further discussed in Section 2.5.4. The following presents brief sketches of these four languages that are usually overlooked in comparative Tungusic linguistics. For recently extinct languages outside of Tungusic, including Lujia or Longjia, and their relevance see Section 2.6.

**Hezhen** various studies

Hezhen is an extinct Nanaic language. In many publications, the language is not clearly distinguished from Kilen (on which see Section 2.4). In China, *hèzhēn* 赫真 and *qílèēn* 奇勒恩 both are treated as dialects of the *hèzhé* 赫哲 language (e.g., An 1986: 79). Hezhen appears to have gone extinct earlier than Kilen, was described mainly in Chinese and, consequently, was not addressed in many comparative Tungusic studies.

Elements from Hezhen may be present in many early descriptions of southern Nanaic, such as Grube (1900), Lattimore (1933), or Jettmar (1937), but the elements are not easily identifiable. The clearest descriptions of Hezhen can be found in Ling (1934: 260-280), NDSSLD (1958: 74-83) (transcribed with Chinese characters), and An (1986: 79-86). In these publications, Hezhen is explicitly contrasted with Kilen, which allows a straightforward identification of the data.

Data from Hezhen are included in several studies in this habilitation thesis (studies 2 Hölzl 2025a: 208-209, 3 Hölzl 2024p: 212, 225-228, 9 Hölzl 2022b: 174, 179, 183, 192, 195, 199, 11 Hölzl 2021e: 183). However, a comprehensive analysis of the available data remains to be done. Preliminary evidence shows that Hezhen and Kilen, while in areal contact with each other, exhibit many differences. On a whole, Hezhen is a southern Nanaic language while Kilen is probably a monophyletic mixed language with a southern Nanaic component (study 8 Hölzl & Payne 2022a: 4). Both languages exhibit similarities to Kili, Ussuri Nanai, and Sikachi Alyan Nanai but the details remain to be worked out (see Section 2.5.4 for some examples).

**Alchuka** various studies

The language Alchuka is named after a place called *alcuka* or *alcuha* in Manchu and *ālèchǔkā* 阿勒楚喀 in Chinese that is located next to Harbin in Heilongjiang province, Northeastern China. Like the following two varieties, the language has mostly been recorded by Mu Yejun 穆晔骏 (1926–1988) and, for some reason, remained almost unknown in comparative Tungusic studies. Most of the available studies of the language are collected in Hölzl (2020b), Zikmundová & Gao Wa (2024), H. Wang (2025), and study 8 (Hölzl & Payne 2022a: 10). Data from the language are used throughout many of the studies in this habilitation thesis.

While concerns over the data's reliability may have caused some scholars to dismiss them, these concerns are not supported by evidence. There is in fact independent confirmation of some of the data from other sources. As pointed out in Section 2.4, Kilen contains clear loanwords from Alchuka or from a closely related language (some examples are discussed in study 2 Hölzl 2025a). Some elements, such as the numeral system, were independently recorded by other scholars (see Hölzl 2017b). Another example for independent evidence for the existence of the Alchuka language is the place name *Harbin* 哈尔滨, the capitol of Heilongjiang province, that must derive from Alchuka *halbin* 'flat' as recorded by Mu (1987: 23). Note that the cognate *halfiyan* 'flat' in Manchu exhibits a different phonological form, which rules it out as a source of the place name (see Hölzl 2024a for the relevant sound changes).

Alchuka contains important features for the comparative study of Tungusic, such as an initial consonant *k-* unknown from Manchu (study 3 Hölzl 2024p). The language also preserves a unique nonfinite verbal suffix *-ʐï* (< *-si) that is addressed in study 7 (Hölzl 2018b). This suffix can take case markers like other participles and is also used as a connegative form in combination with the prohibitive *ɔm(ə)* while Manchu uses the imperfective participle *-rA* instead (see also Hölzl & Hölzl 2019a: 117f.):

(5) a. Alchuka
**ɔm** wak'la-**ʐï**
PROH blame-FFNLV
'don't blame (me)' (Mu 1986: 10, shortened)

b. written Manchu
**ume** wakala-**ra**
PROH blame-FFNLV
'don't blame (me)' (constructed)

Since the language also exhibits loanwords from extinct Para-Mongolic varieties, it is very valuable for the study of these languages as well (Hölzl 2017b). More examples from the language are shown in Section 2.5.4 below.

**Bala** Studies 10 (Hölzl 2020a), 11 (Hölzl 2021e) and various others

This language has mostly been described by Mu Yejun as well, but not exclusively so. All currently known sources are listed in study 8 (Hölzl & Payne 2022a: 10). Study 10 (Hölzl 2020a) provides background information on the language, including the etymology of the name, the sociohistoric context, and the geographic location. The only known text from the language, although heavily influenced by Manchu, is analyzed in study 11 (Hölzl 2021e). Despite being Manchuric, the language appears to have some influence from Kilen, e.g. in the expression of standard negation (study 7 Hölzl 2018b: 251) or in using the word *gərbi* 'name', which ultimately stems from Ewenic (study 4 Hölzl 2022a: 114). Data from the language are also used in many other studies assembled here.

**Chinese Kyakala** Study 12 (Hölzl 2018c) and various others

Hölzl & Hölzl (2019a) coined the term *Chinese Kyakala* to differentiate this Manchuric language from an Udegheic variety with the same name. This latter language, recorded in P. Schmidt (1928a), is referred to as *Russian Kyakala* instead. Originally, the speakers of Chinese Kyakala probably also spoke a form of Udegheic before they shifted to a form of Manchuric (possibly close to Alchuka) and, eventually, to Northeastern Mandarin. The language appears to retain some substrate words, such as the name of the ocean spirit *taimu* 泰木 (Udihe *temu*, Nanai *temu*) unknown from Manchuric (study 8 Hölzl & Payne 2022a: 11). Like the previous two varieties, Chinese Kyakala was also mainly recorded by Mu Yejun. Apart from the references collected in study 8 (Hölzl & Payne 2022a: 10), Chinese Kyakala has also been briefly addressed by Bäcker (1988: 44-50, 242-243). Most of the available data are collected in study 12 (Hölzl 2018c). The only known text, again heavily influenced by Manchu, is analyzed and set into context in Hölzl & Hölzl (2019a), which was excluded from this habilitation for reasons of space.

### 2.5.4 Proto-Tungusic *p- and classification

Studies 1 (Hölzl 2015: 135), 9 (Hölzl 2022b: 185), 10 (Hölzl 2020a), 11 (Hölzl 2021e: 174), 12 (Hölzl 2018c: 115)

**Background** The reconstruction and classification of Tungusic languages has a long history (see Janhunen 2024a for the most recent summary). Nevertheless, there remain many open questions and matters of dispute, especially concerning the higher-level structure of the family (see study 8 Hölzl & Payne 2022a: 2-7 for an overview). A

topic that has been discussed many times and plays a crucial role for the classification of Tungusic languages is the development of the Proto-Tungusic aspirated bilabial plosive *p- [pʰ], especially in word-initial position.[11]

The sound is widely acknowledged to be of considerable importance for the comparative study of Tungusic languages. Whaley & Oskolskaya (2020: 86) go so far as to claim that "[t]he reflexes of PT [Proto-Tungusic] *p- may be the single most significant piece of evidence for the dendritic structure of the family". Consequently, its diachronic developments have been addressed several times by different researchers. In his foreword to the *Grundzüge einer tungusischen Sprachlehre* by Castrén (1856: VIII), Anton Schiefner already observed that Manchu initial *f-* often corresponds to Evenki *h-*. Both are in fact reflexes of PT *p-. As early as 1923, Schmidt described "a phonetic rule that the initial sound *x* in Tungusian, corresponds to *f* in Manchu, *p* in Goldi [Nanai]" (P. Schmidt 1923a: 6). He presents correct correspondences like the following: Evenki ***h**ömun* [həmun], Oroch ***x**ömu* [xəmu], Nanai ***p**ömu* [pʰəmu], Manchu ***f**emen* 'lips' (P. Schmidt 1923a: 19). More extensive comparisons can be found in Cincius (1949: 177), Benzing (1956: 11, 32-34), Doerfer (1978: 8, 10-12), or Georg (2004: 55).

Nevertheless, several points remain unresolved and important datapoints have been neglected. For example, according to Kazama (2003: 7), the change from *p- > *f-* is restricted to, and thus defines, the Manchuric branch. However, not only do older stages (e.g., Old Jurchen) and several modern Manchuric varieties such as Bala preserve the original *p-, but southern Nanaic languages, such as Hezhen, also underwent the same sound change from *p- to *f-*. Janhunen (2024a: 44) recently claimed that "[t]he original strong labial stop *p is preserved as such only in Nanaic". However, as shown in several studies included here, this claim is problematic as evidence for the PT *p- is also preserved in parts of Manchuric until the second half of the 20th century (e.g., Alchuka, Bala, and perhaps Chinese Kyakala). Whaley & Oskolskaya (2020: 86) present two possible scenarios:

> As Georg (2004) argues, the simplest way to account for these facts is that a Southern Branch that includes Nanai maintained the PT **p-*, while in the Northern Branch **p- > *x.* [...] Alternatively, one could take **p- > f-* as a characteristic of the Southern Branch. The Northern Branch (including Nanai) maintains **p-*, and within the Northern Branch, Evenki/Even undergo **p > x.* Within the Evenki/Even languages, there is further *x > h > ∅.* Within the Nanai/Udihe languages, Nanai maintains **p-* and Udihe shows further weakening, i.e. **p- > *x-*.

While both scenarios are theoretically possible, the one proposed by Georg (2004) is much more likely. First, while parts of Manchuric preserve the *p-*, no Ewenic or Udegheic languages show such traces. This pattern is indicative for an older and defining change within all of Northern Tungusic while *p- > f-* in Manchuric and Nanaic is much more recent and not defining for the branches as a whole. In other words, both Manchuric and Nanaic must have originally retained the plosive while there is no evidence for this within Ewenic or Udegheic. In addition, the second scenario contradicts certain evidence for a closer relation between Ewenic and Udegheic on the one hand and Nanaic and Manchuric on the other (e.g., Georg 2004). Thus, the most likely development is the first scenario: Northern Tungusic (> Ewenic and Udegheic) underwent the change *p- > *x- while Southern Tungusic (> Nanaic, Manchuric) simply kept the *p-. The following discusses the development within the four main subbranches of Tungusic.

**Ewenic** Ewenic languages underwent a further change from *x- to *h-* (Table 5). In some languages, this is written <x>, but likely also pronounced [h] in most cases. This initial fricative was reduced to zero in several languages, including Oroqen, Solon, and parts of Khamnigan Evenki that are in areal contact. The same loss must have occurred independently in Arman that was not in close contact with these languages. Several forms with *p-* in Negidal and Samagir are clearly Nanaic borrowings that likely derive from Ulcha or a neighboring variety of Nanai. For instance, Negidal ***p**ikt**ə*** can only have a Nanaic source. The word does not exist in Manchuric and only northern Nanaic preserves both the *p-* and the consonant cluster *-kt-*. The Negidal word for 'window', recorded as ***p**ag**a*** by P. Schmidt (1923a), represents an interesting case. While the initial *p-* again indicates a Nanaic source, it is usually thought that only Ewenic languages regularly preserve an intervocalic *-g-* (e.g., study 8 Hölzl & Payne 2022a: 8-9; Hölzl 2025c). Unless the *-g-* is a secondary innovation in Negidal — which seems unlikely — the form could be from an old or archaic form of Manchuric that preserved both the initial *p-* as well as remnants of the intervocalic *-g-* (see study 11 Hölzl 2021e and Kane & Miyake 2024: 92 for similar examples). However, this seems somewhat less likely on geographical grounds than a Nanaic source.

**Udegheic** Udegheic languages all retained the Northern Tungusic *x- as *x-* (Table 6). As observed by Benzing (1956: 981), Udihe seems to have a few cases in which the initial changed to an *s-* before *i.* For instance, the word for 'child' has the variable form ***x**ita/e* ~ ***s**ita* in Udihe (e.g., Nikolaeva & Tolskaya 2001; study 6 Hölzl 2023a). Similar to Ewenic, there are a few loanwords with an initial *p-*, such as Udihe ***p**a:*, that are most likely from Nanaic. Here the regular loss of the intervocalic *-g-* led to the emergence of a long vowel (or to a diphthong in other cases).

**Nanaic** Nanaic languages preserved the Southern Tungusic *p-. However, the southern Nanaic languages underwent a later development *p- > f-* as in several Manchuric languages. The shared development could be due to

[11] The development in intervocalic position is slightly different (e.g., in Udihe) and will be neglected here (e.g., Benzing 1956: 981).

| Language | 'lip' | 'liver' | 'child' | 'window' | Source |
|---|---|---|---|---|---|
| Arman | _em | _ā-k | _ụt | - | Doerfer & Knüppel 2013 |
| Even, Eastern | həmən | hā-kan | hut | - | Kim 2011 |
| Even, written | hemen | hā-kan | hụt | - | Doerfer, Hesche & Scheinhardt 1980 |
| Evenki, Aoluguya | xəmun | xaa-kin | xutə | - | Tsumagari 1992 |
| Evenki, Southern | hemun | ha-kin | hute | - | Nedjalkov 1997 |
| Evenki, Sakhalin | ? | ha-kin | (h)ute | - | Bulatova & Cotrozzi 2004 |
| Negidal (Upper) | xəmun | xā-kin | (piktə) | (paɣa) | Cincius 1982 |
| Samagir | ? | ? | xutö | (pava) | P. Schmidt 1928b |
| Evenki, Khamnigan | _əmun | haa-kin | _utə | - | Chaoke 2017 |
| Oroqen, Nanmu | - | _aa-hin | _ut | - | Chaoke 2007 |
| Solon, Huihe | _omoŋ | _aa-hiŋ | _ʉt | - | Chaoke 2017 |
| Solon, Ongkor | _omu | _ā-kin | _ut(e) | - | Kałużyński 1971a,b |

Table 5: Ewenic correspondences for PT *p-

| Language | 'lip' | 'liver' | 'child' | 'window' | Source |
|---|---|---|---|---|---|
| Oroch | xəmu(n) | xa-ki(n) | xitə | (pāva) | Avrorin & Lebedeva 1978 |
| Russian Kyakala | xömu | ? | sitö(-mi) | ? | P. Schmidt 1928a |
| Udihe | xeme | x'e-i | xita/e, sita | (pa:) | Nikolaeva & Tolskaya 2001 |

Table 6: Udegheic correspondences for PT *p-

Manchuric (especially Manchu) influence on these languages. Note that some words, such as Kilen *fa* 'window', could either be an autochthonous development or a borrowing from Manchu *fa*. However, for other words, such as Kili *fute* 'child', an explanation in terms of borrowing is ruled out since Manchuric does not have a cognate (Manchu *jui* 'child'). Several Nanaic languages additionally have words starting with velar, uvular, or glottal fricatives that can be identified as borrowings from Northern Tungusic. For instance, words like Kilen *χa-kin* must have been borrowed given the vowel *i* in the suffix that is specific to Northern Tungusic while the expected vowel in Southern Tungusic is *u* (study 9 Hölzl 2022b: 184f.). Sikachi Alyan Nanai, being located at the border of the isogloss, appears to have forms with both *f-* and *p-*, e.g. *patalmeni* ~ *fatalmeni* 'falling star' (Poniatowski 1923: 8).

The absolute chronology of the change is more difficult to ascertain because it is not attested historically. We know from the earliest sources from around 1850 that Nanaic already showed variation between *p-* and *f-*, e.g. *pẹmu* ~ *femu* 'lip, mouth' (Grube 1900).

| Language | 'lip' | 'liver' | 'child' | 'window' | Source |
|---|---|---|---|---|---|
| Ulcha | pömu, pimu | pa | piktö | pāva | P. Schmidt 1923b |
| Uilta | pəmu | paa-xa | puttə | paawa | Ikegami 1997 |
| Samar | pömu | (xā-ki) | piktö | pāva | P. Schmidt 1928b |
| Nanai, Najkhin | pəmun | pa | piktə | pāwa | Ko & Yurn 2011 |
| Nanai, Sikachi Alyan | ? | ? | pikte | ? | Poniatowski 1923 |
| Kili | femu$^{n}$ | (ha-ki$^{n}$) | fute | fā | Sunik 1958 |
| Nanai, Ussuri | femu(n-) | (ha-ki(n-)) | f'ikte | fā | Sem 1976 |
| Hezhen | fəmun | ? | ? | ? | An 1986 |
| Kilen | (xəmun) | (χa-kin) | (xitə) | fa | An 1986 |

Table 7: Nanaic correspondences for PT *p-

**Manchuric** Concerning the development of *p-, Manchuric is perhaps the most complex Tungusic branch. Like Nanaic, Proto-Manchuric must have preserved the original *p-, which speaks in favor of a common retention in these southern Tungusic subbranches. According to the traditional view, the *p-* was retained in (Old) Jurchen as recorded during Jin dynasty (1115-1234) but subsequently changed to *f-* in two varieties of Jurchen (A and B) during Ming dynasty (1368-1644) (Kiyose 2000: 179; Kane & Miyake 2024: 92). The exact dating of the change remains uncertain, however, and is in need of further research in the light of the previously neglected varieties described above (e.g., Hölzl 2017b: 101): Alchuka (Table 8), Bala (study 11 Hölzl 2021e: 173-174), and perhaps Chinese Kyakala (Hölzl & Hölzl 2019a) retain cases of the original *p- (sometimes written <p'> [p$^{h}$]). In all three cases, however, there are also instances of an initial *f-* found in borrowings from varieties of Manchu.

Written Manchu and its modern descendant, written Sibe, both have *f-*. The same is true for all modern spoken

| Alchuka | Manchu | Meaning |
|---|---|---|
| p‘alia | fara | sleigh, … |
| p‘ədɔhɔ | fodoho | willow |
| p‘əili | fuli | jerky, dried meat or fish |
| p‘iniəgə | funiyehe | hair |
| p‘ulun | furun | meat scraped from a bone |
| p‘ut‘ia- | fucihiya- | to cough |

Table 8: The preservation of reflexes of *p- in Alchuka (see Mu 1987: 7, 9-11; Hölzl 2020b)

varieties in Manchuria, e.g. Aihui Manchu (Q. Wang 2005), Sanjiazi Manchu (Enhebatu 1995), Yibuqi Manchu (Zhao 1989), and Dzungaria, i.e. spoken Sibe (Zikmundová 2013). One variety formerly spoken in Qitamuzhen 其塔木镇, the classification of which remains somehow uncertain due to some idiosyncratic features, also exhibits evidence for *f-* (Hölzl & Hölzl 2019a: 132-134). Another mysterious variety recorded somewhere in Heilongjiang province has evidence for an initial affricate *pf-* (<pf‘> [pf$^{h}$]), which is unique among the Tungusic languages (e.g., Ma & Wulaxichun 1993: 5; study 11 Hölzl 2021e: 174).

**Relative chronology and distribution** Figure 1 summarizes the geographical distribution of the changes affecting PT *p-. Their relative chronology is shown in Figure 2. Note that the tree contains several uncertainties as the internal classification of Ewenic, Udegheic, Nanaic, and Manchuric is not entirely clear, which is shown with question marks (see study 8 Hölzl & Payne 2022a; Janhunen 2024b for some discussion and references). Future studies will have to revise certain details. Note that the overall plausibility of these changes is confirmed by similar cases of lenition observed in other languages (e.g., Bybee 2015: 29; Hölzl 2024b, 2025b,c and references therein).

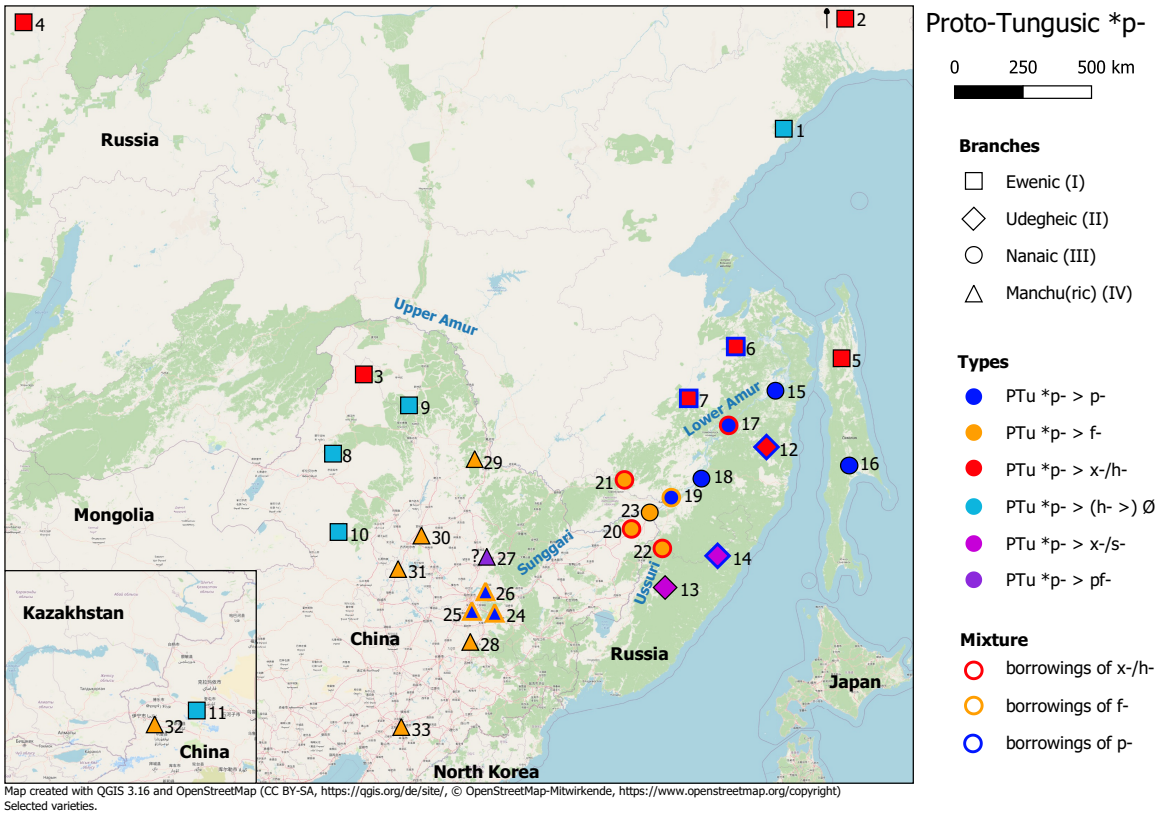


Figure 1: Distribution of PT *p- (simplified), numbers refer to Figure 2

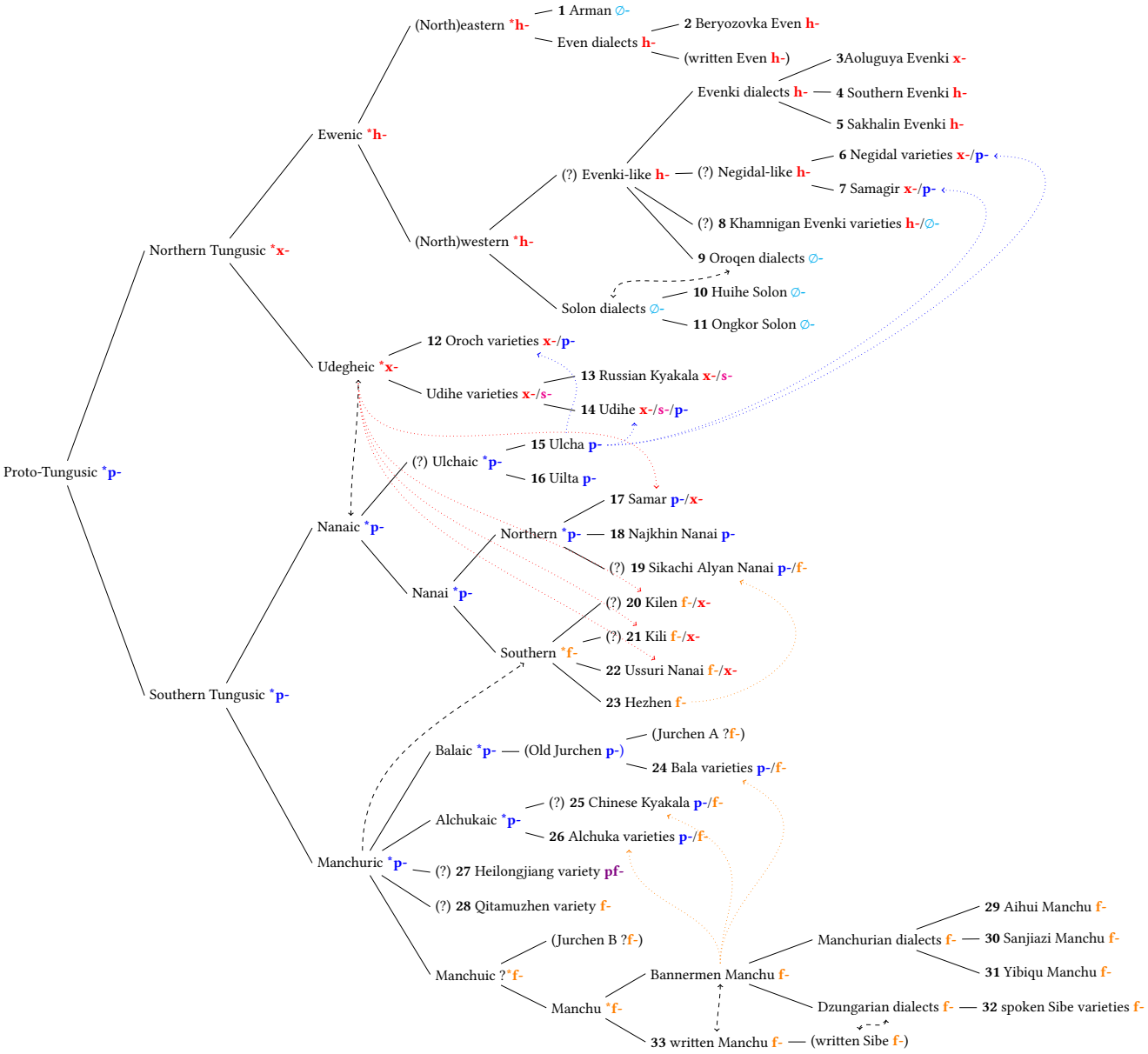


Figure 2: The development of PT *p- in the Tungusic phylogeny (simplified); numbers refer to the map (Figure 1); attested varieties not shown on the map are in parentheses; dashed arrows schematically indicate strong cases of areal convergence or contact; dotted arrows schematically indicate cases of borrowing; question marks show uncertainties in classification and type; only a small selection of the dialectal diversity is shown

## 2.6 Comparative linguistics of the Ta-Li languages

This section discusses three languages in Guizhou province in Southwest China that probably belong to the Sinitic branch of Trans-Himalayan. These languages, referred to as *Ta-Li* languages here, include Caijia, Longjia, and Lujia (or Luren A). Two other languages, here named Nanjing (a form of Yi) and Luren B (possibly an isolate), do not belong to the Ta-Li languages but are closely connected to them areally and historically and will briefly be discussed together with Longjia, and Lujia, respectively. Since the 1980s, the former speakers of the Luren languages are officially — but incorrectly — classified as Manchu minority (study 14 Hölzl & Hölzl 2025), which was the original impetus for working with these languages (see Hölzl & Hölzl 2019a: 135; Hölzl & Hölzl 2019b).

The following presents an introduction to the Ta-Li languages (Section 2.6.1) before addressing available materials for the three languages in turn (Section 2.6.2). The last part is an attempt at a rough typological classification of the languages (Section 2.6.3).

### 2.6.1 Ta-Li languages: Introduction

**Overview** The Ta-Li languages are located in western Guizhou province in Southwest China and include Longjia 龙家, Caijia 蔡家, and Lujia 卢家 (aka Lu, Luren). The name *Ta-Li* was coined by myself (study 13 Hölzl 2021a, Hölzl 2021d) and is now recognized by *Glottolog* where it has been given the Glottocode <tali1265>.[12] It is based on the observation that the three languages are likely genetically related and share the two characteristic words meaning 'two' and 'pig' (Table 9). The two words were chosen because they are, at least in this form, absent from surrounding languages and might preserve features of Old Chinese, such as the lateral (see below). An alternative designation in Chinese that does not, however, make reference to the Lujia language is Cai-Long 蔡龙 (e.g., CLLY 1950s: 1; GMSWSB 1982b: 43).[13]

| Language | Glottocode | Self-designation | 'two' | 'pig' | Study in this thesis |
|---|---|---|---|---|---|
| Ta-Li languages | <tali1265> | | | | |
| - Caijia | <caij1234> | meŋ$^{21}$ni$^{33}$ | ta$^{55}$ | li$^{21}$ | - (cf. Hölzl 2023c) |
| - Longjia | <long1417> | suŋ$^{55}$ȵi$^{55}$(mpau$^{31}$) | ta$^{31}$ | lɛ$^{55}$ | 13 (Hölzl 2021a) |
| - Lujia | <lure1234> | ? | ta$^{31}$ | li$^{31}$ | 14 (Hölzl & Hölzl 2025) |

Table 9: An overview of the Ta-Li languages

The three languages were formerly widespread but have only been recorded in few places (Figure 3). For Caijia, most of the data are from Hezhang and Weining, both located in the western part of the prefecture-level city of Bijie. For Longjia, the majority of the data are from Dafang county in central Bijie although records exist from various other places, including the prefecture-level city of Anshun further to the south. Lujia data stem from places in Central and Northeastern Bijie. See Section 2.6.2 for more details.

**Designations** The three groups do not share a common self-designation. The Caijia in Hezhang refer to themselves as *meŋ$^{21}$ni$^{33}$* (Lü 2022) and those in Weining as *mei$^{21}$ni$^{33}$* (Hsiu 2018). In the 1950s, a different self-designation *pa$^{31}$mi$^{31}$ȵi$^{53}$* was recorded for Weining (CLLY 1950s). The self-appellation of the Longjia in Dafang was recorded as *suŋ$^{55}$ȵi$^{55}$(mpau$^{31}$)*, the last element — seemingly referring to people — being optional (Zhang & Li 1982). The name was recorded as *suŋ$^{33}$ni$^{33}$* in Dafang in the 1950s (CLLY 1950s). In Jiangyizhai, this name was written in Chinese characters as *sōngníbǎo* 松尼保 next to an alternative name *xièlìbǎo* 谢利保 (Hsiu 2013), *lì* being pronounced with an [n] in Guizhou dialect. No reliable self-designation of the Lujia is known (cf. study 14 Hölzl & Hölzl 2025). While the etymology of these names is partly obscure, they all seem to contain a cognate of the classifier for people as a second component. This can also be found attached to other elements, e.g. in the expressions for 'who' (Caijia *lɔ$^{24}$ni$^{33}$*, Longjia *ti$^{33}$ȵi$^{31}$*), or on nouns if the numeral 'one' is elided (study 14 Hölzl & Hölzl 2025):

(6) Longjia

a. pu$^{55}$ ts'uŋ$^{13}$ [__ **ȵi$^{31}$**]
that person __ CLF
'that person' (Zhang & Li 1982: 7(2)[26])

b. pu$^{33}$ ts'uŋ$^{31}$ [i$^{55}$ **ȵi$^{55}$**]
that person one CLF
'that person' (Zhang & Li 1982: 11(2)[43])

(7) Caijia

mɔ$^{33}$ u$^{21}$ts$^{h}$o$^{24}$ [__ **ni$^{33}$**]
that person __ CLF
'that person' (Lü 2022: 41)

The classifier could in turn derive from the Old Chinese word *ni[ŋ] 'person' (> *rén* 人) (Baxter & Sagart 2014: 357). Note that tone sandhi and tone changes are not yet fully understood in both languages.

[12] The current version is Glottolog 5.2: https://glottolog.org/ (2025.06.03)

[13] For a vanished group known as Songjia 宋家 that might have been connected to the Ta-Li speakers, there do not seem to exist any recordings.

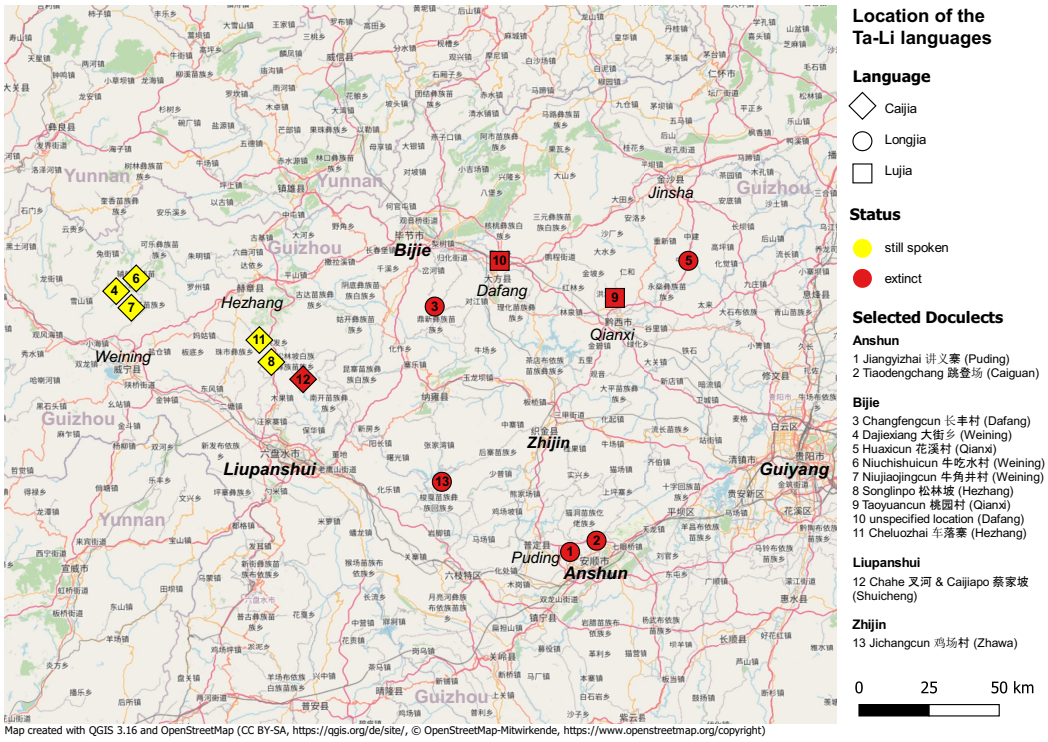


Figure 3: The location of the Ta-Li languages in western Guizhou (selected doculects: 1: Yang 1983, Hsiu 2013, 2: Wu 1989, 3: Zhang & Li 1982; Duan & Zhang 2008, 4/6: ELAR data, 5: Yang 1983, 7: Hsiu 2018, 8: Bo 2004, 9/10: Hölzl & Hölzl 2025, 11: Lü 2022, 12: Che et al. 2003, 13: Zhang 1997)

As opposed to the self-designations, the names given to the Caijia and Longjia in the Trans-Himalayan Yi language — formerly the most important language of the area — are more clearly related. The Yi referred to the Longjia (aka Nanjing) as $a^{33}vu^{33}t'u^{33}$ and to the Caijia as $a^{33}vu^{33}na^{33}$ (Ma & Ying 2001: 90). These names contain attributive adjectives, $t'u^{33}$ meaning 'white' and $na^{33}$ 'black' (the language has NAdj order). The more generic term without the adjective was also sometimes used, e.g. $a^{21}vu^{33}phu^{55}$ 'Nanjing people' in Dafang Yi (Zhai 2021: 29; Luo 2006 [1995]: 134), or $a^{31}vy^{55}po^{44}$ 'Nanjing people' in the Nanjing language (Duan & Zhang 2008: 3043). As in Longjia, the final element appears to be a generic term to refer to people that also occurs in the self-designation of the Yi in Dafang as $ni^{33}su^{33}phu^{55}$ (Zhai 2021: 29). As in this case, the Longjia are not always clearly differentiated from the Nanjing people and their exact difference and origin is a matter of ongoing research.[14]

**Research history and genetic classification** Unlike comparative Tungusic linguistics (Section 2.5), which has a long history going back hundreds of years, a comparative study of the Ta-Li languages does not yet exist. While the three groups have been known for a long time from traditional proto-ethnographic sources (e.g., Wu [Hölzl] 2022: passim, see also study 13 Hölzl 2021a for examples on the Longjia), their languages have only been recorded in the 20th century. However, that the three languages are genetically related has already been suspected since the 1950s based on superficial and mostly lexical comparisons.

The first serious investigation of the three languages was conducted in the **1950s**. But the materials collected for the three languages were never properly analyzed or published. They remain inaccessible for the general public or were lost. A copy of what appears to be a handwritten original of comparative data collected during the 1950s was discovered in 2024 in a bookshop and is now in my personal possession (CLLY 1950s; see study 14 Hölzl & Hölzl 2025 for the dating).

The booklet consist of 21 handwritten pages that contain a brief introduction (pages 1-3), a description of the phonology of Caijia (pages 3-7) and Longjia (pages 7-10), a comparison of clauses, phrases, and numerals in Caijia and Longjia (pages 11-13), a comparative word list for all three languages (pages 13-20), as well as a brief conclusion (page 21). Most of the data are from Caijia and Longjia, Lujia being represented by a few words exclusively (on which see study 14 Hölzl & Hölzl 2025). The first page is shown in Figure 4. The introduction contains a few notes on the classification:

蔡家、龙家、芦人（六喀子）现在和过去曾经使用过他们的相同的语言。这个语言不是汉语，不

[14]Names in other surrounding languages (some of which are summarized in study 13 Hölzl 2021a), require further research.

是彝语，不是布依语，不是苗语，不是仡佬语，也不属于藏缅语族中一支语言，不是洞 [侗] 傣语族中支语，不是苗瑶语族中支语。他们是一种独立的语言。这个独立语言暂定名为“蔡龙语”。(CLLY 1950s: 1)

The Caijia, Longjia, and Lu people (Liukazi), all in the past and present, have used a common language. This language is neither Chinese, Yi, Bouyei, Miao, nor Gelao, nor does it belong to any branch of the Tibeto-Burman language family, the Kam-Tai [Kra-Dai] language family, or the Miao-Yao [Hmong-Mien] language family. They are an independent language [family]. This independent language [family] is tentatively named “Cai-Long language [family].”

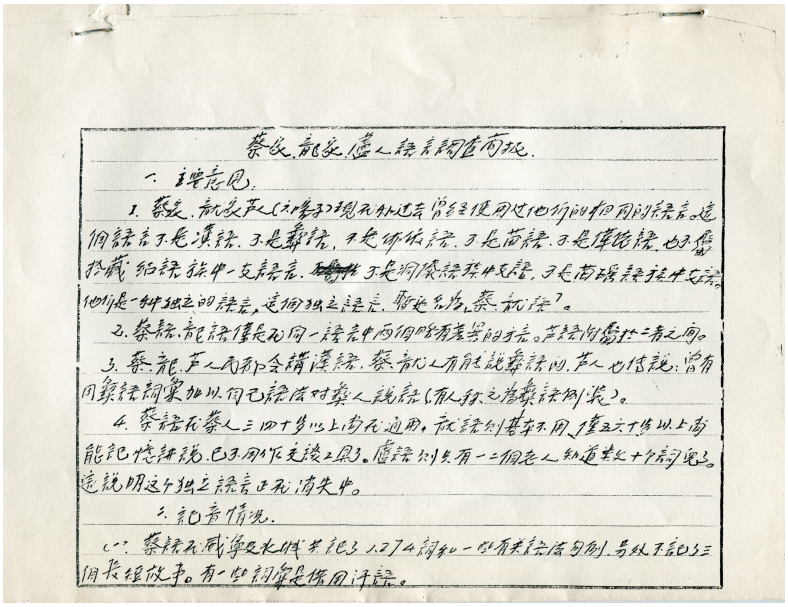

Figure 4: The first page of the comparative sketch from the 1950s (CLLY 1950s)

Although no conclusive evidence for a genetic relationship was presented, the general assessment that the three languages are related remains valid. However, rather than being one common language, the three languages form a branch of Trans-Himalayan that probably belongs to Sinitic.

The main problem for the comparative linguistics of the Ta-Li languages is the limited amount of materials for all three languages. This is due to the fact that already in the 1950s, the three languages were extinct (Lujia), moribund (Longjia), or endangered (Caijia):

> 蔡语在蔡人三四十岁以上尚在通用。龙语则基本不用，仅五六十岁以上尚能记忆讲说，已不用作交谈工具了。卢语则只有一二个老人知道数十个词儿了。这说明这个独立语言正在消失中。(CLLY 1950s: 1)
>
> The Cai[jia] language is still widely spoken among Caijia people over the age of 30-40. The Long[jia] language, however, is almost no longer used, with only those over 50-60 still able to remember and speak it, though it is no longer used as a tool for communication. The Lu[jia] language is known by only one or two elderly individuals, who remember just a few dozen words. This indicates that this independent language [family] is gradually disappearing.

Fieldwork conducted with Yadi Hölzl in 2019 confirmed that Lujia is extinct. In 2025, we confirmed the same for Longjia, although a few people in Changfeng village in Dafang still remember some isolated words or phrases that we were able to record. Andrew Hsiu, who attempted to locate speakers of Longjia in various places in western Guizhou over a decade ago (e.g., Hsiu 2013), also confirmed that the language must be extinct (p.c. 2025). Thus, the only language that is still spoken today is Caijia, but even this language must be considered endangered, or even moribund, by now (Lü 2022).

The first (semi-)published data on the comparison of the three languages, although still inaccessible to most scholars, dates from the **1980s**. They contain typographic problems as well as methodological shortcomings and are not very extensive (e.g., GMSWSB 1982b; GMSWSB 1982a). These studies furthermore exist in several slightly different versions, some of which are handwritten and largely inaccessible to the general public (see study 14 Hölzl & Hölzl 2025 for some examples). Additional publications of only Caijia or Longjia are discussed further below.

Despite the clear and generally correct results in the 1950s and 1980s that the three languages are genetically related, the speakers of the three languages were inexplicably classified separately during the 1980s. The Caijia today belong to a variety of different minorities, the Longjia are classified as Bai (see study 13 Hölzl 2021a), and the

Luren as Manchus (see study 14 Hölzl & Hölzl 2025). Needless to say, none of these classifications has any basis in fact. For instance, Lujia has no similarity whatsoever to Manchuric languages. While all three languages exhibit certain similarities to Baic languages, the exact relationship remains to be investigated. The issue is complicated by the fact that even the linguistic affiliation of Baic languages within Trans-Himalayan remains disputed (see the references in study 5 Hölzl 2024j).

Overall there are very few **recent** comparative studies of the Ta-Li languages. Most approaches limit themselves to discussing potential relationships of Caijia, the most accessible of the three languages, with either Baic (e.g., Zhengzhang 2010) or Waxiang Chinese (e.g., Wu & Shen 2010; Sagart 2011). However, neither comparison is very convincing and leaves the main issues unresolved.

The most sophisticated modern attempt at classifying Caijia can be found in Lee (2023, 2024). These studies address phonological aspects of Caijia (and Longjia) as well as Baic from a Sinitic perspective. They convincingly show that the most plausible external comparisons of the Ta-Li languages are in fact Baic and Sinitic languages. In addition, they strongly support the classification of Baic as Sinitic (see also study 5 Hölzl 2024j).

Study 14 (Hölzl & Hölzl 2025) is the first attempt at a bottom-up approach that aims at demonstrating a relationship between Caijia, Longjia, and Lujia. While phonological details still need to be worked out, it presents clear evidence for the affiliation of Lujia with Caijia and Longjia based on lexical, morphological, and syntactic correspondences. See also Section 2.6.3 below for typological resemblances between the languages.

Clearly, the Ta-Li languages exhibit a large amount of lexical material of Sinitic origin. Furthermore, they appear to preserve several archaisms not found in the majority of Sinitic. For instance, all three languages appear to preserve Old Chinese laterals similar to Waxiang Chinese (e.g., Lee 2023; study 14 Hölzl & Hölzl 2025). An example (absent from Caijia and unattested in Lujia) is OC *lraŋ (> *cháng* 肠) (Baxter & Sagart 2014: 109) and Longjia (*lɛ*$^{55}$)***l***a*ŋ*$^{55}$ 'intestine (of a pig?)' (Zhang & Li 1982: 25). A feature that makes Longjia even more important than either Caijia or Lujia is the presence of prenasalization (e.g., Hölzl 2021b for a preliminary overview). Table 10 shows that prenasalization in Longjia is phonemic.[15] Very likely, this feature can also be traced to Old Chinese. An example is OC *N-t<r>oŋʔ (> *zhòng* 重) (Baxter & Sagart 2014: 58, 118) and Longjia ***nʈ***u*ŋ*$^{31}$(*tɕe*$^{33}$) '(very) heavy' (Zhang & Li 1982: 143). In this case, prenasalization appears to go back to an OC prefix *N- marking intransitive or stative verbs. In other cases, prenasalization could correspond to other Old Chinese features, such as an uncertain loosely attached preinitial (indicated as *Cə- in Baxter & Sagart 2014: 186-191), e.g. OC *Cə.pə[r] (> *fēi* 飞) and Longjia ***mp***e*i*$^{55}$ 'to fly'.

| place of articulation | plain plosive | | prenasalized plosive | |
|---|---|---|---|---|
| bilabial | soldier | **p**ia*ŋ*$^{55}$ | wind | **mp**ia*ŋ*$^{55}$ |
| alveolar | to be located | **t**au$^{55}$ | bean | **nt**au$^{55}$ |
| retroflex | to dig | **ʈ**u$^{55}$ | tree | **nʈ**u$^{55}$ |
| velar | to sit | **k**au$^{55}$ | to ride | **ŋk**au$^{55}$ |

Table 10: Prenasalization as phonemic contrast in Dafang Longjia (Zhang & Li 1982: 3, 5, 12, 70, 167, 173, 176)

This feature makes Longjia special, if not unique, since it might be the only Sinitic language to preserve this kind of complex initial (cf. Caijia *pã*$^{33}$ 'wind', *kɯ*$^{21}$ 'to ride', *pɣ*$^{33}$ 'to fly'). This makes the language of utmost importance for the future study and reconstruction of Sinitic. Caijia only seems to preserve traces of this feature in the sporadic occurrence of voiced plosives (e.g., *ʈo*ŋ$^{33}$ ~ *ɖo*ŋ$^{33}$ 'heavy' or *to*$^{24}$ ~ *do*$^{24}$ 'bean', Lü 2022: passim), but the correspondences still need to be understood more clearly. In Longjia, the plosive is occasionally also assimilated to the preceding nasal and becomes voiced, e.g. ***mp***e*i*$^{55}$ ~ ***mb***e*i*$^{55}$ 'pus' (Zhang & Li 1982: 26).

However, more detailed external comparisons should perhaps only be attempted after the relationship of Caijia, Longjia, and Lujia is more fully understood. One of the most important issues for the moment is a proper description of the individual languages. For instance, there is still no dictionary for either Caijia or Longjia.[16] Given the limited resources for Lujia, the focus should be on these two languages. This first step should be followed by a clarification of which elements in both languages are inherited and borrowed from other languages. This can only be accomplished by a bottom-up comparison of the two languages that proceeds from lexical comparisons through phonological comparisons to sound laws and, if viable, reconstructions of Proto-Ta-Li (i.e., applying the comparative method, see Sections 2 and 2.5). After these initial steps, external comparisons with Sinitic and Baic will be on a stronger footing. The following gives a concise summary of the available materials for the three languages.

[15] In the suffix *(n)tsɿ*$^{33}$, prenasalization is not phonemic but triggered by a preceding nasal, e.g. *na*ŋ$^{55}$***nt***s*ɿ*$^{33}$ 'pocket, bag' (Zhang & Li 1982: 34). Anshun Longjia had an additional prenasalized alveopalatal plosive.

[16] A Caijia word list assembled by Lee Man Hei is available here: https://zenodo.org/records/7822357 (2025.06.01).

### 2.6.2 Available materials

**Longjia** Study 13 (Hölzl 2021a)

Longjia is an almost entirely undocumented and extinct language. Study 13 (Hölzl 2021a) is a preliminary study of Longjia that presents some background information, describes the data available at the time of writing, and summarizes previous studies of the language. It represents the first generally available publication on Longjia and the first publication in English. The main results for the investigation of Longjia will be published separately as a forthcoming grammar of Longjia.

Future study of Longjia is only possible through the rediscovery of extensive and reliable handwritten fieldnotes from the 1980s that are in my personal possession. These fieldnotes were collected by Zhang Jimin 张济民 and Li Juewei 李珏伟 in 1982 from one of the last speakers named Xie Yongxiu 谢永秀. This speaker was already 83 years old at the time of the fieldwork and was living in a village called Changfeng 长丰 in Dafang (see study 13 Hölzl 2021a: 25-30 for more details). We were recently able to visit her family that still lives in the village (Figure 5). Results from our fieldwork will be published together with all available materials in the grammar.

These fieldnotes and most previous studies of the Longjia language were described in study 13 (Hölzl 2021a: 25-30). However, since the publication, additional data have become available that are briefly summarized here (CLLY 1950s; Wu 1989: 116-119; Zhang 1997; Duan & Zhang 2008).

One source not yet mentioned are the recently discovered materials from the 1950s mentioned above (CLLY 1950s). These mostly represent the same dialect from Dafang (known as Dading at the time of writing) as the more extensive fieldnotes from the 1980s, and possibly some data from Puding. Prenasalized plosives here were usually noted as voiced plosives instead, e.g. ***g**au*$^{55}$ (CLLY 1950s: 11), ***ŋk**au*$^{55}$ 'to ride' (Zhang & Li 1982: 167) or ***b**iaŋ*$^{55}$ (CLLY 1950s: 15), ***mp**iaŋ*$^{55}$ 'mouth' (Zhang & Li 1982: 22). This is either a dialectal difference or represents a slight inaccuracy in the recording since prenasalization is clearly present in the fieldnotes from the 1980s, in independent records of Longjia (e.g., Wu 1989), and in one of our own recordings from Changfeng village made in 2025. However, prenasalization is occasionally absent in the fieldnotes from the 1980s as well, e.g. *mpei*$^{55}$ ~ *pei*$^{55}$ 'to fly' (Zhang & Li 1982: 155, 8(3)), although this could be a typo.

A more recent word list from Changfeng village in Dafang from 2004 was published in Duan & Zhang (2008: 3057-3059). These data were provided by Zhao Defu 赵德富, who was 81 years old at the time and has since passed away. The data are generally reliable and only contain minor mistakes (e.g., the words for 'to open' and 'to close' were confused), although they seem to represent a reduced state of the language that was no longer spoken and subject to attrition. Among other things, it seems to have lost features such as prenasalization. In 2025, we were able to locate the daughter of the informant, who remembers isolated phrases from the language (Figure 5).

Figure 5: The present author (mid front), Yadi Hölzl (right front), and inhabitants of Changfeng village with an old drum formerly used in rituals (photo taken by Wu Jun in Changfeng, 2025)

One unreliable record mentioned in Hölzl (2021a) is a brief description of Longjia in Caiguan (CGZZ 2004). It was recently discovered that these records are a copy of a somewhat more reliable recording of the language from Anshun that was published in Wu (1989: 116-119). Although this version also contains several typographic

problems, their number is much reduced compared to CGZZ (2004).

A word list from a previously unknown variety in Zhijin — geographically located between the recordings in Dafang in the North and from Anshun in the South (Figure 3) — was published in Zhang (1997: 161). This list is entirely written in Chinese characters that are in need of an additional interpretation before being usable. The language is mistakenly referred to as a form of Bai.

The only lexical item found in all extant records of the language is the verb 'to eat', which is also among the most characteristic items in Longjia not found in Lujia, Caijia, or any other language, for that matter (Table 11).

| Variety | Date | to eat | Source |
|---|---|---|---|
| Dading | 1920s | wǎ 瓦 | Zhao et al. 1985 [1926] |
| Dading | 1950s | wa$^{33}$ | CLLY 1950s |
| Dafang | 1982 | wa$^{31}$ | Zhang & Li 1982 |
| Dafang | 2004 | wa$^{31}$ | Duan & Zhang 2008: 3058 |
| Qianxi | 1980s | ua$^{31}$ | GMSWSB 1982b: 22 |
| Puding | 1980s | ua$^{31}$ | GMSWSB 1982b: 22 |
| Puding | ? | wá 娃 | Hsiu 2013 |
| Anshun | 1989 | wa$^{31}$ | Wu 1989: 117 |
| Zhijin | 1997 | wǎ 瓦 | Zhang 1997: 161 |

Table 11: The characteristic word for 'to eat' in different recordings of Longjia

Given that Dafang Longjia was recorded at different times, it is possible to establish several regular sound changes. For instance, from the 1950s to the 1980s the diphthong *ai* apparently changed to a monophthong *ɛ*: *ŋuai*$^{33}$ > *ŋuɛ*$^{33}$ 'month', or *ŋo*$^{31}$*-lai*$^{31}$ > *ŋo*$^{31}$*-nɛ*$^{31}$ 'we'. An example for a change from the 1980s to 2004 is from the final *-uŋ* to a nasalized vowel *-õ* (~ *o*): *ts'uŋ*$^{33}$ > *tshõ*$^{31}$ 'person', or *muŋ*$^{33}$*luŋ*$^{33}$ > *mo*$^{55}$*lõ*$^{55}$ 'insect, …'. These regular correspondences between independent recordings of the language also confirm their overall accuracy.

According to traditional accounts, the Longjia historically merged with another population that is usually referred to as **Nanjing** 南京 (study 13 Hölzl 2021a). Because of this, the Nan(jing) and Long(jia) are also sometimes referred to as Nan-Long (or Long-Nan), a name combining the initial syllables of both groups — a typical Chinese word formation strategy. The reliability of the traditional account about the origin of these people is questionable and requires further research.

Duan & Zhang (2008: 3043-3056) present data from a language referred to as Nanjing language (*nánjīng huà* 南京话) or Awupo language (*āwǔpō huà* 阿武泼话) that were still unknown to me during the writing of study 13 (Hölzl 2021a). Since the latter name also refers to the Longjia and Caijia (see Section 2.6.1), this language will be referred to here as *Nanjing* language to avoid confusion. The data were collected in 2004 from an 80 year old speaker (who allegedly did not use the language for 50 years) and the language may now be extinct, and currently, these remain the only known data. Unless further speakers are found, future research needs to be based on these published materials.

The language is claimed to be similar to the local Yi language but sufficiently distinct to be incomprehensible. The two languages share some typological features with Caijia and Longjia (e.g., tones) but are genetically and structurally distinct. For example, whereas Longjia is a language with basic VO order like Caijia, the Nanjing language, has OV order like Yi instead (cf. Dafang Yi *ʑi*$^{21}$ 'water' *ndo*$^{21}$ 'to drink', Zhai 2021: 186, 174).

(8) Longjia
**ʑuŋ**$^{55}$ ɕi$^{31}$
drink water
'to drink water' (Zhang & Li 1982: 157)

(9) Caijia
**ã**$^{21}$ sɿ$^{55}$
drink water
'to drink water' (Lü 2022: 38)

(10) Nanjing
ʑi$^{33}$ **duo**$^{33}$
water drink
'to drink water' (Duan & Zhang 2008: 3058)

However, both Nanjing and Dafang Yi contain a lexical overlap with Longjia which could be due to instances of borrowing between the languages. An example for this is Dafang Yi *ɣʊ*$^{55}$ (Zhai 2021: 185), Nanjing *hõ*$^{55}$ (Duan & Zhang 2008: 3045), and Longjia *hu*$^{55}$ (Zhang & Li 1982: 1), all of which mean 'vegetables'. Nanjing also appears to have certain unique characteristics not found in either Dafang Yi or Longjia. The exact nature of and relationship between the three languages can only be clarified with further comparative research that necessarily also includes surrounding languages and Caijia.

**Caijia** While the earliest data for Longjia are from the 1920s, the earliest documented description of Caijia is from the 1950s. Caijia is, therefore, not a newly discovered language, as is sometimes claimed. The original fieldnotes from the 1950s (some of which are known to exist) are currently unavailable. These were probably the basis of the Caijia data found in the handwritten record of Caijia, Longjia, and Lujia from the 1950s described above (CLLY 1950s). The first semi-published materials for Caijia were edited in the 1980s (GMSWSB 1982b; GMSWSB 1982a). Until relatively recently, the only widely available data were found in a brief sketch of Caijia by Bo (2004). An often overlooked short word list of Caijia from Liupanshui that was transcribed with Chinese characters can be found in Che et al. (2003). Over the last years a relatively large number of articles has been published that address individual features of the language and cannot be reviewed here individually (e.g., Lü 2017, 2018, 2019a,b; Hsiu 2018 among several others).[17] Without doubt, Caijia by now is the best described of the three Ta-Li languages. It is the only language for which a reference grammar is available (Lü 2022). This was recently reviewed in Majewicz (2023) and Hölzl (2023c). Since the language is still spoken (albeit endangered), fieldwork is still a possibility.

**Lujia** Study 14 (Hölzl & Hölzl 2025)

The least known of the Ta-Li languages is Lujia. Previously, this language has been referred to as *Lu* or *Luren* language. However, both of these names are somewhat problematic. Lu is perhaps too short to be useful and the latter part of Luren actually means 'person', which is why the term is more plausibly applied to the former speakers of the language. Additionally, there are probably two languages with this name (see below). To avoid these issues, study 14 (Hölzl & Hölzl 2025) employed the term Lujia instead, a name found in the comparative sketch from the 1950s (CLLY 1950s). Study 14 (Hölzl & Hölzl 2025) argues that the available data for Luren actually represent two and not one language. For lack of a better designation, these are preliminarily referred to as Luren A (or Lujia) and Luren B. The classification of the latter language that was recorded in Jinsha county remains unresolved and it could be a so-called language isolate.[18] Not much can currently be said about this language given the limited available data and lack of related languages. The language is typologically similar to the Ta-Li languages in having tones, numeral classifiers, etc. However, unlike Luren A (or Lujia), it exhibits final nasals and appears to have a larger inventory of tones. A few lexical items have been identified as loanwords.

All currently known studies and all reliable data on the two Luren languages are collected in study 14 Hölzl & Hölzl (2025). The main source are six word lists of which three represent Luren A and two Luren B. The sixth word list is the most extensive but also the most unreliable as it contains many typographic problems and mistakenly combines data from both varieties into one list. Three of the word lists are mistakenly referred to as containing data on the Manchu language.

### 2.6.3 Areal typology

The three Ta-Li languages all have typical properties of the Mainland Southeast Asian area (cf. Enfield & Comrie 2015b). These include, among others:

- comparatively large vowel systems: 11 in Hezhang Caijia (Lü 2022: 33), 8 in Dafang Longjia (cf. Zhang & Li 1982), and at least 7 in Lujia (study 14 Hölzl & Hölzl 2025), not including diphthongs
- many initial but few final consonants (tendency for open syllables): only *-n* and *-ŋ* in Hezhang Caijia (Lü 2022: 35) and Dafang Longjia from the 1980s (Zhang & Li 1982: passim), none in Lujia (study 14 Hölzl & Hölzl 2025) and Dafang Longjia as recorded in 2004 (Duan & Zhang 2008: 3057-3059)
- many monosyllabic words like Caijia *pɣ*$^{21}$ (Lü 2022), Longjia *mpei*$^{55}$ 'fat', albeit with exceptions like Caijia *məŋ*$^{21}$*nʊŋ*$^{21}$ (Lee 2023: 11), Longjia *muŋ*$^{33}$*luŋ*$^{33}$ 'insect, …' (Zhang & Li 1982)
- (complex) tone systems: 4 or 5 in Hezhang Caijia (Lü 2022: 36-37), 4 in Dafang Longjia (cf. Zhang & Li 1982), and at least 4 in Lujia (study 14 Hölzl & Hölzl 2025), not including sandhi phenomena that mostly remain unclear for now
- morphological tone changes that remain to be investigated in detail (Lee 2023: 87; Lü 2022: 43-44)
- "isolating" morphology and a general lack of inflection, with exceptions like plural marking in the personal pronouns
- numeral classifiers in all three languages (see study 14 Hölzl & Hölzl 2025)

In terms of syntax, Caijia and Longjia exhibit many similarities and some rare patterns whereas almost nothing can be said about Lujia. Caijia and Longjia have basic SVO word order (see 8, 9) but can have marked SOV order following the typical Sinitic or Baic *positional differential object marking* (Seržant et al. preprint) using a prenomial

[17] The Endangered Languages Archive (ELAR) contains materials on Caijia collected by Li Lan, Wang Jian, Lü Shanshan, Fu Kang, and Sun Shun: https://www.elararchive.org/dk0548/ (2025.06.06)

[18] If accurate, the stock it represents could be named *Jinshaic*.

flag as object marker derived from a verb meaning 'to take' (Caijia *a*$^{33}$, Longjia *a*$^{55}$, see Lü 2022 and study 14 Hölzl & Hölzl 2025). Caijia and Longjia exhibit at least two cross-linguistically rare word order patterns (Table 12). The first is VO in combination with prenominal relative clauses (RelN).

(11) Caijia
mɔ$^{21}$ [zɣ$^{21}$ zɣ$^{33}$ **sɪ$^{21}$**] u$^{33}$ts$^{h}$o$^{33}$
that eat meal REL person
'the person who is eating' (Lü 2022: 336)

(12) Longjia
[a$^{55}$ kɛ$^{33}$ **ku$^{31}$**] la$^{31}$wo$^{55}$
take chicken REL eagle
'an eagle taking a chicken' (Zhang & Li 1982: 5(3))

Languages with this type of combination are usually found in Southeast Asia (e.g., Sinitic, Baic, Hmong-Mien, some Austronesian languages on Taiwan). Unlike many of these languages, however, Caijia and Longjia do not usually employ the same marking for relative clauses as they use for adnominal possession. This makes the pattern even more unusual (Hölzl 2024k). Both languages have related possessive markers, e.g. Longjia *ŋu$^{31}$huŋ$^{31}$* 'my, mine' (Zhang & Li 1982: 5(2)), Caijia *ŋo$^{21}$hɱ$^{55}$* 'my, mine' (Lü 2022: 63) (recorded as *ɣuŋ$^{55}$* in Bo 2004: 76). And both languages can use a word for 'family' as genitive (Caijia *ŋo*$^{33}$ '1SG' + *hɛ*$^{21}$ 'family' > *ŋuɛ*$^{21}$).

(13) Caijia
**ŋuɛ$^{21}$** pa$^{33}$
my.family grandfather
'my grandfather' Lü 2022: 66)

(14) Longjia
[ŋu$^{31}$ **ɕi$^{55}$**] pa$^{55}$mpei$^{33}$
1SG family m.grandfather
'my grandfather' (Zhang & Li 1982: 4(2))

The second rare word order combination is Dem-Adj-N-Num(-Clf) (see Dryer 2018). Note that in the following Caijia example, the numeral 'one' is elided (cf. *k$^{h}$ui$^{55}$* [*ji$^{21}$ k$^{w}$ɔ$^{21}$*] 'dog one CLF', Lü 2022: 362).

(15) Caijia
mɔ$^{21}$ pia$^{21}$ k$^{h}$ui$^{55}$ [__ k$^{w}$ɔ$^{21}$]
that white dog __ CLF
'that (one) white dog' Lü 2022: 272)

(16) Longjia
tsi$^{33}$ ma$^{55}$ lɛ$^{55}$ [i$^{55}$ ku$^{33}$]
this small pig one CLF
'this one small pig' (Zhang & Li 1982: 11(2))

Possibly, this word order is cross-linguistically infrequent due to the combination of prenominal adjectives and postnominal numerals (AdjN & NNum), which appears to be dispreferred overall. In Dryer's global sample, while at least some other languages exhibit either RelN & VO or AdjN & NNum, only Bai shares both of these rare word order patterns (Dryer 2013a,b,c,d). This similarity could be either due to a common inheritance or due to the location in a very similar ecological setting (Hölzl 2022c).

There are further typological similarities to Baic languages. For instance, both Caijia and Longjia exhibit head-initial order for sexus markers.

(17) Caijia
kɪ$^{33}$ **bu$^{33}$**
chicken male
'rooster' (Bo 2004: 70)

(18) Longjia
kɛ$^{55}$ **pau$^{33}$**
chicken male
'rooster' (Zhang & Li 1982: 15)

The same is attested in Baic languages, e.g. Shalang Bai *kɛ*$^{34}$ ***po*$^{55}$** 'rooster' (F. Wang 2012: 78). But there are also certain differences. For instance, while in both Caijia and Longjia the demonstrative precedes the noun as shown above, it follows in Baic, e.g. Shalang Bai *khuã*$^{33}$ *kɯ*$^{31}$ *tɯ*$^{42}$ 'dog this CLF' (F. Wang 2012: 83).

Table 12 summarizes the word order properties of Caijia and Longjia discussed in this section. The strong typological similarities add to the evidence accumulated in study 14 (Hölzl & Hölzl 2025) that the two languages are genetically related and therefore share many phonological, lexical, and grammatical features. Future studies will have to include a larger number of features and attempt a more detailed comparison with Sinitic and Baic languages.

| Feature | Caijia | Longjia | Headedness |
|---|---|---|---|
| transitive clause | VO (OV) | VO (OV) | initial (final) |
| relative clause | RelN | RelN | final |
| adnominal possession | GenN | GenN | final |
| demonstrative | DemN | DemN | final |
| adjective | AdjN | AdjN | final |
| numeral & classifier | N(Num)Clf | N(Num)Clf | initial |
| sexus marker | NSex | NSex | initial |

Table 12: Selected word order features in Hezhang Caijia (Lü 2022) and Dafang Longjia (Zhang & Li 1982: passim)

# 3 Conclusion

What the studies assembled in this habilitation thesis have in common apart from addressing aspects of a (*General / Panchronic*) *Comparative Linguistics* that combines aspects of synchronic, diachronic, and other approaches to language structure (Section 2.1), East Asia (Section 2.2), and a general focus on Tungusic languages (Section 2.5), is a holistic point of view, aiming for a complete understanding of the respective areas. In sum, the most important aspects discussed in this thesis include:

- The development of a **systemic approach** to several domains of language, including the deictic day name system (with the topology of a line) and the interrogative system (with a complex topology and self-similar or fractal relationship to the rest of the language).
- One of the first scientific studies of the "**personal name question**" ('What's your name?') that could be among the few universal questions found in all or most languages. This also includes the establishment of the possibly universal "personal name frame" as underlying background knowledge for the question, although there are cultural and linguistic differences in the details.
- Ideas for an "**Areal Construction Grammar**" that addresses areal phenomena from a constructionist perspective (e.g., shared constructionalization, constructionalization areas). In the case of Manchu and Mongolian, this also includes a common constructional network (e.g., for negation).
- Ideas for a distinction between *polyphyletic* (or "true") mixed languages (e.g., Media Lengua) and *monophyletic* **mixed languages** (e.g., Kilen, Wutun) and a potential scale between them corresponding to the degree of relatedness. The proposed common denominator for the two categories is an unresolved genetic classification, in language families in the first case and in branches in the second.
- Advances in the classification of the **Tungusic** languages, including the hypothesis that Manchuric could be a contact language (more precisely a semi-creole), a comprehensive treatment of Proto-Tungusic *mg as well as *p-, and an analysis of several monophyletic mixed languages (especially Kilen).
- The confirmation of the **etymological origin** of the name *Manchu*, as having a likely source in Proto-Tungusic *mamgo 'large river, lower Amur'. Previously, countless other etymologies had been proposed, none of which was convincing on linguistic grounds. While the etymology addressed here had already been put forward by several of the leading experts in the field, many details of the phonological and semantic development remained uncertain and disputed.
- The development of foundational ideas for **a study of recently extinct languages**. While the significance of such a field of research was only briefly illustrated with the help of some Tungusic languages and languages of Guizhou, the general research questions can also be applied to other cases from around the world. The methodology might have to be adapted according to the individual cases. Given that a large number of currently endangered languages will disappear within this century, the issue will remain relevant and will perhaps even become more important as more and more languages vanish.
- Some of the first comprehensive studies in English of several extinct and **neglected Tungusic languages** (e.g., Alchuka, Bala, Chinese Kyakala) and a demonstration of their significance for the historical development of Tungusic languages, including both retentions and innovations.
- The rediscovery of extensive materials of Caijia, Lujia, and especially the extinct **Longjia** language and some preliminary results concerning their importance for the study of Sinitic languages (e.g., preservation of laterals in all three languages, prenasalization in Longjia) and the linguistic evolution of EA. More details concerning Longjia will be presented in a forthcoming monograph.
- The first publicly available study of Longjia in English and the first scientific study of the two **Luren** languages ("Luren A"/Lujia and "Luren B"), one of which could be a stock of its own. If accurate, this would be the first isolate ever discovered in Guizhou, represeting a previously unknown stock.
- A demonstration that the **official classification** of the Luren as Manchus in the 1980s is without scientific basis. While the study included here focuses on linguistic evidence, a forthcoming companion paper will demonstrate the same based on ethnographic, historical, and genetic data.
- One of the first comprehensive comparative studies of Caijia, Longjia, and Lujia and the coining of the term "**Ta-Li**" (Glottocode <tali1265>) as a name for the Sinitic subclade they seem to constitute.

In conclusion, this habilitation thesis attempts to contribute meaningfully to the study of East Asian languages and the broader field of (General) Comparative Linguistics. By addressing some new areas of research and making accessible several previously almost unknown languages, it offers a fresh perspective and lays the groundwork for further exploration into the linguistic diversity of East Asia and its historical development.

## Division of labor and corridenda to the studies

**Study 2** Corrigenda:

- "more unanalyzable" should be "more analyzable" (Hölzl 2025a: 196)

**Study 3** Division of labor:

The study was written, reviewed, and accepted around 2018. Due to the delay of the project for many years, I became an additional editor in June 2024 and completed the book in the beginning of July so that the volume could finally appear (Enke et al. 2024). The peer review process for my chapter was completed before I became a coeditor.

**Study 4** Division of labor:

This study appeared in a volume edited by myself and Thomas E. Payne (Hölzl & Payne 2022b). The peer review process for my chapter was organized by my coeditor.

**Study 8** Division of labor:

This is an introduction to the volume *Tungusic languages: Past and present* (Hölzl & Payne 2022b). The planning and write-up of the chapter was mostly done by myself, except for a few summaries of individual chapters published in the volume that were written by my coeditor Thomas E. Payne.

**Study 12** Division of labor:

This study appeared in a special issue of the journal *International Journal of Diachronic Linguistics and Linguistic Reconstruction* edited by myself and Peter-Arnold Mumm (Hölzl & Mumm 2018). The peer review process of my paper was organized by my coeditor and the journal editor.

Corrigenda:

- "some three hundred years" should be "some one hundred years" (p. 121)
- "Mu Yuejun" should be "Mu Yejun" (p. 127)
- "tent | maika | ?" should be "tent | maika | maiga" (p. 126)
- "not mentioned in Nikolaeva & Tolskaya (2001)" should be "also mentioned in Nikolaeva & Tolskaya (2001: 925)" (p. 128)

**Study 13** Corrigenda:

- "Zhang & Ji" should be "Zhang & Li" (correct in bibliography)
- "*mpao*$^{31}$" should be "*mpau*$^{31}$" (p. 13, 15)
- "Luezi" should be "Liuezi" (p. 30)

**Study 14** Division of labor:

The linguistic analysis and the write-up was mostly done by myself. My coauthor Yadi Hölzl helped with the data collection, the fieldwork, the conceptual structure of the paper, the background information, and the proofreading.

# References


Adam, Nina & Andreas Hölzl. 2024. Subjects in word order and alignment typology. In Artemis Alexiadou, Doreen Georgi, Fabian Heck, Gereon Müller & Florian Schäfer (eds.), *Gisbert Fanselow's contributions to syntactic theory* (Linguistische Arbeitsberichte 96), 311–322. Leipzig: Universität Leipzig.

Aikhenvald, Alexandra Y. & R. M. W. Dixon. 2001. Introduction. In Alexandra Y. Aikhenvald & R. M. W. Dixon (eds.), *Areal diffusion and genetic inheritance: Problems in comparative linguistics*, 1–26. Oxford: Oxford University Press.

Alexiadou, Artemis, Doreen Georgi, Fabian Heck, Gereon Müller & Florian Schäfer (eds.). 2024. *Gisbert Fanselow's contributions to syntactic theory* (Linguistische Arbeitsberichte 96). Leipzig: Universität Leipzig.

Allentoft, Morten E. et al. 2024. Population genomics of post-glacial western Eurasia. *Nature* 625(7994). 301–311.

An, Jun. 1986. *Hezheyu jianzhi*. Beijing: Minzu chubanshe.

Anderson, Gregory D.S. 2015. Overview of the Munda languages. In Paul Sidwell & Mathias Jenny (eds.), *The handbook of Austroasiatic languages* (Grammars and Language Sketches of the World's Languages), 364–414. Leiden: Brill.

Anttila, Raimo. 1989. *Historical and comparative linguistics*. 2nd edn. Amsterdam: John Benjamins.

van der Auwera, Johan. 2012. From contrastive linguistics to linguistic typology. *Languages in Contrast: International Journal for Contrastive Linguistics* 12(1). 69–86.

Avrorin, Valentin A. & Elena P. Lebedeva. 1978. *Oročskije teksty i slovar'*. Leningrad: Nauka.

Ayafor, Miriam & Melanie Green. 2017. *Cameroon Pidgin English: A comprehensive grammar* (London Oriental and African Language Library 20). Amsterdam: John Benjamins.

Bäcker, Jörg (ed.). 1988. *Märchen aus der Mandschurei*. Munich: Eugen Diederichs Verlag.

Bakker, Peter. 2017. Typology of mixed languages. In Alexandra Y. Aikhenvald & R. M. W. Dixon (eds.), *The Cambridge handbook of linguistic typology*, 217–253. Cambridge: Cambridge University Press.

Baxter, William H. & Laurent Sagart. 2014. *Old Chinese: A new reconstruction*. Oxford: Oxford University Press.

Beekes, Robert S. P. 2011. *Comparative Indo-European linguistics: An introduction*. 2nd edn. Amsterdam: John Benjamins.

Benzing, Johannes. 1953. Remarques sur les langues tongouses. *Ural-Altaische Jahrbücher* 25. 109–118.

Benzing, Johannes. 1956. *Die tungusischen Sprachen: Versuch einer vergleichenden Grammatik*. Wiesbaden: Akademie der Wissenschaften und der Literatur.

Bickel, Balthasar & Johanna Nichols. 2007. Inflectional morphology. In Timothy Shopen (ed.), *Language typology and syntactic description, Vol 2: Complex constructions*, 2nd edn., 169–240. Cambridge: Cambridge University Press.

Bo, Wenze. 2004. Caijiahua gaikuang. *Minzu Yuwen* 2. 68–81.

Bopp, Franz. 1857. *Vergleichende Grammatik des Sanskrit, Zend, Armenischen, Griechischen, Lateinischen, Litauischen, Altslavischen, Gothischen und Deutschen*. 2nd edn. Vol. 1. Berlin: Ferdinand Dümmler.

Bowern, Claire. 2017. Comparison and relationship of languages. In Jared Klein, Brian Joseph & Matthias Fritz (eds.), *Handbook of comparative and historical Indo-European linguistics* (Handbücher zur Sprach- und Kommunikationswissenschaft 41.1), 1–7. Berlin: de Gruyter Mouton.

Brugmann, Karl. 1904. *Kurze vergleichende Grammatik der indogermanischen Sprachen*. Strassburg: Trübner & Co.

Bulatova, Nadezhda & Stefano Cotrozzi. 2004. *Ewenkische Erzählungen aus der Insel Sachalin*. Bonn: Verlag für Kultur und Wissenschaft.

Bybee, Joan L. 2015. *Language change* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

Campbell, Lyle. 2013. *Historical linguistics: An introduction*. 3rd edn. Edinburgh: Edinburgh University Press.

Campbell, Lyle. 2018. How many language families are there in the world? *Anuario del Seminario de Filología Vasca "Julio de Urquijo"* 52. 133–152.

Campbell, Lyle. 2024. *The indigenous languages of the Americas: History and classification*. Oxford: Oxford University Press.

Castrén, Matthias Alexander. 1856. *Grundzüge einer tungusischen Sprachlehre*. St. Petersburg: Kaiserliche Akademie der Wissenschaften.

CGZZ. 2004. Baizu yuyan. In Wang, Bo et al. (ed.), *Caiguanzhen zhi*, 493–496. Guiyang: Guizhou renmin chubanshe.

Chaoke, D. O. 2007. *Nanmu elunchunyu yanjiu*. Beijing: Minzu chubanshe.

Chaoke, D. O. 2017. *Ewenke zu san da fangyan cihui bijiao*. Beijing: Shehui kexue wenxian chubanshe.

Chaoke, D. O. & Kalina. 2017. *Arong ewenkeyu*. Beijing: Shehui kexue wenxian chubanshe.

Che, Guangxian et al. 2003. *Liupanshuishi zhi: minzu zhi*. Guiyang: Guizhou renmin chubanshe.

Chien, Yuehchen. 2024. Yilan Creole. In Paul Jen-kuei Li, Elizabeth Zeitoun & Rik De Busser (eds.), *Handbook of Formosan languages. Vol. 3: The indigenous languages of Taiwan*, 83–122. Leiden: Brill.

Cincius, Vera I. 1949. *Sravnitel'naja fonetika tunguso-man'čžurskich jazykov*. Leningrad: Uchpergiz.

Cincius, Vera I. 1982. *Negidal'skij jazyk: issledovanija i materialy*. Leningrad: Nauka.

Cincius, Vera I. 1975/77. *Sravnitel'nyj slovar' tunguso-man'čžurskich jazykov*. 2 vols. Leningrad: Nauka.

CLLY. 1950s. *Caijia, longjia, luren yuyan diaocha jianbao*. (Anonymous, undated, and unpublished manuscript probably from the 1950s.)

Collinge, N.E. 1995. History of comparative linguistics. In E. F. K. Koerner & R.E. Asher (eds.), *Concise history of the language sciences: From the Sumerians to the Cognitivists*, 195–202. Oxford: Pergamon.

Comrie, Bernard. 1989. *Language universals and linguistic typology*. 2nd edn. Chicago: The University of Chicago Press.

Croft, William. 2003. *Typology and universals*. 2nd edn. (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

Croft, William. 2016. Comparative concepts and language-specific categories: Theory and practice. *Linguistic Typology* 20(2). 377–393.

Croft, William. 2022. *Morphosyntax: Constructions of the world's languages* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.
Croft, William, Keith Denning & Suzanne Kemmer. 1990. Typology and diachrony in the work of Joseph H. Greenberg. In William Croft, Suzanne Kemmer & Keith Denning (eds.), *Studies in typology and diachrony: Papers presented to Joseph H. Greenberg on his 75th birthday*, ix–xviii. Amsterdam: John Benjamins.
Crossley, Pamela K. 1997. *The Manchus*. Oxford: Blackwell.
Delbrück, Berthold. 1893. *Vergleichende Syntax der indogermanischen Sprachen*. Vol. 1. Strassburg: Trübner.
Diessel, Holger. 2003. The relationship between demonstratives and interrogatives. *Studies in Language* 27(3). 635–655.
Ding, Hongdi & Sicong Dong. 2024. Interrogatives of reason in Tibeto-Burman languages of Sichuan. *Language and Linguistics* 25(2). 187–233.
Doerfer, Gerhard. 1978. Classification problems of Tungus. In Gerhard Doerfer & Michael Weiers (eds.), *Beiträge zur nordasiatischen Kulturgeschichte* (Tungusica 1), 1–26. Wiesbaden: Harrassowitz.
Doerfer, Gerhard, Wolfram Hesche & Hartwig Scheinhardt. 1980. *Lamutisches Wörterbuch*. Wiesbaden: Harrassowitz.
Doerfer, Gerhard & Michael Knüppel. 2013. *Armanisches Wörterbuch*. Nordhausen: Verlag Traugott Bautz.
Dryer, Matthew S. 2007. Word order. In Timothy Shopen (ed.), *Language typology and syntactic description, volume i: Clause structure*, 2nd edn., vol. 1, 61–131. Cambridge: Cambridge University Press.
Dryer, Matthew S. 2011. The evidence for word order correlations. *Linguistic Typology* 15(2). 335–380.
Dryer, Matthew S. 2013a. Order of adjective and noun (v2020.4). In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Zenodo: https://doi.org/10.5281/zenodo.13950591.
Dryer, Matthew S. 2013b. Order of numeral and noun (v2020.4). In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Zenodo: https://doi.org/10.5281/zenodo.13950591.
Dryer, Matthew S. 2013c. Order of object and verb (v2020.4). In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Zenodo: https://doi.org/10.5281/zenodo.13950591.
Dryer, Matthew S. 2013d. Order of relative clause and noun (v2020.4). In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Zenodo: https://doi.org/10.5281/zenodo.13950591.
Dryer, Matthew S. 2017. Word order in Sino-Tibetan languages from a typological and geographical perspective. In Graham Thurgood & Randy LaPolla (eds.), *The Sino-Tibetan languages* (Routledge Language Family Series), 70–82. London: Routledge.
Dryer, Matthew S. 2018. On the order of demonstrative, numeral, adjective, and noun. *Language* 94. 798–833.
Dryer, Matthew S. 2019. Grammaticalization accounts of word order correlations. In Karsten Schmidtke-Bode, Natalia Levshina, Susanne Maria Michaelis & Ilja A. Seržant (eds.), *Explanation in typology: Diachronic sources, functional motivations and the nature of the evidence* (Conceptual Foundations of Language Science 1), 63–95. Berlin: Language Science Press.
Duan, Ling & Xinglian Zhang. 2008. Yunnan, guizhou, hunan baizu qige yuyan diaocha ziliao. In Xu Lin (ed.), *Dali congshu: baiyu pian*, vol. 4, 2902–3076. Kunming: Yunnan minzu chubanshe.
Elliott, Mark C. 2001. *The Manchu way*. Stanford: Stanford University Press.
Ellis, Jeffrey. 1966. *Towards a general comparative linguistics*. Berlin: De Gruyter Mouton.
Enfield, N. J. 2014. *Natural causes of language: Frames, biases, and cultural transmission* (Conceptual Foundations of Language Science 1). Berlin: Language Science Press.
Enfield, N. J. 2018. *Mainland Southeast Asian languages: A concise typological introduction*. Cambridge: Cambridge University Press.
Enfield, N. J. 2021. *The languages of Mainland Southeast Asia* (Cambridge Language Surveys). Cambridge: Cambridge University Press.
Enfield, N. J. & Bernard Comrie. 2015a. *Languages of Mainland Southeast Asia* (Pacific Linguistics 649). Berlin: Mouton de Gruyter.
Enfield, N. J. & Bernard Comrie. 2015b. Mainland Southeast Asian languages: State of the art and new directions. In N. J. Enfield & Bernard Comrie (eds.), *Languages of Mainland Southeast Asia* (Pacific Linguistics 649), 1–27. Berlin: Mouton de Gruyter.
Enhebatu, Merden. 1995. *Manyu kouyu yanjiu*. Hohot: Neimonggu daxue chubanshe.
Enke, Dankmar, Thilo Weber, Larry Hyman, Johanna Nichols, Guido Seiler & Andreas Hölzl (eds.). 2024. *Language change for the worse* (Studies in Diversity Linguistics 33). Berlin: Language Science Press.
Fortson, Benjamin W. 2010. *Indo-European languages and culture: An introduction*. 2nd edn. Malden: Blackwell.
Fox, A. 2006. Historical and comparative linguistics in the 19th century. In Keith Brown (ed.), *Encyclopedia of language & linguistics*, 2nd edn., 317–326. Oxford: Elsevier.
Fulk, Robert D. 2018. *A comparative grammar of the early Germanic languages* (Studies in Germanic Linguistics 3). Amsterdam: John Benjamins.
von der Gabelentz, Georg. 1869. Ideen zu einer vergleichenden Syntax. *Zeitschrift fur Völkerpsychologie und Sprachwissenschaft* 6. 376–384.
von der Gabelentz, Georg. 1875. Weiteres zur vergleichenden Syntax. *Zeitschrift für Völkerpsychologie und Sprachwissenschaft* 8. 129–165.
von der Gabelentz, Georg. 1894. [T]ypologie der Sprachen, eine neue Aufgabe der Linguistik. *Indogermanische Forschungen* 4. 1–7.
Georg, Stefan. 2004. Unreclassifying Tungusic. In Carsten Naeher (ed.), *Proceedings of the First International Conference on Manchu-Tungus Studies*, vol. 2 (Tunguso Sibirica 9), 45–57. Wiesbaden: Harrassowitz.
GMSWSB, Guizhousheng minzu shiwu weiyuanhui shibie bangongshi. 1982a. Caijia de yuyan. In *Guizhou minzu shibie ziliaoji, vol. 8 longjia, caijia*, 110–125. Guiyang.
GMSWSB, Guizhousheng minzu shiwu weiyuanhui shibie bangongshi. 1982b. Longjiayu diaocha baogao. In *Guizhou minzu shibie ziliaoji, vol. 8 longjia, caijia*, 22–43. Guiyang.

Gnecchi-Ruscone, Guido Alberto et al. 2022. Ancient genomes reveal origin and rapid trans-Eurasian migration of 7th century Avar elites. *Cell* 185(8). 1402–1413.
Gnecchi-Ruscone, Guido Alberto et al. 2025. Ancient genomes reveal trans-Eurasian connections between the European Huns and the Xiongnu empire. *Proceedings of the National Academy of Sciences* 122(9). e2418485122.
Goddard, Cliff. 2005. *The languages of East and Southeast Asia: An introduction*. Oxford: Oxford University Press.
Grube, Wilhelm. 1900. *Goldisch-deutsches Wörterverzeichnis*. Sankt Petersburg: Kaiserliche Akademie der Wissenschaften.
Hackstein, Olaf. 2017. The morphological and constructional evolution of OHG *huuanta* and Dutch *want*. *Indo-European Linguistics* 5. 1–30.
Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687.
Haspelmath, Martin. 2014. Comparative syntax. In Andrew Carnie, Yosuke Sato & Dan Siddiqi (eds.), *The Routledge handbook of syntax*, 490–508. London: Routledge.
Haspelmath, Martin. 2019. Indexing and flagging, and head and dependent marking. *Te Reo* 62(1). 93–115.
Haspelmath, Martin. 2025. *Against grammatical categories (and what to replace them with)*. Paper presented at the Colloquium of Slavic Linguistics, University of Potsdam, 2025.04.16, Potsdam.
Hawkins, John A. 1986. *A comparative typology of English and German: Unifying the contrasts*. London, Sydney: Croom Helm.
Heggarty, Paul & Colin Renfrew. 2014. East Asia: Languages. In Colin Renfrew & Paul Bahn (eds.), *The Cambridge world prehistory*, vol. 2 (Cambridge Histories), 870–896. Cambridge: Cambridge University Press.
Her, One-Soon. 2017. Deriving classifier word order typology, or Greenberg's Universal 20A and Universal 20. *Linguistics* 55(2). 265–303.
Hock, Hans Henrich & Brian D. Joseph. 2009. *Language history, language change, and language relationship: An introduction to historical and comparative linguistics*. 2nd edn. (Trends in Linguistics: Studies and Monographs 218). Berlin: Mouton de Gruyter.
Hockett, Charles F. 1987. *Refurbishing our foundations: Elementary linguistics from an advanced point of view* (Current Issues in Linguistic Theory). Amsterdam: John Benjamins.
Holm, John & Peter Patrick. 2007. *Comparative creole syntax: Parallel outlines of 18 creole grammars* (Westminster Creolistics Series 7). London: Battlebridge.
Hölzl, Andreas. 2012. Negation und die Stellung des Mandschurischen innerhalb der tungusischen Sprachen. In Daniel Leese (ed.), *Hou Han(s') Shu 後漢斯書: Festschrift zum 50. Geburtstag von Hans van Ess*, 35–67. Munich: Institute of Sinology.
Hölzl, Andreas. 2015. A typology of negation in Tungusic. *Studies in Language* 39(1). 118–159.
Hölzl, Andreas. 2016. Towards a new typology of questions. *Bavarian Working Papers in Linguistics* 5. 17–28.
Hölzl, Andreas. 2017a. *Kilen: Synchronic and diachronic profile of a mixed language*. Paper presented at 24. LIPP-Symposium "Sprachen im Kontakt: gestern – heute – morgen", Ludwig Maximilian University of Munich, 2017.06.21–23, Munich.
Hölzl, Andreas. 2017b. New evidence on Para-Mongolic numerals. *Journal de la Société Finno-Ougrienne* 96. 21–37.
Hölzl, Andreas. 2018a. *A typology of questions in Northeast Asia and beyond: An ecological perspective* (Studies in Diversity Linguistics 20). Berlin: Language Science Press.
Hölzl, Andreas. 2018b. Constructionalization areas: The case of negation in Manchu. In Evie Coussé, Peter Andersson & Joel Olofsson (eds.), *Grammaticalization meets Construction Grammar* (Constructional Approaches to Language 21), 241–276. Amsterdam: Benjamins.
Hölzl, Andreas. 2018c. Udi, Udihe, and the language(s) of the Kyakala. *International Journal of Diachronic Linguistics and Linguistic Reconstruction* 15. 111–146.
Hölzl, Andreas. 2020a. Bala (China): Language snapshot. *Language Documentation and Description* 19. 162–170.
Hölzl, Andreas. 2020b. *Language obsolescence in retrospect: The case of Alchuka*. Paper presented at the First Conference on the Endangered Languages of East Asia (CELEA 1), University of Venice, 2020.09.02-03, Venice.
Hölzl, Andreas. 2021a. Longjia (China): Language contexts. *Language Documentation and Description* 20. 13–34.
Hölzl, Andreas. 2021b. *Prenasalization in Longjia*. Paper presented at the 34th Paris Meeting on East Asian Linguistics (JLAO), Centre de recherches linguistiques sur l'Asie orientale (CRLAO), 2021.07.07–09, Paris.
Hölzl, Andreas. 2021c. *The Eynu language*. Paper presented at Language contact in the Altaic world: A multiperspective approach, Univerzita Karlova, 2021.11.19–20, Prague.
Hölzl, Andreas. 2021d. *The Lu(ren) languages of Guizhou, China*. Paper presented at 9th International Constrastive Linguistics Conference (ICLC), Università di Genova, 2021.05.20–21, Santa Margherita-Portofino.
Hölzl, Andreas. 2021e. The only known text from Bala, an extinct Tungusic language. *Studia Orientalia Electronica* 9(1). 173–191.
Hölzl, Andreas. 2022a. 'What's your name?' in Tungusic and beyond. In Andreas Hölzl & Thomas E. Payne (eds.), *Tungusic languages: Past and present* (Studies in Diversity Linguistics 32), 89–148. Berlin: Language Science Press.
Hölzl, Andreas. 2022b. The etymology of "Manchu": A critical evaluation of the riverside hypothesis. *International Journal of Eurasian Linguistics* 4(2). 160–208.
Hölzl, Andreas. 2022c. *Unusual word order features in Longjia and Caijia*. Paper presented at Syntax Semantics Colloquium, University of Potsdam, 2022.07.26, Potsdam.
Hölzl, Andreas. 2023a. Diachronic changes in constructional networks: Evidence from Manchuric nominal morphology. *Constructions and Frames* 15(2). 160–186.
Hölzl, Andreas. 2023b. *Manchu as a contact language*. Paper presented at Lecture Series on "Language Contact in Eurasia", Seminar für Turkologie und Zentralasienkunde, University of Göttingen, 2023.12.14, Göttingen.
Hölzl, Andreas. 2023c. Review of Lü Shanshan. 2022. A reference grammar of Caijia: An unclassified language of Guizhou (Sinitic Languages of China 8.). Berlin: de Gruyter Mouton. *Bulletin of the School of Oriental and African Studies* 86(2). 398–399.
Hölzl, Andreas. 2024a. *Ethnolinguistic aspects of birch trees: Tungusic and beyond*. Paper presented at Linguistic prehistory and ecology in the Northern Pacific Rim, Max Planck Institute of Geoanthropology, 28–29.08.2024, Jena.

Hölzl, Andreas. 2024b. Historische Phonologie (Synopseartikel). In Mechthild Habermann & Ilse Wischer (eds.), *Historische Sprachwissenschaft* (WSK Online). Berlin: Mouton de Gruyter.
Hölzl, Andreas. 2024c. Isolierende Sprache. In Mechthild Habermann & Ilse Wischer (eds.), *Historische Sprachwissenschaft* (WSK Online). Berlin: Mouton de Gruyter.
Hölzl, Andreas. 2024d. Kontaktsprache. In Mechthild Habermann & Ilse Wischer (eds.), *Historische Sprachwissenschaft* (WSK Online). Berlin: Mouton de Gruyter.
Hölzl, Andreas. 2024e. Kreolsprache. In Mechthild Habermann & Ilse Wischer (eds.), *Historische Sprachwissenschaft* (WSK Online). Berlin: Mouton de Gruyter.
Hölzl, Andreas. 2024f. Mischsprache. In Mechthild Habermann & Ilse Wischer (eds.), *Historische Sprachwissenschaft* (WSK Online). Berlin: Mouton de Gruyter.
Hölzl, Andreas. 2024g. Paläo-asiatische Sprachen. In Mechthild Habermann & Ilse Wischer (eds.), *Historische Sprachwissenschaft* (WSK Online). Berlin: Mouton de Gruyter.
Hölzl, Andreas. 2024h. Pidginsprache. In Mechthild Habermann & Ilse Wischer (eds.), *Historische Sprachwissenschaft* (WSK Online). Berlin: Mouton de Gruyter.
Hölzl, Andreas. 2024i. Polysynthetische Sprache. In Mechthild Habermann & Ilse Wischer (eds.), *Historische Sprachwissenschaft* (WSK Online). Berlin: Mouton de Gruyter.
Hölzl, Andreas. 2024j. Postnominal flagging and OV in Sinitic: Areal and typological perspectives. *Studies in Language* 49. 1–45.
Hölzl, Andreas. 2024k. *Prenominal relative clauses in VO and postnominal genitives in OV languages.* Paper presented at Syntax of the World's Languages 9, Pontificia Universidad Católica del Perú, 23–26.07.2024, Lima.
Hölzl, Andreas. 2024l. Review of Alexander Vovin†, José Andrés Alonso de la Fuente, Juha Janhunen (eds.). 2024. The Tungusic languages (Routledge Language Family Series). London: Routledge. *Bulletin of the School of Oriental and African Studies* 87(3). 580–582.
Hölzl, Andreas. 2024m. Sprachtyp. In Mechthild Habermann & Ilse Wischer (eds.), *Historische Sprachwissenschaft* (WSK Online). Berlin: Mouton de Gruyter.
Hölzl, Andreas. 2024n. Sprachtypologie (Synopseartikel). In Mechthild Habermann & Ilse Wischer (eds.), *Historische Sprachwissenschaft* (WSK Online). Berlin: Mouton de Gruyter.
Hölzl, Andreas. 2024o. *The architecture of interrogative systems.* Paper presented at Linguistic Lecture Series, Centre Language – Variation – Multiligualism, University of Potsdam, 2024.05.15, Potsdam.
Hölzl, Andreas. 2024p. The complexification of Tungusic interrogative systems. In Dankmar Enke, Thilo Weber, Larry Hyman, Johanna Nichols, Guido Seiler & Andreas Hölzl (eds.), *Language change for the worse* (Studies in Diversity Linguistics 33), 211–241. Berlin: Language Science Press.
Hölzl, Andreas. 2025a. A typology of the deictic day name system in Manchuric. *Eurasiatica: Quaderni di studi su Balcani, Anatolia, Iran, Caucaso e Asia Centrale* 22. 189–217.
Hölzl, Andreas. 2025b. Konsonantenschwächung. In Mechthild Habermann & Ilse Wischer (eds.), *Historische Sprachwissenschaft* (WSK Online). Berlin: Mouton de Gruyter.
Hölzl, Andreas. 2025c. Konsonantenschwund. In Mechthild Habermann & Ilse Wischer (eds.), *Historische Sprachwissenschaft* (WSK Online). Berlin: Mouton de Gruyter.
Hölzl, Andreas & Chundra Cathcart. 2019. *Sortal numeral classifiers in Central Asia.* Paper presented at 13th Conference of the Association for Linguistic Typology (ALT), Università di Pavia, 04–06.09.2019, Pavia.
Hölzl, Andreas & Yadi Hölzl. 2019a. A wedding ceremony of the Kyakala in China: Language and ritual. *International Journal of Diachronic Linguistics and Linguistic Reconstruction* 16. 87–144.
Hölzl, Andreas & Yadi Hölzl. 2019b. *The endangered languages of the Manchus.* Paper presented at Endangered languages in Northern Asia, Free University of Berlin, 2019.11.29-30, Berlin.
Hölzl, Andreas & Yadi Hölzl. 2025. The lost voices of Guizhou: Rediscovering the languages of the Luren. *Cahiers de Linguistique Asie Orientale*. 1–39.
Hölzl, Andreas & Peter-Arnold Mumm (eds.). 2018. *Klasse Person. Festschrift für Wolfgang Schulze.* Special issue of *International Journal of Diachronic Linguistics and Linguistic Reconstruction* 15.
Hölzl, Andreas & Thomas E. Payne. 2022a. Introduction. In Andreas Hölzl & Thomas E. Payne (eds.), *Tungusic languages: Past and present* (Studies in Diversity Linguistics 32), 1–20. Berlin: Language Science Press.
Hölzl, Andreas & Thomas E. Payne (eds.). 2022b. *Tungusic languages: Past and present* (Studies in Diversity Linguistics 32). Berlin: Language Science Press.
Hsiu, Andrew. 2013. *New endangered Tibeto-Burman languages of southwestern China: Mondzish, Longjia, Pherbu, and others.* Paper presented at ICSTLL 46, Dartmouth College, 2013.08.08, New Hampshire.
Hsiu, Andrew. 2018. *The Niujiaojing dialect of Caijia in Weining county, Guizhou, China.* (Unpublished manuscript).
Huber, Magnus & the APiCS Consortium. 2013. Order of subject, object, and verb. In Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.), *The atlas of pidgin and creole language structures.* Oxford: Oxford University Press. https://apics-online.info/parameters/1.
von Humboldt, Wilhelm. 1946 [1820]. *Über das vergleichende Sprachstudium in Beziehung auf die verschiedenen Epochen der Sprachentwicklung*. Leipzig: Felix Meiner.
Ikegami, Jirō. 1997. *Uirutago jiten.* Sapporo: Hokkaidō University Press.
Ioannidis, Alexander G. et al. 2020. Native American gene flow into Polynesia predating Easter Island settlement. *Nature* 583(7817). 572–577.
Jacques, Guillaume. 2011. A panchronic study of aspirated fricatives, with new evidence from Pumi. *Lingua* 121(9). 1518–1538.
Jakobson, Roman. 1955. *Slavic languages: A condensed survey.* 2nd edn. New York: King's Crown Press.

Janhunen, Juha. 2005. Tungusic: An endangered language family in Northeast Asia. *International Journal of the Sociology of Language* 173(173). 37–54.
Janhunen, Juha. 2023. The unity and diversity of Altaic. *Annual Review of Linguistics* 9. 135–154.
Janhunen, Juha. 2024a. Proto-Tungusic. In Alexander Vovin, José Andrés Alonso de la Fuente & Juha Janhunen (eds.), *The Tungusic languages* (Routledge Language Family Series), 35–75. London: Routledge.
Janhunen, Juha. 2024b. Tungusic as a language family. In Alexander Vovin, José Andrés Alonso de la Fuente & Juha Janhunen (eds.), *The Tungusic languages* (Routledge Language Family Series), 1–18. London: Routledge.
Jettmar, H. M. 1937. Der Stamm der oberen Golden (Die Sungari-Ussuri-Gruppe). *Mitteilungen der Anthropologischen Gesellschaft in Wien* 5. 245–276.
Jin, Ning. 1993. *Sibe English conversations*. Wiesbaden: Harrassowitz.
Kałużyński, Stanisław. 1971a. Solonisches Wörterverzeichnis nach F.V. Muromskis handschriftlichen Sprachaufzeichnungen. *Rocznik Orientalistyczny* 34(1). 39–77.
Kałużyński, Stanisław. 1971b. Solonisches Wörterverzeichnis nach F.V. Muromskis handschriftlichen Sprachaufzeichnungen (Fortsetzung und Schluß). *Rocznik Orientalistyczny* 34(2). 15–56.
Kane, Daniel & Marc Hideo Miyake. 2024. Jurchen. In Alexander Vovin, José Andrés Alonso de la Fuente & Juha Janhunen (eds.), *The Tungusic languages* (Routledge Language Family Series), 76–102. London: Routledge.
Kayne, Richard. 2013. Comparative syntax. *Lingua* 130. 132–151.
Kazama, Shinjiro. 1998. Genetical position of Hezhen. In Osahita Miyaoka & Minoru Oshima (eds.), *Languages of the North Pacific Rim, vol 4*, 115–131. Kyōto: Kyoto: ELPR.
Kazama, Shinjiro. 2003. *Basic vocabulary (A) of Tungusic languages* (Publications on Tungus Languages and Cultures 25). Kyōto: Osaka: ELPR.
Kim, Juwon. 2011. *A grammar of Ewen* (Altaic Languages Series 6). Seoul: Seoul National University Press.
Kiparsky, Paul. 2006. The Amphichronic Program vs. Evolutionary Phonology. *Theoretical Linguistics* 32(2). 217–236.
Kiparsky, Paul. 2015. New perspectives in historical linguistics. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics*, 64–102. London: Routledge.
Kiyose, Gisaburo N. 2000. Genealogical relationship of Jurchen dialects and literary Manchu. *Central Asiatic Journal* 44. 177–189.
von Klaproth, Julius H. 1831. *Asia polyglotta*. 2nd edn. Paris: Verlag von Heideloff & Campe.
Ko, Dongho & Gyudong Yurn. 2011. *A description of Najkhin Nanai* (Altaic Languages Series 7). Seoul: Seoul National University Press.
Krahe, Hans. 1970. *Einleitung in das vergleichende Sprachstudium* (Innsbrucker Beiträge zur Sprachwissenschaft). Innsbruck: Universität Innsbruck.
Krzeszowski, Tomasz P. 1984. Tertium comparationis. In Jacek Fisiak (ed.), *Contrastive linguistics: Prospects and problems*, 301–312. Berlin: De Gruyter Mouton.
Latham, Robert Gordon. 1862. *Elements of comparative philology*. London: Walton & Maberly.
Lattimore, Owen. 1933. The Gold tribe of the lower Sungari. *Memoires of the American Anthropological Association* 40. 1–77.
Lazaridis, Iosif et al. 2025. The genetic origin of the Indo-Europeans. *Nature* 615(7951). 1–9.
Lee, Man Hei. 2023. Phonological features of Caijia that are notable from a diachronic perspective. *Journal of Historical Linguistics* 13. 82–114.
Lee, Man Hei. 2024. Bai and Old Western Chinese. *Bulletin of the School of Oriental and African Studies* 87(3). 533–555.
Lefebvre, Claire. 2015. *Relabeling in language genesis*. Oxford: Oxford University Press.
Lehmann, Christian. 1989. Language description and general comparative grammar. In Gottfried Graustein & Gerhard Leitner (eds.), *Reference grammars and modern linguistic theory* (Linguistische Arbeiten 226), 133–183. Berlin: De Gruyter.
Lehmann, Winfred P. 1992. Comparative linguistics. In Edgar C. Polomé & Werner Winter (eds.), *Reconstructing languages and cultures*, 3–22. Berlin: De Gruyter Mouton.
Ling, Chunsheng. 1934. *Songhuajiang xiayou de hezhezu*. Nanjing: Guoli zhongyang yanjiuyuan lishi yuyan yanjiusuo.
Lord, Robert. 1966. *Comparative linguistics*. London: English Universities Press.
Lü, Shanshan. 2017. Caijiahua qizhong youbiju de leixingxue kaocha. *Fangyan* 3. 341–351.
Lü, Shanshan. 2018. Two locative constructions in Caijia from the typological perspective of Asian languages. *Studies in Language* 42(3). 600–640.
Lü, Shanshan. 2019a. Caijiahua dengbiju de leixingxue kaocha. *Yuyan kexue* 99. 160–175.
Lü, Shanshan. 2019b. Equative and similative constructions in Caijia (Sino-Tibetan). *Faits de Langues* 50(1). 203–225.
Lü, Shanshan. 2022. *A reference grammar of Caijia: An unclassified language of Guizhou, China* (Sinitic Languages of China 8). Berlin: Mouton.
Luo, Xianren. 2006 [1995]. Dafang yiyu diaocha. In Guizhousheng minzu shiwu weiyuanhui (ed.), *Guizhou "liushan liushui"' minzu diaocha ziliao xuanbian, minzu yuyan juan*, 126–142. Guiyang: Guizhou minzu chubanshe.
Ma, Lisan & Chen Ying. 2001. *Dian chuan qian gui yiwen zidian*. Kunming: Yunnan minzu chubanshe.
Ma, Xueliang & Wulaxichun. 1993. Manyuzhi yuyan zhong de songqi qing cayin. *Minzu yuwen* 6. 4–9, 37.
Mackenzie, J. Lachlan. 2009. Content interrogatives in a sample of 50 languages. *Lingua* 119. 1131–1163.
Majewicz, Alfred F. 2023. Shanshan Lü 2022. A reference grammar of Caijia. An unclassified language of Guizhou. Sinitic Languages of China – Typological Descriptions volume 8. Berlin/Boston: De Gruyter Mouton. Pp. xxviii, 600. *Rocznik Orientalistyczny* 76(1). 176–184.
Mallory, James P. 2025. *The Indo-Europeans rediscovered: How a scientific revolution is rewriting their story*. London: Thames & Hudson.
Mallory, James P. & Douglas Q. Adams. 2006. *The Oxford introduction to Proto-Indo-European and the Proto-Indo-European world* (Oxford Linguistics). Oxford: Oxford University Press.

Markey, T.L. 1982. Afrikaans: Creole or non-creole? *Zeitschrift für Dialektologie und Linguistik* 49. 169–207.
Matras, Yaron. 2020. *Language contact*. 2nd edn. (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.
Mazzoli, Maria & Eeva Sippola. 2021. Mixed languages: From core to fringe. In Maria Mazzoli & Eeva Sippola (eds.), *New perspectives on mixed languages: From core to fringe*, 1–26. Berlin: De Gruyter Mouton.
McElvenny, James. 2024. *A history of modern linguistics: From the beginnings to World War II*. Edinburgh: Edinburgh University Press.
McMahon, April. 1994. *Understanding language change*. Cambridge: Cambridge University Press.
McWhorter, John. 2007. *Language interrupted: Signs of non-native acquisition in standard language grammars*. Oxford: Oxford University Press.
Meillet, Antoine. 1908. *Introduction à l'étude comparative des langues indo-européennes*. 2nd edn. Paris: Hachette.
Moravcsik, Edith A. 2013. *Introducing language typology* (Cambridge Introductions to Language and Linguistics). Cambridge: Cambridge University Press.
Morpurgo Davies, Anna. 1992. Comparative-historical linguistics. In William Bright (ed.), *International encyclopedia of linguistics*, vol. 2, 159–163. Oxford: Oxford University Press.
Mu, Yejun. 1986. Alechuka manyu de shuci yu gezhuci. *Manyu yanjiu* 1986(1). 2–17.
Mu, Yejun. 1987. Balayu. *Manyu yanjiu* 1987(2). 2–31, 128.
Muysken, Pieter & Norval Smith. 1990. Question words in pidgin and creole languages. *Linguistics* 28(4). 883–903.
NDSSLD, Neimeng dongbei shaoshuminzu shehui lishi diaochazu. 1958. *Heilongjiangsheng fuyuanxian xiabacha hezhe minzuxiang qingkuang (hezhezu diaocha cailiao zhi yi)*. Beijing: Quanguo renmin daibiao dahui minzu weiyuanhui bangongshi.
Nedjalkov, Igor. 1994. Evenki. In Peter Kahrel & René Berg (eds.), *Typological studies in negation* (Typological Studies in Language 29), 1–34. Amsterdam: John Benjamins.
Nedjalkov, Igor. 1997. *Evenki* (Descriptive Grammars Series). London: Routledge.
Nettle, Daniel. 1999. *Linguistic diversity* (Oxford Linguistics). Oxford: Oxford University Press.
Nichols, Johanna. 1997. Modeling ancient population structures and movement in linguistics. *Annual Review of Anthropology* 26(1). 359–384.
Nikolaeva, Irina & Maria Tolskaya. 2001. *A grammar of Udihe* (Mouton Grammar Library 22). Berlin: Mouton de Gruyter.
Norquest, Peter K. 2016. *A phonological reconstruction of Proto-Hlai* (Languages of Asia). Leiden: Brill.
Ostapirat, Weera. 2005. Kra-Dai and Austronesian: Notes on phonological correspondences and vocabulary distribution. In Laurent Sagart, Roger M. Blench & Alicia Sanchez-Mazas (eds.), *The peopling of East Asia*, 107–131. London: Routledge.
Ostapirat, Weera. 2018. Macrophyletic trees of East Asian languages reexamined. *Senri Ethnological Studies* 98. 107–121.
Pallas, Peter Simon. 1786. *Linguarum totius orbis vocabularia comparativa*. Vol. 1. Saint Petersburg: Typis Iohannis Caroli Schnoor.
Pallas, Peter Simon. 1789. *Linguarum totius orbis vocabularia comparativa*. Vol. 2. Saint Petersburg: Typis Iohannis Caroli Schnoor.
Palmer, Bill. 2018. Language families of the New Guinea Area. In Bill Palmer (ed.), *Papuan languages and linguistics* (The World of Linguistics 4), 1–20. Berlin: Mouton.
Poniatowski, Stanisław. 1923. *Materials to the vocabulary of the Amur Gold* (Bibliotheca Universitatis Liberae Polonae 10). Warsaw: Cura et sumptibus Universitatis liberae Polonae.
Potter, Ben A. 2023. Integrating ancient genetic and archaeological patterns for the peopling of Northeast Asia and the Americas. *Anthropological Papers of the University of Alaska* 7(1-2). 41–62.
Raff, Jennifer. 2022. *Origin: A genetic history of the Americas*. New York: Twelve.
Ramsey, Robert S. 1987. *The languages of China*. Princeton: Princeton University Press.
Rankin, Robert L. 2003. The comparative method. In Brian D. Joseph & Richard D. Janda (eds.), *The handbook of historical linguistics*, 183–212. Malden: Blackwell.
Reich, David. 2018. *Who we are and how we got here: Ancient DNA and the new science of the human past*. New York: Pantheon Books.
Ross, Malcolm & Mark Durie. 1996. Introduction. In Mark Durie & Malcolm Ross (eds.), *The comparative method reviewed: Regularity and irregularity in language change*, 3–38. Oxford: Oxford University Press.
Sagart, Laurent. 2011. *Classifying Chinese dialects/Sinitic languages on shared innovations*. Paper presented at Séminaire Sino-Tibétain du Centre de Recherches Linguistiques sur l'Asie Orientale, 2011.03.28, Nogent-sur-Marne.
Sandman, Erika. 2016. *A grammar of Wutun*. Helsinki: University of Helsinki dissertation.
Sandman, Erika. 2021. Wutun as a mixed language. In Maria Mazzoli & Eeva Sippola (eds.), *New perspectives on mixed languages: From core to fringe*, 325–359. Berlin: De Gruyter Mouton.
von Schlegel, Friedrich. 1808. *Ueber die Sprache und Weisheit der Indier: Ein Beitrag zur Begründung der Alterthumskunde*. Heidelberg: Mohr und Zimmer.
Schleicher, August. 1848. *Sprachvergleichende Untersuchungen*. Bonn: H. B. König.
Schmidt, P. 1923a. The language of the Negidals. *Acta Universitatis Latviensis* 5. 3–38.
Schmidt, P. 1923b. The language of the Olchas. *Acta Universitatis Latviensis* 8. 231–288.
Schmidt, P. 1928a. The language of the Oroches. *Acta Universitatis Latviensis* 17. 17–62.
Schmidt, P. 1928b. The language of the Samagirs. *Acta Universitatis Latviensis* 19. 218–249.
Schmidt, P. W. 1926. *Die Sprachfamilien und Sprachenkreise der Erde*. Heidelberg: Carl Winter's Universitätsbuchhandlung.
Schulze, Wolfgang. 2005. Towards a history of Udi. *International Journal of Diachronic Linguistics and Linguistic Reconstruction* 2. 55–99.
Schulze, Wolfgang & Andreas Hölzl. 2013. Review of Andrej L. Malchukov & Lindsay J. Whaley (eds.). 2012. Recent advances in Tungusic linguistics (Turcologica 89). Wiesbaden: Harrassowitz. *Studies in Language* 37(2). 454–462.
Sem, Lidija I. 1976. *Očerki dialektov nanajskogo jazyka: bikinskij (ussurijskij) dialekt*. Leningrad: Izdat. Nauka.

Seržant, Ilja, Sergey Say, Andreas Hölzl, Aigul Zakirova, Gao Xinyi & Andreas Pregla. preprint. A typology of positional differential argument marking. *Linguistic Typology*.
Sidwell, Paul. 2020. Nicobarese comparative grammar. In Mathias Jenny, Paul Sidwell & Mark Alves (eds.), *Austroasiatic syntax in areal and diachronic perspective*, 82–104. Leiden: Brill.
Sihler, Andrew L. 1995. *New comparative grammar of Greek and Latin*. Oxford: Oxford University Press.
Slonimska, Anita & Seán G. Roberts. 2017. A case for systematic sound symbolism in pragmatics: Universals in wh-words. *Journal of Pragmatics* 116. 1–20.
Song, Jae Jung. 2018. *Linguistic typology*. Oxford: Oxford University Press.
SUISJ. 1984. *Sibe uksurai irgen siden jube*. Vol. 1. Urumqi: Xinjiang renmin chubanshe.
Sunik, Orest P. 1958. *Kur-urmijskij dialekt: issledovanija i materialy po nanajskomu jazyku*. Leningrad: Uspedgiz.
Syea, Anand. 2017. *French Creoles: A comprehensive and comparative grammar* (Routledge Comprehensive Grammars). London: Routledge.
Syea, Anand. 2024. *A comprehensive and comparative grammar of English Creoles* (Routledge Comprehensive Grammars). London: Routledge.
Szemerényi, Oswald. 1990. *Einführung in die vergleichende Sprachwissenschaft*. 4th edn. Darmstadt: Wissenschaftliche Buchgesellschaft.
Szeto, Pui Yiu & Chingduang Yurayong. 2021. Sinitic as a typological sandwich: Revisiting the notions of Altaicization and Taicization. *Linguistic Typology* 25(3). 551–599.
Tan, Gan-ling. 2023. A new view on 'Yilan Creole'. *Journal of Pidgin and Creole Languages* 38(2). 320–388.
Theisen, Joachim. 2016. *Kontrastive Linguistik: Eine Einführung*. Tübingen: Narr.
Thomason, Sarah G. 1997. A typology of contact languages. In Arthur K. Spears & Donald Winford (eds.), *The structure and status of pidgins and creoles*, 71–88. Amsterdam: John Benjamins.
Trask, Robert L. 2000. *The dictionary of historical and comparative linguistics*. Edinburgh: Edinburgh University Press.
Trask, Robert L. 2015. *Trask's historical linguistics*. Robert McColl Millar (ed.). 3rd edn. London: Routledge.
Tsumagari, Toshiro. 1992. A basic vocabulary of Khamnigan and Oluguya Ewenki in northern Inner Mongolia. *Bulletin of the Institute for the Study of North Eurasian Cultures* 21. 83–103.
Tsumagari, Toshiro. 1997. Linguistic diversity and national borders of Tungusic. *Senri Ethnological Studies* 44. 175–186.
Vajda, Edward J. 2010. A Siberian link with Na-Dene languages. In James Kari & Ben A. Potter (eds.), *The Dene-Yeniseian connection*, 33–99. Fairbanks: Department of Anthropology, University of Alaska Fairbanks.
Vajda, Edward J. 2022. The Dene-Yeniseian hypothesis. In Michael Fortescue & Edward J. Vajda (eds.), *Mid-Holocene language connections between Asia and North America*, 237–480. Leiden: Brill.
Veluppilai, Viveka. 2012. *An introduction to linguistic typology*. Amsterdam: John Benjamins.
Veselinova, Ljuba N. 2015. Special negators in the Uralic languages. In Matti Miestamo, Anne Tamm & Beáta Wagner-Nagy (eds.), *Negation in Uralic languages* (Typological Studies in Language 108), 547–600. Amsterdam: John Benjamins.
Vovin, Alexander. 2006. Why Manchu and Jurchen look so non-Tungusic? In Alessandra Pozzi, Juha Janhunen & Michael Weiers (eds.), *Tumen jalafun jecen akū: Manchu studies in honour of Giovanni Stary*, 255–266. Wiesbaden: Harrassowitz.
Vovin, Alexander, José Andrés Alonso de la Fuente & Juha Janhunen (eds.). 2024. *The Tungusic languages* (Routledge Language Family Series). London: Routledge.
Wang, Feng. 2012. *Kunming xishan shalang baiyu yanjiu*. Beijing: Zhongguo shehui kexue chubanshe.
Wang, Haibo. 2025. Aruchukago goishū: Mu Yejun shiryō no manshūgo bungo sakuin. *Hoppō gengo kenkyū* 15. 237–276.
Wang, Qingfeng. 2005. *Manyu yanjiu*. Beijing: Minzu chubanshe.
Weiss, Michael. 2015. The comparative method. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics*, 127–146. London: Routledge.
Whaley, Lindsay & Sofia Oskolskaya. 2020. The classification of the Tungusic languages. In Martine Robbeets & Alexander Savelyev (eds.), *The Oxford guide to the Transeurasian languages*, 81–91. Oxford: Oxford University Press.
Wu, Changsheng. 1989. *Anshunxian minzuzhi*. Anshun: Anshunxian minzu shiwu weiyuanhui.
Wu, Yunji & Ruiqing Shen. 2010. *Xiangxi guzhang waxianghua diaocha baogao*. Shanghai: Shanghai jiaoyu chubanshe.
Wu [Hölzl], Yadi. 2022. *Deguo he jieke cang guizhou miaotu*. Xi'an: Shanxi shifan daxue chuban zongshe.
Yang, Dehua. 1983. Guanyu longjia yuyan diaocha baogao (taolun gao). In Guizhousheng minzu shibie gongzuodui longjiazu (ed.), *Longjia zushu wenti diaocha baogao*. (Unpublished manuscript.)
Zhai, Huifeng. 2021. *Yiyu dongbu fangyan cankao yufa*. Beijing: Zhongguo shehui kexue chubanshe.
Zhang, Chengkun. 1997. *Zhijin xianzhi*. Beijing: Fangzhi chubanshe.
Zhang, Jimin & Juewei Li. 1982. *Yuyan diaocha dagang: dafang longjiayu*. Guizhousheng minzu shiwu weiyuanhui. (Unpublished field notes.)
Zhao, Jie. 1989. *Xiandai manyu yanjiu*. Beijing: Minzu chubanshe.
Zhao, Xian'guo et al. 1985 [1926]. *Dadingxian zhi*. Chongqing: Yuxin yinshuachang.
Zhengzhang, Shangfang. 2010. Caijiahua baiyu guanxi ji cigen bijiao. In Pan Wuyun & Shen Zhongwei (eds.), *Yanjiu zhi le*, 389–400. Shanghai: Shanghai jiaoyu chubanshe.
Zikmundová, Veronika. 2013. *Spoken Sibe: Morphology of the inflected parts of speech*. Prague: Karolinum.
Zikmundová, Veronika & Gao Wa. 2024. Spoken Manchu. In Alexander Vovin, José Andrés Alonso de la Fuente & Juha Janhunen (eds.), *The tungusic languages*, 463–482. London: Routledge.