Communicative efficiency and institutional pressure on language

by

Sihan Chen

B.S. Mechanical Engineering, University of Miami, 2019.

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN COGNITIVE SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2026

Authored by: Sihan Chen
Department of Brain and Cognitive Sciences
August 7, 2026

Certified by: Edward Gibson
Professor, Thesis Supervisor

Accepted by: Nancy Kanwisher
Walter A. Rosenblith Professor
Graduate Officer, Department of Brain and Cognitive Sciences

THESIS COMMITTEE

Thesis Supervisor

Edward Gibson
Professor
Department of Brain and Cognitive Sciences

Thesis Readers

Roger Levy
Professor
Department of Brain and Cognitive Sciences, MIT

Joshua Tenenbaum
Professor
Department of Brain and Cognitive Sciences, MIT

Michael Ramscar
Professor
Department of Psychology, University of Tuebingen

Damián Blasi
Professor
Center for Brain and Cognition, Pompeu Fabra University

Communicative efficiency and institutional pressure on language

by

Sihan Chen

Submitted to the Department of Brain and Cognitive Sciences
on August 7, 2026 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN COGNITIVE SCIENCE

ABSTRACT

How do languages become the way they are today? An influential view is that languages are shaped by pressures towards communicative efficiency: languages optimize getting the meanings across with only a few distinct words to memorize, and they minimize difficulty in production and comprehension. It has been proposed that the mechanism towards efficient communication is bottom-up: language users do not explicitly optimize for efficiency; instead, they favor a less costly form if it conveys a similar meaning, which becomes conventionalized through repeated individual choices. In this thesis, we investigate the robustness of this bottom-up mechanism in three studies, examining cases where top-down institutional pressures are imposed on language use. The first study is a baseline case, where we show how systems of spatial demonstratives (e.g., "here", "from there") vary across languages by pressure towards communicative efficiency. The second study investigates how regulations on personal name systems may have affected the information structure of names. Despite surface differences between East Asian and Western traditions, historical data reveal a shared information structure: the first element (e.g., "Chen", "Edward") is less informative than the second (e.g., "Sihan", "Gibson"), enabling efficient individual distinction. However, sociopolitical processes transformed different components into hereditary patronyms, first in East Asia, second in the West, reshaping name informativeness. The third study concerns the effectiveness of government efforts to simplify legal language. Legal language is often difficult to understand because of long-distance syntactic dependencies, but it remains unclear why laws are written in such a convoluted way. We hypothesize that it is syntactically transparent to describe a complex legal concept in one complex sentence, which is harder to comprehend. Indeed, we found like modern American laws, legalese texts in ancient China also contain more convoluted structures than contemporary non-legalese ones, possibly because these legal texts often include the offense and the punishment in the same sentence. On the other hand, we found a much smaller difference in dependencies between the two genres in modern Chinese than previously observed in American English. This is possibly because China went through a drastic change from ancient China in legal tradition, and lawmakers adopted simpler alternative ways to convey the same meaning. Taken together, our results suggest the mechanism towards communicative efficiency can be interfered with by institutional constraints.

Thesis supervisor: Edward Gibson
Title: Professor

Acknowledgments

Graduate school can be long, stressful, and tiring, but I feel immensely lucky to have had an amazing experience during my PhD. Below I would like to express my sincere gratitude to a (large) number of individuals, including but not limited to the following:

I would like to thank my advisor Ted Gibson. Ted has been an amazing supervisor along the way, helping me transform from an engineer to a scientist. When I'm hyper-focused on a lot of details, he's the person reminding me to pay attention to the big picture, to think about why I am doing the analysis that I am doing, to think about what kind of alternatives / baselines to compare with. In addition to helping me become a more rigorous researcher, Ted has shown me how to become a better writer and how to convey an idea in a clear and simple way. I am also grateful to Ted for his flexibility, for letting me do my work wherever I want. Last but not least, thanks to Ted for all the Dixit games, and all the extra food and wine that I always ask for.

I would also like to thank my thesis committee: Roger Levy, Josh Tenenbaum, Michael Ramscar, and Damián Blasi, for their continued support. I thank Roger for his rigor, for encouraging me to think deeply and thoroughly about a theory, its mechanism, and the hypothesis. I thank Josh for his way of thinking, for inspiring me to engage and evaluate science in a Bayesian manner. I thank Michael for his passion and for opening up a new way for me to see language, based on which I saw a unifying theme of all my projects. I thank Damián for his broad perspective on human language, culture, and evolution, for motivating me to think about how human language as a whole interacts with other factors.

Next I would like to thank my thesis collaborators: Richard Futrell, Kyle Mahowald, Frank Mollica, Tristan Brown, and Eric Martínez. I thank Richard, Kyle, and Frank for always being a Slack message / email away when I need help with turning my vague ideas into something doable, for giving me detailed feedback at every stage of the project. I thank Tristan for his expertise in Chinese history and his knowledge on classical Chinese literature. I thank Eric for sparking my interest in law, for influencing my way of speaking, and for being a great friend and fellow lab member for 4 years.

I would also like to express my gratitude to two faculty members at BCS: Ev Fedorenko and Nancy Kanwisher. Their way of thinking and presenting science deeply influenced me. Ev gave me a new perspective on how to think about language in relation to other activities in the human mind. Nancy made me more aware of the confounding factors when evaluating scientific work. Both of them deeply influenced me on how to present, how to make complicated ideas easy to understand for everyone. Also, thanks Ev (and Ted) for all the house parties, and Nancy for letting me loiter in her lab for an extended period of time.

I'd also like to thank my collaborators outside my thesis work, especially Antonio Benítez-

Burraco and David Gil, for introducing me to the world of linguistic typology and exploring cool connections with me between languages and societies. I would also like to thank Ljiljana Progovac, Rachel Ryskin, Bevil Conway, Tamar Regev, Noga Zaslavsky, Vera Kempe and Meilin Zhan, for their mentorship and feedback. I owe a lot of people here project updates - I will get back to them as soon as I finish the thesis.

I'd like to acknowledge two professors from my university for leading me to pursue a PhD. I thank Weiyong Gu for giving me an opportunity to work with him, which stimulated my interest in conducting research. I thank Caleb Everett for his fascinating class on introduction to linguistic anthropology: it opened up an entire new world for me. Largely because of his class, I decided to pursue a PhD in language research. I also thank him for his mentorship, from which I had a first experience in doing research on language, and from the experience it became certain to me that language research is my passion.

Thanks to my friends and colleagues in TedLab and EvLab, for supporting me in the past six years, and for their passion for language and for sharing many similar interests with me. Also thanks to everyone in Kanwisher Lab for letting me hang around in their lab space at 3am and for accepting me as an unofficial lab member.

Thanks to all my UROPs that have helped me immensely along the way: Rachel Zhu, Mary Zhou, Sophia Zhang, Erica Zhang, Lia Washington, Deeraj Pothapragada, Lucy Lyu, Vivian Kou, Annie Guo, Jorie Fleming, Nicole Ding, Jonathan Dase, Iris Chen, and Phoebe Bo. Many of them have done an amazing job in analyzing the legalese and non-legalese documents in classical Chinese, which is very difficult and time consuming.

My gratitude also goes to the friends I have as a PhD student, including but not limited to: Aryan Zoroufi, Serena Bono, Mahnia Asadollahi, Bowen Zheng, Laura Nicolae, Selena She, Jess Chomik-Morales, Fernanda de la Torre, Pramod RT, Nishad Gothoskar, Shivali Choksi, Thomas Hikaru Clark, Alicia Chen, Ria Das, Ian Griffith, Aixiu An, Anna Ivanova, Moshe Poliak, Peng Qian, Sam Hutchinson, Michal Fux, Charlie Shvartsman, Nikos Moustroufis, Alex Berne, Anna Lindner, Dünya Bulut, Anja Bradić, Jana Bebek, Anita Shanker, João Barateiro, Twan Wormgoor, Chris Grevener, Sandra Nolvi, Harry Briffa, Matthew Blackie, Alex Scarr Hall, Ori Shaked, Lukas Schuecker, and Orestis Simakou. You are the people I turn to when I go for random, niche expeditions, when I share things that no one else cares about, when I need advice, and when I want to party. All the gossip sessions at 5am, dinners we've had, food we've cooked, drinks we've invented, trips we've made, 6h long monthly zoom calls we have had, and the BCS-specific lingo we've implemented, will all be forever remembered by me. Thank you all so much for being such great friends. I would also like to thank Sofia Skopetos for adopting me into the Greek culture. Ευχαριστώ πάρα πολύ για την ευγένεια και την φιλοξενία σας!

I feel privileged to be part of the broader Building 46 community, supported by an amazing team who keeps everything running smoothly. I'd like to thank Julianne Ormerod for all the support the moment I submitted my BCS application until now. I would also like to thank Jamie Wiley and Sierra Vallin for their help when I was TAing and when I was recruiting UROPs. I also want to thank the two custodians in the building who I talk to every day and night: Diego Arango and Diana Patino, along with their coworkers, for keeping our workspace clean and shiny.

I would also like to thank many artists whose music has been a constant companion throughout my graduate school: Nikos Oikonomopoulos, Anastasia Dimitra, Petros Iakovidis,

Natasa Theodoridou, Katerina Lioliou, Nikos Vertis, Giorgos Sabanis, and Željko Joksimović. Their voices have filled many hours of coding, debugging, and writing. In fact, much of this thesis was written with their music blasting in the background.

Last but certainly not least, I would love to thank my parents, my extended family, my grandma, my aunts and uncles, my cousins, for taking great care of me, and for always supporting me. I know I can always count on them when I make major decisions of my life, including deciding to go to grad school and move to a different country for a new job. When I was young, I grew up with my maternal grandparents a lot. They were both teachers, and they inspired me to be curious about everything, especially about knowledge, and here I am in grad school. Sadly, they both passed away, but I'd like to acknowledge their influence in me, for making me who I am today.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The use of language to convey information is ubiquitous to ordinary human existence: language is used to socialize, to give and receive instructions, to learn about current affairs, and to share thoughts and experiences across space and time. The communicative function of language [1] thus enables knowledge to accumulate across individuals, and is thought to be a key component to the evolution of human society and culture [2].

Language conveys information in many diverse ways. For example, speakers of sign languages communicate with each other using hand signs. Inhabitants of La Gomera in the Canary Islands use a special whistle language called Silbo Gomero, which enables them to talk to each other while being miles apart. While most of us convey and process information through speaking and writing, how we achieve this also varies. Some languages utilize only two vowels, such as Buwal [3], whereas others can have up to 14, such as Danish [4]. Some languages divide up nouns into many classes, such as Swahili [5]. Some languages put the verb first in a sentence, such as Irish [6], while others put the verb last in a sentence, like Turkish [6]. This prompts us to ask: what affects the range and diversity of human language?

Researchers have examined many different factors affecting language diversity. One class of such factors is the type of interactions between humans and their physical environment. For example, climate conditions have been proposed to influence the sounds that languages prefer to use. Languages spoken in colder or drier areas are less likely to develop complex tones [7] or to include more vowels in their phonetic inventory [8], since aridity hinders vocal cord movement, which is necessary to articulate these sounds. Languages spoken in areas with dense vegetation are found to favor sounds characterized by lower frequencies, as higher frequency sounds are more likely to be absorbed by the surroundings and hence fail to reach the listeners [9]. As a second example, subsistence level has also been studied: the development of agriculture alters human bite configuration, making certain sounds easier to produce and therefore increasing their prevalence across languages [10,11]. A second kind of factor beyond interactions with the environment involves how interactions occur among humans. For example, languages spoken by more people or with more L2 speakers are suggested to have simpler morphology [12–14] (but cf. [15]). And vocabulary size in semantic domains is suggested to be influenced by the communicative need: societies where colors are more useful in distinguishing objects tend to develop more fine-grained color categorization [16], and societies where expressing exact quantities is not necessary may not have number words [17].

In addition, because language is a tool for communication [1], an influential view is that the structures of language are shaped by pressures towards efficient communication (See [18] for a review), where the notion "efficiency" has been used in two kinds of ways. One sense of efficiency is to partition a meaning space into words such that language users can convey meanings accurately with only a few distinct words to memorize. Many studies have been conducted on how different languages label colors (e.g. [19–22]). For example, Zaslavsky et al. [21] formulated efficiency as optimal trade-off between two competing pressures: informativity and simplicity. A maximally informative color system requires a unique name for each humanly distinguishable color, which will result in millions of color words to memorize and use. Meanwhile, in a maximally simple color system, all possible colors are labeled by the same word, which is not informative at all. Zaslavsky et al. [21] showed that attested color systems achieve maximal informativity, given a complexity. In addition to color systems, studies applying such a framework on other semantic domains also yielded the same results: kinship systems [23], personal pronouns [24], tense [25], modal verbs [26], and numerals [27].

According to the second sense of efficiency, language users minimize production and comprehension difficulty. One metric of processing difficulty is the length of syntactic dependencies [28]. The meaning of a sentence is formed by its syntactic dependencies: the connections between words to give rise to compositional meanings. Dependency length measures how far apart words are when they combine to form meanings. It is easier for the producer and comprehender if the words that combine meanings are close together, as longer dependencies add to production and comprehension cost [29]. Indeed, many studies (e.g. [28,30–34]) have shown that languages prefer structures that minimize dependency length.

How does communicative efficiency emerge? It has been proposed that efficiency emerges from an organic process, as if languages were shaped by an "invisible hand" [35]: a language user chooses the less costly form if multiple forms can express the same meaning [36]. For example, in a context where a long word is predictable from context, speakers tend to articulate a shorter form of the word, such as saying "math" instead of "mathematics" [37]. If many language users make the same choice, the less costly form may become conventionalized in a language. Iterative-learning experiments have been used to study how features of languages get conventionalized in artificial languages (e.g. [38–40]). Participants in [40] were asked to learn an artificial language where both word order and cases were required to mark the agent and the patient, but after several iterations, participants became less likely to produce case markings, since word order alone was already sufficient to convey the information.

The mechanism towards efficient communication is organic: each speaker does not have the overall goal of maximizing communicative efficiency in their mind. It is also individual: each speaker independently favors a less costly form when it conveys a similar meaning. However, we are not just a collection of individuals; instead, we live in societies, which are kept in order by various institutions. These institutions operate in a top-down manner, and each of them has an explicit goal. For example, in the US, the Internal Revenue Service was established to collect taxes, and the Environmental Protection Agency was set up to protect the environment. They also operate beyond an individual level, by establishing rules that people are expected to follow, regardless of their preferences. Language is one domain where the two mechanisms can come into conflict, as there are language users spontaneously choosing how to express their meanings, as well as institutions that may prescribe how people

speak, according to objectives unrelated to communicative efficiency.

In this thesis, we ask: to what extent can the bottom-up mechanism towards communicative efficiency withstand the top-down institutional pressures? We focus on the institutional pressure implemented in the form of policies. Many policies are being implemented in order to make rules in various aspects of our lives. For instance, some policies are intended to lay out how much tax should be collected from everyone in order to pay for public services. Notably, there are policies that have been carried out that have impact on how languages are used. For example, after the French Revolution in 1789, the dialect of French spoken in Paris was made the standard French, which became the language of education, in place of other languages spoken in France such as Occitan. A second example comes from Türkiye in 1928, where a reform was implemented to introduce a new writing system, and as a result, Turkish changed from being written in an Arabic-based script to a Latin-based script. While many studies have been conducted on how language policies interact with factors such as literacy rates (e.g. [41]), it still remains unclear on how such policies can potentially affect the communicative efficiency of a language. After policies are implemented, can language users still continue to find less costly ways to express the same meaning, or will the policies restrict their abilities to do so?

Specifically, here we study two policies that affected how language is used. The first policy regulates personal names. Historically, across cultures, personal names were flexible and context-dependent: the same person "John" may have been called "John Smith" or "John Short" so that he could be distinguished from other John's who were bakers or were tall [42,43]. Later, in order to better collect taxes and account for properties, governments in the West passed laws to make the second part of the name hereditary: for example, the descendants of "John Smith" all carried the name "Smith" as their second part of the name, regardless of their occupations. The second policy attempts to simplify legal language. Throughout history, laws have been subject to numerous complaints on how hard they are to understand [44–48]. As a result, many governments have tried to make laws easier to understand, such as the British Parliament in the 1850s [44] and the U.S. government in the 1970s [47].

How have these policies affected the communicative efficiency of language? This thesis provides three preliminary results. First, in Chapter 2, we investigated a baseline case: how systems of **spatial demonstratives** (i.e. "here" and "from there") vary across languages by pressure towards communicative efficiency. Some languages have simple systems, only differentiating distances; in contrast, other languages have more complex systems, differentiating both distances and orientations (e.g. "to", "at", "from" a location). Using tools from information theory and a database of 200 genealogically and geographically diverse languages, we show that among all the theoretically possible spatial demonstrative systems, those adopted by real languages lie near an efficient frontier, balancing between two competing communicative pressures: informativity and complexity. Moreover, languages also vary in the relative weight they place on informativity and complexity.

Then, in Chapter 3, we investigate how regulations on **personal name systems** may have affected the information structure of names. Despite differences in naming traditions in East Asia and the West, using census records and historical datasets, we show from an information-theoretic perspective that East Asian names and Western names historically shared a similar structure: the name appearing first (e.g. "Chen", "Franklin") is less informative than the name appearing second (e.g. "Sihan", "Roosevelt"). This way enables a large number of individuals

to be distinguished, without needing a large number of names. Such information structure also allows full usage of the communication channel. However, regulations in these societies transformed different parts of the names into hereditary patronyms: the first name in East Asia and the second name in the West. We show how these changes altered such efficient information structure of names in order to keep distinguishing a large number of individuals possible. In scientific publishing, following the Western conventions, authors are indexed by their given name initials and hereditary patronyms. While such a practice makes sense for Western authors, it undermines the distinguishability of East Asian authors by compressing the more informative given names into initials.

Finally, in Chapter 4, we study the effectiveness of government efforts to simplify **legal language**. American legal texts are notoriously difficult to understand, largely due to complex sentence structures. What is the origin of such complexity? Is it possible to make American laws simpler to read? In this chapter, we hypothesized the complex meanings that laws need to convey naturally lead to complex grammar. To evaluate this hypothesis, drawing on corpora from the 7th century Tang Code, we examined linguistic complexity of legal texts in ancient China, which were created independently of the laws in Western civilizations. We found that classical Chinese legalese exhibits substantially greater grammatical complexity (longer inter-word dependencies) than non-legalese texts from the same time period, which suggests that the complexity may stem from the complexity of the intended meanings. Then, we analyzed linguistic complexity of legal texts in modern China, which went through a drastic change from ancient China in legal traditions in the early 20th century. Our analysis of modern Chinese legalese texts shows that grammatical complexity is much reduced compared to both the Tang Code and to modern American English legalese. This reduced complexity is likely due to modern-day Chinese lawmakers adopting alternative drafting strategies aimed at reducing syntactic complexity while conveying the same meaning, by breaking down single sentences into multiple simpler clauses. This is possibly a result of the Chinese government's broader effort to make Chinese more accessible to the general public.

Chapter 2

An information-theoretic approach to the typology of spatial demonstratives

Note This chapter is adapted from the following publication (published under a CC-BY-4.0 license):

Chen, S., Futrell, R., & Mahowald, K. (2023). An information-theoretic approach to the typology of spatial demonstratives. *Cognition*, 240, 105505.

Abstract We explore systems of spatial deictic words (such as 'here' and 'there') from the perspective of communicative efficiency using typological data from over 200 languages [49]. We argue from an information-theoretic perspective that spatial deictic systems balance informativity and complexity in the sense of the Information Bottleneck [21,50]. We find that under an appropriate choice of cost function and need probability over meanings, among all the 21,146 theoretically possible spatial deictic systems, those adopted by real languages lie near an efficient frontier of informativity and complexity. Moreover, we find that the conditions that the need probability and the cost function need to satisfy for this result are consistent with the cognitive science literature on spatial cognition, especially regarding the source–goal asymmetry. We further show that the typological data are better explained by introducing a notion of consistency into the Information Bottleneck framework, which is jointly optimized along with informativity and complexity.

2.1 Introduction

When Shakespeare's Hotspur says "Whither I go, thither shall you go too," he's using *whither* as an interrogative meaning "to where" and *thither* as a spatial demonstrative meaning "to there." When Edgar in *King Lear* says "Men must endure / Their going hence, even as their coming hither," *hence* means "from here" and hither "to here." This suite of spatial demonstrative (*here*, *hither*, *hence*, *there*, *thither*, *thence*, *where*, *whither*, *whence*) is largely lost in modern English—aside from *here* and *there*, some specialized use cases, and some fossilized expressions like "hither and yon" and "henceforth." English speakers now use *there* both to refer to a static location and for the "to" directional meaning. That is, one says "I

was going there," not "I was going to there." In effect, the old meaning of *thither* has been entirely subsumed under the auspices of the word *there*.

Spatial demonstrative systems are a source of cross-linguistic variation [49,51–53] and have been studied as part of a broader body of work exploring how cross-linguistic variation affects, and is affected by, spatial cognition across cultures [54–60]. Some languages have more complex spatial demonstrative systems than others. Whereas English has only two levels to reflect distance in spatial demonstrative (*here* and *there*), Spanish has *aquí* (here), *ahí* (there, close by), *allí* (there, medium distance), and *allá* (there, far away). Other languages, like Malagasy, have systems that draw distinctions between spatial locations in which an item is visible or invisible. For instance, the suffix *-èto* is used to mean "here" when the object is visible but *-àto* when it is invisible. Moreover, some languages define spatial words in terms of landmarks (mountains, rivers, etc.), whereas others define space in terms of people (e.g., *here* means near the speaker and *there* means near the hearer), whereas others define spatial words in reference to a hypothetical "deictic center" [51], meaning what constitutes *here* and *there* varies based on the imagined center of the particular discourse.

Why should languages converge on similar solutions? At the same time: why should some languages have more complicated spatial word systems than others? We argue that this convergence is an instance of a general functional pressure for efficiency in language [18,61]. We argue that as in other semantic domains [e.g., 19,22,25,62–65], there is a tradeoff between complexity and informativity. This tradeoff can be quantified using the information bottleneck [22,50,66,67]. Drawing on a typological database of spatial demonstratives across languages [49], we undertake an information-theoretic analysis of the cross-linguistic variation of place demonstrative systems across 5 major world regions.

To study the communicative efficiency of spatial demonstratives, we have to make a number of assumptions about both the world and about language users. Specifically, we have to estimate quantities like: How often do people refer to items that are close as opposed to far away? How often do people talk about where things were, as opposed to where they are going? How costly is it to confuse a word like "here" with a word like "there", as opposed to confusing a word like "here" with a word like "hence" ("from here")?

Here, we ask whether spatial demonstrative systems are optimized relative to statistical baselines and, if so, what assumptions must hold for that to be true. To fully characterize spatial demonstratives of the world's languages, we show that we must account for something not previously dealt with in information bottleneck work: a preference for consistency in paradigms. Introducing a notion of consistency, we show that a number of plausible, communicatively efficient systems are ruled out because they lack the consistency that characterizes real-world systems. We also show that, in an efficiency-based theory, the strong source/goal asymmetry found in many natural languages does not appear to originate from usage frequency, but rather from a stronger communicative penalty for confusing source words with place words than for confusing goal words with place words, this pattern emerges.

2.2 Background

2.2.1 Background on spatial demonstrative

The elements of language that we are concerned with are the class of deictic expressions used for space: meanings like HERE and THERE—which are closely related to deictic spatial interrogatives (e.g., where/whether/whence) and to deictic demonstratives (e.g., this/that/these/those) [68–81]. While there has been considerable work on and debate about what it means to be a deictic expression, it is generally agreed that deictic are sensitive to context and involve joint attention between the communicators [74,75].

Although these words may be naively interpreted in terms of referring to configurations in actual physical space, a variety of evidence from discourse analysis and experiments [69,73,82] suggests that they rather refer to positions within a subjective space defined by a particular discourse and in particular the body of the speaker (but see [83] for arguments against the body-centered view, a position going back to [81]). An important part of this position is that attention in a physical scene (e.g., eye gaze, pointing) has a major effect on the deictic expressions used [84,85]. And, when physical expressions are less available, deictic language is affected [85–87].

This body of work has also shown a particularly prominent boundary between the more proximal levels and all other distal levels, perhaps because the more proximal language refers in particular to that which is within reach [69,73,88,89].

For our purposes, we use a notion of distance to formalize the meanings underlying deictic words. But this distance need not be thought of as a purely physical distance, but can instead represent the subjective distance between distal levels.

Following [52] and [49], we define the notion of "spatial deictic demonstratives" as expressions that encode relative spatial properties. Many of these expressions are adverbs (e.g. *here*, *there* in English; *odavde* in Serbo-Croatian), but can also be adpositional phrases (e.g. *from here* in English; *cóng zhè lǐ* in Mandarin Chinese). The spatial relation encoded in these expressions can be divided into 2 dimensions: **distal level**—the distance of the referent with respect to the speaker, listener, or both, and **orientation**—the relative movement of the referent with respect to that deictic level. The key orientations we consider include PLACE, in which the referent is at a given distal level (e.g., *here*, *there*), GOAL, in which the referent is moving towards a distal level (e.g., *hither* and *thither* in archaic English), and SOURCE (e.g., *hence* and *thence*), in which the referent is moving *from* a distal level. We will discuss each of these 2 dimensions in detail below.

In Table 2.1, we show examples for two languages (English and Maltese) which, despite having very different words, have similar deictic systems. Both Maltese and English share the same strategy to partition the 3-by-3 meaning space using these terms: they use one term to represent D1-PLACE and D1-GOAL (the word 'here' in English), one term for PLACE and GOAL in both D2 and D3 (the word 'there' in English), one term only for D1-SOURCE ('from here'), and one term for SOURCE in both D2 and D3 ('from there').

Variation of distal levels in demonstrative systems Much work on typological deixis [49,52,68,90] draws a distinction between person-oriented and deictic-center-oriented systems. A deictic-oriented system posits a deictic center around which the conversation is centered

Table 2.1: English (left) and Maltese (right)) spatial demonstratives

	goal	place	source
D3	there	there	from there
D2	there	there	from there
D1	here	here	from here

	goal	place	source
D3	hemm	hemm	minn hemm
D2	hemm	hemm	minn hemm
D1	hawn	hawn	minn hawn

and that which is proximal is proximal to that center. Other systems are based more clearly on the position of the listeners. For instance, in Tagalog, "dito" is used when the referent is at a position near the speaker; "diyan" is used if the referent is at a position near the listener; and "doon" is used when the referent is far from both the speaker and the listener. [74] complicates this picture, showing that systems with more than 2 deictic levels can have more complicated relationships with the position of the speaker and listener.

Moreover, some languages (e.g., Khwarshi, a language spoken in the Caucasus) use a combination of modalities. Different words are used if the referent is close by in general, close to the speaker, close to the listener, far away in general, far away from the speaker, and far away from the listener, respectively. The word also takes different forms depending on the grammatical genders. There are various other ways to categorize distal levels based on spatial and geographic features, such as altitude, upstream/downstream of a river, and visibility. We leave it to future work to incorporate speaker/listener systems and more geographic-based systems into this modeling framework.

For simplicity, we focus our analysis here on deictic-centric systems as categorized by [49], in which spatial demonstratives are used relative to an imagined deictic center. Note that, by shoehorning all such systems into the same modeling framework, we are necessarily collapsing over meaningful variation that exists among languages which are "deictic-centered". In future work, we believe it will be possible to more richly incorporate speaker/listener systems into this framework, as well as variation among languages on these dimensions.

Orientation: Place, Goal, and Source In these spatial demonstrative systems, the form most likely to be formally marked is the SOURCE form (e.g., "from here", "from there"). When there is syncretism, it is most likely to be between PLACE and GOAL, or between all three. This pattern is not limited to just spatial demonstratives: across domains, languages tend to be more likely to mark sources than goals [91–95]. [96] gives a diachronic overview of how this played out historically in Greek, where goal markers were lost earlier than source markers. This pattern may fall out of a more general asymmetry between movement from a source and movement towards a goal.

There is robust evidence of a relationship between linguistic spatial reference systems and cognitive ones [51,54,56,97–102]. Thus, the linguistic distinction is likely related to the fact that humans have an overall cognitive bias towards goals, as opposed to sources. They describe goals with more fine-grained distinctions [103,104], and, in experiments, are more likely to focus on goal-directed movement than source movement [92,105,106]. Children in particular seem to display a strong goal bias [107] and overextend goal-directed meanings (e.g., assuming "weed the garden" means putting weeds into the garden, even when that contradicts evidence from world knowledge [108]). Children also seem to produce GOAL

markers earlier than SOURCE markers (See [108], for Greek and English, [109] for Hungarian, and [110] for Hebrew).

Nikitina [93], drawing on cross-linguistic evidence, suggests that "the meaning of Goal seems to be 'closer' to the meaning of Place than to the meaning of Source." In our work, we operationalize this by placing PLACE, GOAL, and SOURCE on a line, with GOAL and SOURCE as endpoints and with PLACE in the middle; we find that this configuration gives the best fit to the typological data. But, as Nikitina [93] notes, there is an asymmetry. Do et al. [111] show that, although goals are conceptually privileged, the frequency of source mentions increases when source is not in the common ground. Thus, there are likely pragmatic/communicative factors that underlie the source/goal asymmetry, while Chen et al. [112] points out that the asymmetry might be due to an online attention bias, because it seems both SOURCE and GOAL are encoded in memory.

Does this mean that there is, in general, a greater penalty for confusing source words with place/goal words than for confusing place and goal? Or does the pattern fall out of the empirical need probability of discussing source events (which are, overall, less likely than goal events)? This is a question we address using the Information Bottleneck approach: whether the observed distribution of spatial adverb paradigms across languages can be explained merely by the prior probability of the various categories or whether there is evidence for a cognitive cost that differentially penalizes confusing source words with place words.

2.2.2 Past work on the efficient structure of semantic spaces

A large body of work has explored the way that languages efficiently break up semantic categories. This work typically quantifies a trade-off between complexity and informativity—earlier by measuring complexity and informativity explicitly [20] and, more recently, through the information bottleneck [22,50,66,67]. There is strong evidence that languages efficiently navigate this tradeoff by being maximally informative, given the complexity of the system.

One rich domain for these explorations has been color words [16,19,21]. Some languages have more color words than others, with more precise boundaries. As a result, these languages can more precisely pick out particular parts of the color space. But that precision comes at a cost: greater complexity. If languages are efficient, there should be no languages that have more complex color systems but have less precision. And, broadly, this seems to be the case.

Through typological comparison, it has been shown that there is efficient structure in the semantic spaces of kinship terms [23], numerals [27], names of animals [64], and season words [65]. There has also been work showing evidence for efficient structure in the organization of various grammatical systems, including indefinite pronouns [113], tense systems [25,63], quantifiers [114], and person systems [24].

2.2.3 Typological database

We use a database of spatial demonstratives that appears in [49]. Their work's methodology draws on [52], which explores spatial interrogatives (e.g., *where*, *whither*, *whence*) typologically. Nintemann et al. [49] report on the spatial demonstrative systems of languages across 5 major world regions (Africa, Americas, Asia, Europe, Oceania), drawing on reference grammars for the majority of evidence. For each language, they report the wordforms for each relevant

distal level and by place, goal, and source. They also annotate whether the system has simple distal levels (e.g., proximal, medial, distal) or if it further sub-divides the spatial system along other dimensions, such as visible/invisible.

Nintemann et al. [49] were interested, in part, in comparing the spatial demonstrative system to the spatial interrogative system. Because we are primarily interested in the structure of the spatial demonstrative system, we do not consider the spatial interrogative system. Nintemann et al. [49] test several hypotheses in their work, focusing in particular on syncretism in orientations: that is, whether and how different orientations are referred to by the same form.

Nintemann et al. [49] use the following schema for identifying syncretism in the spatial demonstrative system:

1. $P \neq G \neq S$
2. $P = G \neq S$
3. $P \neq G = S$
4. $P = S \neq G$
5. $P = G = S$

In the list above, an equal sign indicates that the two orientations on both sides of the sign are referred to by the same demonstrative. For instance, $P = G \neq S$ means that Place (P) and Goal (G) are referred to by the same demonstrative, but not Source (S) and Goal (G). For example, in English, 'here' can refer to both 'at here' (Place) and 'to here' (Goal), but 'from here' only refers to Source.

Nintemann et al. [49] hypothesize that languages employ the same syncretism pattern in spatial demonstratives across different distal levels. For instance, if a language has syncretism for Place and Goal at one distal level, it tends to also have syncretism at other distal levels. We will address this prediction and its relation to framework in Section 2.6.1. They also hypothesize that most languages employ one of three syncretism patterns: using the same demonstrative for all orientations (Pattern 5 above), using different demonstratives for all orientations (Pattern 1 above), or using the same demonstrative for Place and Goal but another for Source (Pattern 2 above). As a third hypothesis, they offer that there is a rise in length of forms from Place via Goal to Source.

We argue that certain of these hypotheses fall out of information-theoretic motivations and can be tested using our framework. The last of these, that construction length rises from Place to Goal to Source, falls straightforwardly out of the general relationship between frequency and length [115,116] and, because Nintemann et al. [49] show this relationship robustly in their work, we do not focus on it here. But we note that this pattern is broadly consistent with communicatively efficient patterns. We focus on two key hypotheses: that certain paradigms are much more common across languages than others and that languages prefer consistency in their syncretism patterns within spatial demonstratives. This second point, which we refer to as **consistency**, requires an additional step beyond the pure information bottleneck approach that we develop in Section 2.6.1.

Below, we re-formulate these hypotheses into two key factors that we aim to explain using an information-theoretic approach.

Orientation Syncretism Patterns As mentioned above, the third hypothesis states that Patterns 1, 2, and 5 are widely attested in world languages. It is indeed the case in [49]: it is common for languages to have a three-way split between Place, Goal, and Source (Pattern 1), common to have no split (Pattern 5), and common for syncretism between Place and Goal with Source marked separately (Pattern 2). Meanwhile, it is rare for there to be syncretism between source and goal (with place as an outlier) or between source and place (with goal as an outlier). In spite of being rare, they are indeed attested in [49]: Balese for Pattern 3, and Northern Saami for Pattern 4. We show that in the information bottleneck approach, if a language values informativity more compared with complexity, Pattern 1 emerges, whereas if a language values complexity more than informativity, Pattern 5 emerges. Pattern 2 is likely to stem from the cognitive phenomenon of source-goal asymmetry mentioned in previous sections and operationalized here as a higher penalty for confusing Place and Source than for confusing Place and Goal.

Consistency The second hypothesis states that the same syncretism pattern is likely to occur in both near distals and far distals. That is, it is unlikely (but not impossible) to be the case that a language follows one pattern (e.g., $P = G = S$) in the near distals and another pattern (e.g., $P \neq G = S$) in the far distals. We show that the information bottleneck approach, as currently proposed, does not necessarily lead to this result. Thus, in our third experiment, we propose adding a consistency component to the system. By adding a preference for consistency, this pattern emerges.

2.3 Information-theoretic formulation

We aim to model spatial demonstratives using the Information Bottleneck (IB) framework which was introduced into linguistics by [21]. The IB model has its ultimate origins in the physics literature [50] and is a special case of the general theory of lossy compression [117–119]. In this section, we review the framework and how we will apply it to our domain. We also discuss a number of extensions to the Information Bottleneck that will prove necessary to get an adequate model of the typological data.

2.3.1 Basic information bottleneck

Applied to natural language, the Information Bottleneck is a model of a communicative system: a mapping from mental representations of meaning to discrete forms. These discrete forms may be distinguished from each other by syntactic, morphological, or lexical means. On its own, the Information Bottleneck fundamentally provides a model of *what distinctions are made* without specifying *how* they are made.

The Information Bottleneck formalizes the intuition that an 'optimal' communicative system balances informativity on one hand with complexity on the other. Schematically, an optimal system should minimize an **objective function** of the form:

$$\text{Complexity} - \beta \cdot \text{Informativity}, \qquad [2.1]$$

where β is a scalar value that determines how much a unit of Complexity should be traded off with a unit of Informativity. The scalar β can be seen as a conversion factor that converts Informativity into a common currency with Complexity. The meaning of Eq. 2.1 is that languages minimize Complexity and maximize Informativity.

The Information Bottleneck allows us to give precise definitions for both Informativity and Complexity in terms of information theory. Thus, it allows us (1) to evaluate the comparative optimality of real systems, by plugging them into Eq. 2.1, and (2) to derive mathematically optimal systems to which real systems can be compared, by finding minima of Eq. 2.1. The key information-theoretic concept behind the IB framework is mutual information, defined below.

Mutual information Both the Informativity and Complexity terms in the IB framework will be defined in terms of **mutual information**: a statistical quantity that gives the most general measure of dependence between two random variables [120]. Given two random variables X and Y, the mutual information between X and Y is an average log-likelihood ratio:

$$\mathrm{I}[X : Y] = \sum_x \sum_y p(x, y) \log \frac{p(y \mid x)}{p(y)}, \qquad [2.2]$$

where the x and y are possible values of the random variables X and Y respectively.

Intuitively, Eq. 2.2 quantifies how much the uncertainty about Y decreases when you know the value of X. When X and Y are independent—such that knowing X gives no information at all about Y—their mutual information is zero. The more dependent they are on each other, the higher their mutual information. As we will see below, mutual information admits a number of different interpretations, allowing it to serve both as a measure of Informativity and Complexity when applied to different variables.

Mathematical setup Following the convention of [21], we define the information bottleneck using three random variables:

1. U, a random variable over **world states**. In the case of deictic demonstratives, a world state is a pair $\langle r, \theta \rangle$ of a distal level r, consisting of one of a set of D discrete distal levels, and orientation θ, consisting of one of PLACE, GOAL, and SOURCE. Thus, there are $3 \times D$ world states. In this work we set the number of distal levels to $D = 3$.[1] The world state may refer to an objective physical space, or to a shared subjective space within a discourse.

2. M, a random variable over **meanings**: mental representations of world states. Each meaning corresponds to a distribution on world states, parameterized as below in Section 2.3.2. We assume that the relationship between world states and meanings is fixed and independent of language; presumably it is set by perceptual systems that map between world states and mental representations.

[1] In order to model actual world states, the distal levels r should likely vary continuously. We use a discrete number of distal levels for mathematical tractability.

3. W, a random variable over discrete **words**. A **communicative system** q consists of a stochastic mapping from meanings to words:

$$q : M \to W,$$

or equivalently a conditional distribution $q(w|m)$ on words given meanings. The Information Bottleneck allows us to derive optimal systems q for a given M and U.

Following previous work [21], we assume that the 'need distribution' on meanings $p(m)$ and the conditional distribution on world states given meanings $p(u|m)$—which reflects the perceptual relationship between cognitive meanings and world states—are fixed across languages. In reality, different cultural or geographic constraints may mean that the need distribution is not fixed across languages. Indeed, at least in the domain of color, studies such as [121] do suggest the need distribution varies across languages, and such variations seem to be related to geographic location and ecologic region. We leave it to future work to incorporate these differences into this approach.

Given this setting, the IB optimality of a system q with respect to meanings M and world states U is the difference of mutual information.

$$J_{\text{IB}}[q] = \underbrace{\text{I}[M : W]}_{\text{Complexity}} - \beta \cdot \underbrace{\text{I}[W : U]}_{\text{Informativity}} . \qquad [2.3]$$

In the equation, the mutual information between words and world states $\text{I}[W : U]$ plays the role of Informativity, while the mutual information between *meanings* and words $\text{I}[M : W]$ plays the role of Complexity. Below, we review the meanings and motivations for these terms.

Motivation: Informativity The Informativity term uses the interpretation of mutual information as quantifying the amount of information contained in one variable about another. In this case, it gives the amount of information in the word W about the world state U. This interpretation is valid because mutual information quantifies the average reduction in uncertainty about the world state U that happens upon observation of a word W.

Motivation: Complexity The complexity of a system, on the other hand, is defined using the mutual information of *meanings* and words. In its appearance here, mutual information represents the complexity of the *mapping* between meanings and words. It can be interpreted the number of distinctions about meaning encoded in the system.

Mutual information has been used in this sense in neuroscience as a general measure of complexity for action policies, that is, mappings from states to actions as implemented by agents (See [**lai2021policy**,122] for a review). It correlates with empirical measures of cognitive effort [123], and plays a role in models of the complexity of language production [124]. Under this measure, a communicative system has complexity 0 when all meanings are mapped to a single word—in that case, no computation at all is required to specify the word. A system has maximal complexity when each meaning is mapped to an individual unique word (See Figure 2.1 for an illustration).

In the plain Information Bottleneck, complexity is quantified only by mutual information between meanings and words. However, it is possible that a more complete theory of

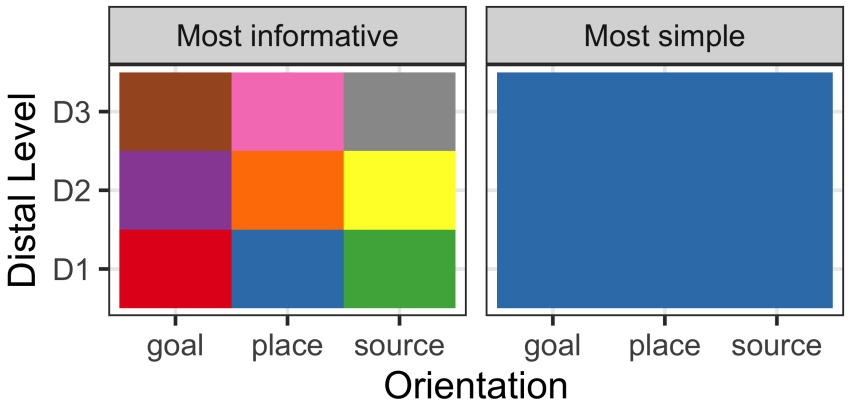


Figure 2.1: An illustration of the complexity-informativity tradeoff. The horizontal axes indicate the orientation (PLACE/GOAL/SOURCE distinctions. The vertical axes indicate the distal level. Each color represents a word. **Left**: a system with maximal informativity. In this case, each meaning has its own unique word. For a listener, there will be no confusion on what a speaker means when she utters a word. However, this system is also maximally complex, as it contains a total of 9 words. **Right**: a system with maximal simplicity (or minimal complexity). In this case, all 9 meanings are expressed by only 1 word. This system is the simplest but also the least informative.

communicative systems in natural language will require some more elaborate notion of complexity. Below, we will see that there is evidence that a full model of deictic words will require further constraints on *nondeterminism* and *consistency*, which can be implemented as an additional terms in an extended optimization objective.

Notion of optimality We study the optimality of systems where optimality is defined using Eq. 2.3. This is a **multi-objective** optimization problem, meaning that multiple notions of optimality (informativity and complexity) are being optimized simultaneously. Furthermore, since informativity and complexity are related to each other mathematically, they trade off with each other. In particular, informativity is upper bounded by complexity:

$$\underbrace{\mathrm{I}[W:U]}_{\text{Informativity}} \leq \underbrace{\mathrm{I}[M:W]}_{\text{Complexity}},$$

so it is not possible to achieve arbitrarily high informativity with a system of low complexity. Essentially, by maximizing the mutual information of words and world states while minimizing the mutual information of meanings and words, we are finding a system which encodes *only* those distinctions of meaning which are relevant for distinguishing world states.

When we plot a space defined by informativity and complexity, as we will do in Fig. 2.5, we see two regions: (1) an **achievable** region of possible systems below the black line, and (2) an **unachievable** region above the black line where there is no possible system q that simultaneously achieves the given value of informativity and complexity. The black line between these regions is the **efficient frontier**, defining a set of systems which are the best possible within the bounds of what is achievable in terms of IB optimality.

2.3.2 Parameterization

Summarizing the above, an Information Bottleneck model of a communicative system requires that we formulate two distributions: (1) a prior distribution on meanings $p(m)$ indicating how often a speaker needs to express a meaning, and (2) a conditional distribution on world states given meanings $p(u \mid m)$. We can then study the optimality of different systems $q(w \mid m)$.

Need distribution $p(m)$ We model the set of possible meanings using a three-way distinction of orientation (place vs. goal vs. source) and a three-way distinction of distance (proximal, distal, and far-distal), giving $3 \times 3 = 9$ total possible meanings.

We set the prior probabilities on meanings $p(m)$ empirically, estimating the probability of each meaning $p(m)$ from the frequencies of the corresponding words in Finnish in Lexiteria.[2] We use these word frequencies because Finnish has a full unambiguous 3×3 distinction in its deictic words. In particular, it does not have syncretism of place and goal. The resulting probabilities are shown in Table 2.2. We use Finnish data in our main analyses for convenience, but it should not be assumed that the distribution of these terms will be the same across languages as they are in Finnish. The particular choice of our prior is not crucial to our findings, as discussed below, where we compare among priors.

	goal	**place**	**source**
D3	tuonne 0.073	tuolla 0.030	tuolta 0.006
D2	sinne 0.185	siellä 0.072	sieltä 0.017
D1	tänne 0.393	täällä 0.160	täältä 0.065

Table 2.2: The need distribution $p(m)$ based on spatial deictic demonstrative frequency in Finnish, provided along with the actual spatial demonstratives.

Conditional distribution on world states $p(u \mid m)$ A world state is a tuple of a distance d and an orientation n. We model the distribution on world states conditional on meanings using **cost functions** which define a cost for confusing one distance d with another distance d', or one orientation n with another orientation n'.

The cost for confusing distances d and d' is simply the absolute value of the difference between them:

$$C_{dd'} = |d - d'|. \qquad [2.4]$$

The cost for confusing two orientations n and n' is given by three cost values C_{PG}, C_{PS}, and C_{GS} which are non-negative real numbers that define the cost for confusing place and goal, place and source, and goal and source respectively:

$$C_{nn'} = \begin{cases} 0 & \text{if } n = n' \\ C_{PG} & \text{if } n = \text{place and } n' = \text{goal or vice versa} \\ C_{PS} & \text{if } n = \text{place and } n' = \text{source or vice versa} \\ C_{GS} & \text{if } n = \text{goal and } n' = \text{source or vice versa.} \end{cases}$$

[2] https://lexiteria.com/word_frequency_list.html

Parameter	Meaning
β	Tradeoff between informativity and complexity
μ	Decay in $p(u\|m)$
C_{PG}	Cost for confusing PLACE and GOAL
C_{PS}	Cost for confusing PLACE and SOURCE
C_{GS}	Cost for confusing GOAL and SOURCE

Table 2.3: Free parameters of the information bottleneck and our model of the semantic domain of spatial demonstratives.

In addition, we have the constraint that the three cost values fall on a line (that is, the maximal cost of the three must be equal to the sum of the smaller two). We experiment with different values for the orientation costs in Experiment 2.

Finally, the costs are combined to form the probability distribution on world states given meanings using exponential decay with a decay rate parameter μ:

$$p(u \mid m) \propto \mu^{C_{dd'}+C_{nn'}}. \quad [2.5]$$

This means that a meaning m which corresponds to distance d and orientation n will give probability primarily to world states with matching d' and n', and also to other world states that are similar in distance and orientation. A sample distribution $p(u \mid m)$ under two different decay parameters is illustrated in Figure 2.2.

Summary of parameters Table 2.3 shows the free parameters of the model. The need distribution $p(m)$ has additional parameters which are set empirically.

2.3.3 Generalizing the Information Bottleneck with Further Constraints

In order to model spatial demonstratives, the simple Complexity constraint of the basic Information Bottleneck will not be enough. This is for a combination of reasons involving the way that spatial demonstrative data is coded, as well as genuine linguistic phenomena that depart from the predictions of the basic Information Bottleneck.

Determinism The Information Bottleneck generally predicts that optimal systems are **nondeterministic**: the mapping from a meaning to a word is a probabilistic function. This is an advantage in the case of semantic domains such as color words, where the mapping from perceptual space to color categories is indeed nondeterministic both across and within speakers: many speakers are unable to produce consistent color names for regions 'on the boundary' between color categories, and this variability is reflected straightforwardly in data sources such as the World Color Survey which give trial-level information on color labels provided by participants to color stimuli [125].

In the case of spatial demonstratives, however, the available data sources such as [49] provide only the mapping from distal levels and orientations to words. While in many

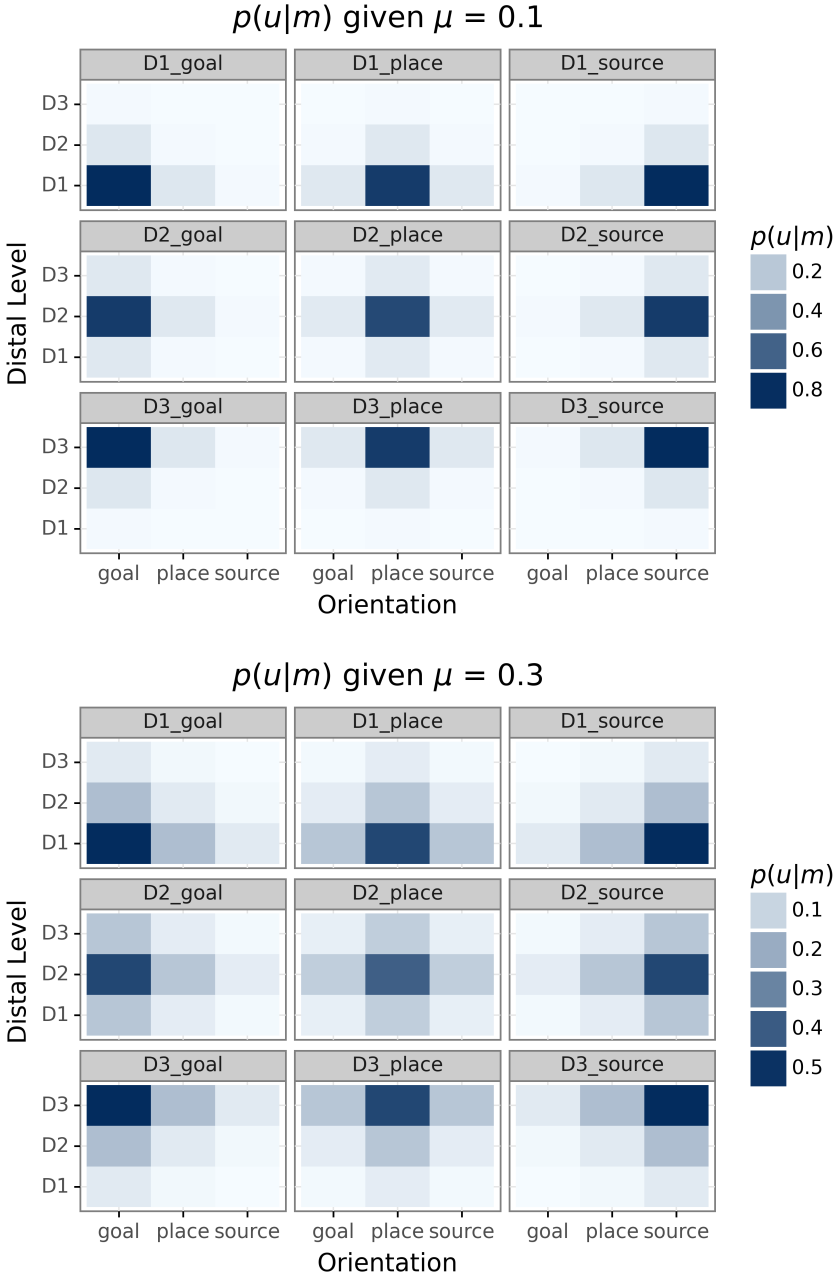


Figure 2.2: Sample distributions of world states U (in x- and y-axes), conditioned on meaning M (in each facet), when the decay parameter $\mu = 0.1$ (top) and $\mu = 0.3$ (bottom). The color represents the conditional probability, from 0 (bright) to 1 (dark). Values in each facet sum to 1.

cases this is a nondeterministic one-to-many mapping, we do not have data from which we could estimate the probability of using one particular word given one particular location described. The general IB framework does make fine-grained probabilistic predictions about how words will be used to describe meanings stochastically, but the available data do not allow these predictions to be tested. We believe that the full probabilistic structure of spatial demonstrative systems could only be investigated through quantitative experiments with a highly controlled meaning space.

In order to model the existing categorical descriptions of spatial demonstrative systems, we introduce a **determinism** constraint into the IB framework: we constrain systems to have a deterministic mapping from meanings to words. This Deterministic Information Bottleneck has been studied in the machine learning literature by [67]. When considering only deterministic systems, the Information Bottleneck objective reduces to

$$J_{\text{DIB}} = \text{H}[W] - \beta \cdot \text{I}[W : U], \tag{2.6}$$

where the Complexity term is replaced with the entropy over words produced, $\text{H}[W]$.[3]

In order to find optimal deterministic systems, we can simply iterate through all the possible mappings from meanings to words and find the ones that score the highest on J_{DIB}. The number of all possible unique systems mapping m meanings to words can be calculated as the Stirling number of the second kind (Sequence A008277 in the Online Encyclopedia of Integer Sequences, in [126])

For example, the number of possible systems given 9 meanings (3 distal levels and 3 orientations) is 21,146. We enumerate all these systems and compute their respective efficiency and informativity. Since these are all possible systems under the fixed number of meanings, the real systems will be a subset of them.

Consistency Another factor that constrains spatial demonstrative systems, but which is not encompassed in the basic IB framework, is consistency or naturalness [127]. This notion captures the idea that not all paradigms are equally easy for humans to learn and use, even if they have similar complexity. In particular, it has been shown that paradigms with various kinds of similarity in their structure (what are often called "natural patterns", as in [128–130]) are more common [131,132] and easier to learn [128,133–138], which has been posited to drive typological patterns [53,127,133,134,139–142]. These constraints reflect a kind of 'system pressure' which may be distinct from communicative pressures [143].

We operationalize these ideas as **consistency**. We say a deictic system is consistent when it has the same pattern of distinctions in each distal level and in each orientation. For example, in English, the word "here" is used to refer to both "place" and "goal" in the proximal level; similarly, the word "there" is used to also refer to both "place" and "goal" in the distal level. In addition, to indicate the orientation SOURCE, at both distal levels, the preposition "from" is used; similarly, to indicate orientations SOURCE or GOAL, no preposition

[3]The entropy $\text{H}[W] = -\sum_w p(w) \log p(w)$ represents the uncertainty in the random variable W. The IB objective reduces to Eq. 2.6 for deterministic systems because (1) the mutual information $\text{I}[M : W] = \text{H}[W] - \text{H}[W \mid M]$ and (2) for deterministic systems, $\text{H}[W \mid M] = 0$. The DIB objective can also be thought of as adding another term $\text{H}[W \mid M]$ to the plain IB objective in Eq. 2.3, thus penalizing systems with nondeterminism [67].

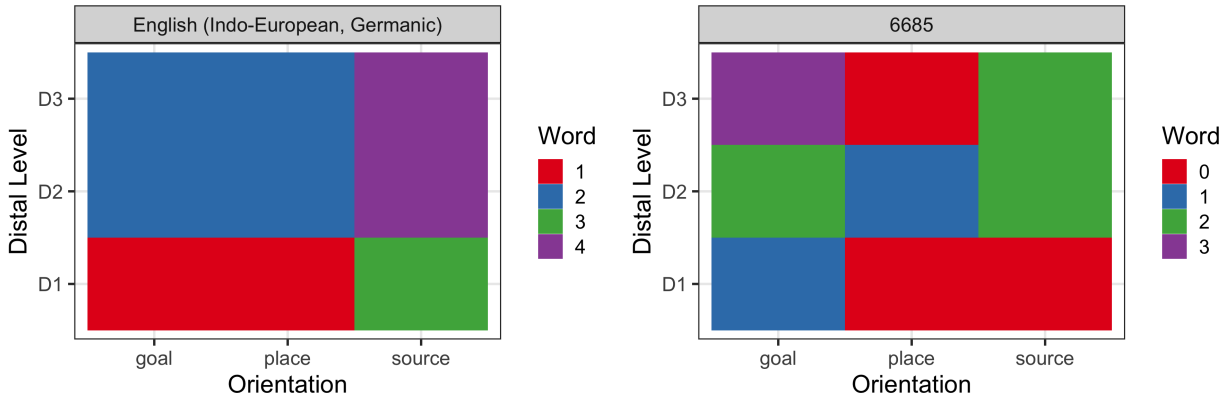


Figure 2.3: Examples of consistent (left) and inconsistent (right) paradigms. The horizontal axis indicates the PLACE/SOURCE/GOAL distinctions. The vertical axis denotes the distal level. English (left) has syncretism for PLACE and SOURCE at all 3 distal levels, whereas a simulated paradigm (right) does not.

is required. Therefore, English spatial demonstratives are consistent. In contrast, most of the enumerated deterministic systems are not consistent. An example is shown Figure 2.3: here the consistent paradigm of English is juxtaposed with a random inconsistent paradigm with the same number of words.

It is not necessarily the case that optimization of the IB objective will produce consistent systems. This is because in the IB framework, complexity of a system is measured using mutual information, which does not explicitly penalize inconsistency. We will ultimately find that the attested linguistic systems are best modeled by an extended IB optimization that includes consistency as an additional constraint.

In order to quantify consistency of real and simulated systems, we introduce a **consistency score**, defined as the sum of the number of unique SOURCE / PLACE / GOAL patterns plus the number of unique distal level patterns in a given language, similar to the enumerative complexity in [144]. For instance, since English only has the pattern of "ABB" (using the same word to refer to both PLACE and GOAL) in all deictic levels and "CDD" in all orientations, English has a consistency score of 2, whereas the simulated system in Figure 2.3 has a consistency score of 6. Hence, a lower consistency score indicates higher degree of consistency.

We use the consistency score for two purposes: (1) to quantify the actual consistency of attested system when compared to random simulated systems and to optimal simulated systems, and (2) to derive optimal systems where consistency is included as an additional constraint. The (deterministic) IB objective including consistency is

$$J_{\text{Consistency}} = \mathrm{H}[W] - \beta \cdot \mathrm{I}[W : U] + \gamma \cdot \mathrm{S}[M : W], \qquad [2.7]$$

where $\mathrm{S}[M : W]$ is the consistency score, and γ is a scalar parameter indicating how strongly inconsistency is penalized. When $\gamma = 0$, the consistency constraint has no effect. To preview the results below, we find a good fit to the attested deictic systems with $\gamma = 1$.

2.4 Experiment 1: Basic Information Bottleneck

In the first experiment, we compute the information plane for each of the real systems in our data set, as well as for simulated systems. The information plane plot reveals how close real deictic systems are to optimal systems as predicted by the basic Information Bottleneck, as described in Section 2.3.1.

We used the Appendices from [49] in order to construct a machine-readable database of place demonstratives. We only analyzed languages where spatial demonstratives are used solely relative to an imaginary deictic center (Section 2.2.1) and left out languages that employ a speaker/listener-centric system or those that adapt a combination of deictic-centric and speaker/listener-centric approach. For instance, languages such as English and Maltese are included in our analysis, whereas languages such as Japanese are not. This left 220 languages for analysis (See Figure 2.4 for the geographical distribution of the 220 languages).

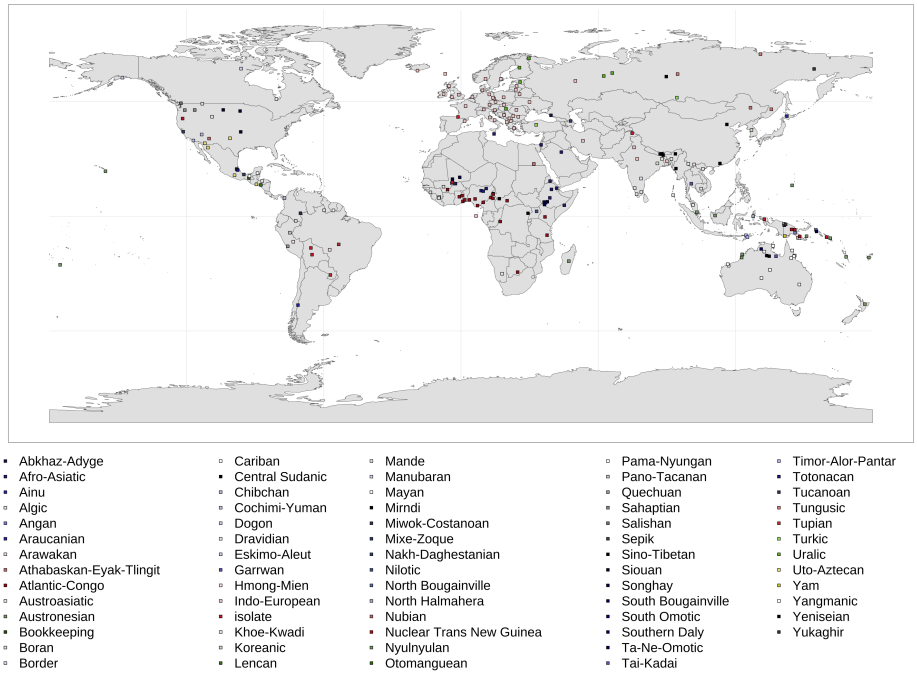


Figure 2.4: Geographic distribution of the 220 languages investigated in this study, generated in R [145]. Colors denote language families. The coordinates are taken from Glottolog [146].

2.4.1 Data processing

When there are multiple options for a particular cell, we concatenate those options together. For instance, Distal II in Tok Pisin is "longwe liklik" or "longwe" for each of place, goal, and source. When we concatenate these together, we see that there is syncretism between place, goal, and source. In Yuracaré, on the other hand, proximal Place is "ani" and Source is "ani" or "an=chi." Because the language has the option to distinguish place and source, we consider these cells as separate. Note that, for the analysis, all that matters is whether a particular cell is the same or different from another cell.

We highlight one coding choice which will have relevance for interpreting our figures: when a language has fewer than three distal levels, we assume as a convention that the second distal level encompasses both D2 and D3. This was motivated purely by the need to have a convention for the representation of two-level languages, and not motivated by any particular linguistic consideration.

Table 2.4a shows the deictic system for Tamil as presented by [49], and Table 2.4b shows the same information re-coded, as it is presented to our model: the only thing that matters is whether a particular form is the same or different to another form in the paradigm. For our purposes, distinctions of form may be syntactic (involving multiple words) or morphological.

Table 2.4: Tamil place demonstratives (a) and coded demonstratives (b)

	goal	**place**	**source**
distal	angee	angee	angeyrundu
proximal	ingee	ingee	ingeyrundu

	goal	**place**	**source**
distal	C	C	D
proximal	A	A	B

2.4.2 Choice of prior and parameters

We set the need distribution over meanings using the frequencies derived from Lexiteria for Finnish. We chose Finnish because it has canonical separation between three distal levels and between place/goal/source, with a single word associated with each. We consider a number of alternative need distributions below, and two alternative corpora besides Lexiteria in Appendix A.3.

In our main analysis in Section 2.4.3, we first set our maximum number of distal levels to be 3, decay parameter μ to be 0.2, orientation confusion cost C_{PG} and C_{PS} to be 0.8 and 1.32, respectively. These parameters are selected because they broadly reflect the properties of attested spatial deictic systems. The maximum number of distal levels are set as 3 because most of our attested systems have fewer than 3 distinct distal levels. The confusion cost C_{PS} is set to be larger than C_{PG} because of the rich literature on the asymmetry between SOURCE and GOAL, as discussed in Section 2.2.1. The specific numbers of C_{PS} and C_{PG} are fitted to the data, that is, they are the ones that place the attested systems as close to the optimal frontier as possible. The choice of decay parameter μ does not affect the results for $\mu < 0.7$, as demonstrated in Appendix A.1. Later in Section 3.2.2, we will explore the effect of each parameter separately and see how they affect the efficiency of attested deictic systems.

2.4.3 Results: Efficient frontier

In Figure 2.5, we plot the information plane for the set of 220 real systems, along with 21,146 random simulated ones. The black line indicates the efficient frontier if we allow the systems to be non-deterministic. As also pointed out in [21], allowing non-determinism extends the efficient frontier by offering additional degrees of freedom that can be used to generate more efficient systems than are available under a deterministic approach. We also calculated the expected Informativity among simulated systems, under a given Complexity, which is plotted as the blue dashed line in the figure.

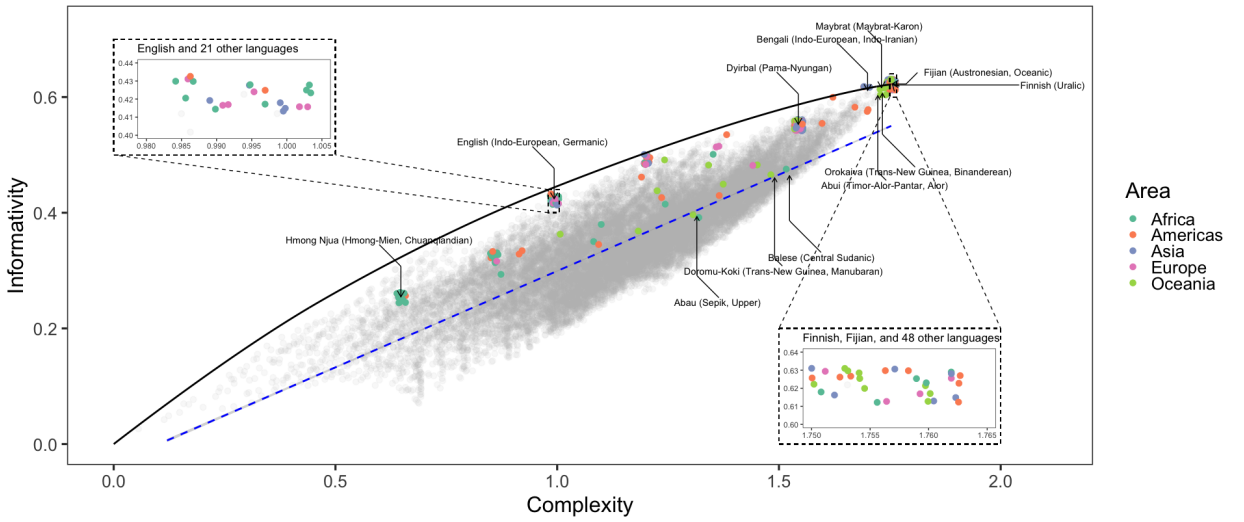


Figure 2.5: Each colored point on the information plane represents an attested deictic system, and the gray points represent all the 21,146 hypothetically possible deterministic systems. The efficient frontier is marked by the black line. The points are jittered to avoid overlap. The blue dashed line represents the expected informativity among simulated systems given a complexity. Some languages close to the frontier or far from the frontier are labeled as an illustration. Since the points are heavily clustered, a zoomed-in view of two of the clusters are presented in the two panels: spatial deictic systems sharing similar complexity and informativity of English (left) and Finnish / Fijian (right).

The real systems typically fall close to the optimal frontier, although there are some exceptions. Of the 220 systems in the sample, only 4 of them have below-expected Informativity relative to their Complexity ($I[M : W]$). These languages are Balese, Doromu-Koki, Bunoge Dogon, and Abau. Balese, the language that falls farthest from optimal on the plot, is noteworthy for being the only language in [49]'s sample in which goal/source syncretism is required. Doromu-Koki, Bunoge Dogon, and Abau all have varying degrees of goal/source syncretism, as well. As we will see in Experiment 2, the reason that these languages appear sub-optimal is that we assume higher cost for confusing GOAL with SOURCE than for any other orientation confusion, and so a language with goal/source syncretism is dispreferred.

Relative to the majority of simulated systems, what causes the apparent optimality of real systems? One major factor is that natural language deictic systems rarely have jumps in distal levels. That is, a language is unlikely to use the same word for "here" as for "far over there" but a different word for "there." From the perspective of our information-theoretic framework, that makes sense since the cost of confusing meanings which are spatially distant is high. But the cost of confusing two nearby distal levels is relatively lower. This analysis further raises the question: which of the parameters in our model are driving these relationships? In Experiment 2, we undertake a series of explorations to uncover which choice of parameters would make the real systems appear most optimal.

In addition to asking what makes deictic systems closer to the efficient frontier than the random systems, we can ask why real deictic systems are often not exactly on the efficient

frontier. We will take up this question in Section 2.6.1 where we discuss an additional constraint that appears to be operative in natural language—the constraint of consistency.

2.5 Experiment 2: Exploring factors that affect optimality

Experiment 1 found that natural deictic systems are very often closer to the efficient frontier than random systems for a particular setting of the IB model parameters. But what drives these results? Does the apparent efficiency of real systems emerge mostly from assumptions about the need distribution of the words? Or does it depend more on the cost parameters that penalize confusing, e.g., place with goal or goal with source? We take up these questions here.

In order to better characterize how different parameters affect the apparent optimality of communicative systems, we perform here comprehensive searches through parameter space, and study how the apparent optimality of systems changes under different parameter settings. Our goal here is to determine the minimal requirements on the parameters such that natural languages are well-modeled by the IB optimization.

In order to compare different model configurations, we need a metric for the optimality of a set of systems under those configurations. With such a metric, we can compare the score across different sets of parameters. We adopt the metric of **a generalized version of Normalized Information Distance** (gNID) proposed in [21]. gNID, bounded by 0 and 1, reflects the distance between an attested system and the closest optimal system.

As shown in [21], we can match each attested spatial deictic system $q_\ell(w|m)$ representing language ℓ with an optimal, non-deterministic spatial deictic system $q_{\beta_\ell}(w|m)$ by finding the trade-off parameter β_ℓ that minimizes the difference in the efficiency score between the attested system and the optimal system (Equation 2.8):

$$\beta_\ell = \underset{\beta}{\operatorname{argmin}}[(\mathrm{I}_{q_\beta}[W;M] - \beta \cdot \mathrm{I}_{q_\beta}[U;M]) - (\mathrm{I}_{q_\ell}[W;M] - \beta \cdot \mathrm{I}_{q_\ell}[U;M])]. \qquad [2.8]$$

Then, the gNID between $q_\ell(w|m)$ and $q_{\beta_\ell}(w|m)$ can be calculated by Equation 2.9, where W'_ℓ and W'_{β_ℓ} denote random variables of other possible spatial deictic demonstratives that could be produced given a meaning m:

$$\mathrm{gNID}(W_l, W_{\beta_\ell}) = 1 - \frac{\mathrm{I}[W_l, W_{\beta_\ell}]}{\max\{\mathrm{I}[W_l, W'_l], \mathrm{I}[W_{\beta_\ell}, W'_{\beta_\ell}]\}} \qquad [2.9]$$

The lower gNID an attested system has, the closer it is to the optimal frontier. For example, Kodiak Alutiiq (see Figure 2.5) has a gNID of 10^{-7}, corresponding to the fact that it lies nearly on the optimal frontier, occupying the point of maximal informativity and maximal complexity. On the other hand, Balese (see Figure 2.5) has a gNID of 0.446, corresponding to the fact that it is far away from the optimal frontier.

Using the methods described above, for each set of parameters, we compute the average gNID over all attested spatial deictic systems and then compare gNID across different sets of parameters. Below we report how varying each parameter will affect the average gNID, or the information distance to the optimal frontier, of attested spatial deictic demonstrative systems.

Factor 1: PLACE/GOAL/SOURCE Costs First, we can ask about the cost functions for confusing PLACE, GOAL, and SOURCE. Based on the prior literature, there are two major claims to consider. First, there is reason to believe that PLACE is intermediate between GOAL and SOURCE, meaning the penalty for confusing SOURCE and GOAL should be high [93]. Second, we predict that it is less costly to confuse GOAL with PLACE than to confuse SOURCE with PLACE. This prediction falls out of the cognitive science literature [103,106,147] suggesting that source-directed movement tends to be more marked.

Factor 2: Choice of need distribution In Experiment 1, the need distribution $p(m)$ is based on Finnish word frequency data. The resulting distribution exhibits two important properties: 1) the frequency decreases as the distal level increases, indicating that we tend to talk about things happening further away less, and 2) the frequency of PLACE is greater than the frequency of GOAL, which is itself greater than the frequency of SOURCE. To investigate the role of the need distribution in our model of deictic systems, we measure the average gNID among all attested systems under different permutations of the marginal PLACE / GOAL / SOURCE distribution, while keeping the marginal distal level distribution constant. For instance, suppose the marginal distribution used in Experiment 1 is $p(\text{PLACE}) = a$, $p(\text{GOAL}) = b$, and $p(\text{SOURCE}) = c$, where $a > b > c$ (henceforth denoted as PLACE > GOAL > SOURCE). One of the permutations could be SOURCE > PLACE > GOAL, where $p(\text{SOURCE}) = a$, $p(\text{PLACE}) = b$, and $p(\text{GOAL}) = c$.

2.5.1 Methods

In this analysis, we keep all other parameters identical to those in Experiment 1 and grid search different combinations of PLACE/GOAL/SOURCE confusion costs and need distribution permutations.

For the confusion costs, instead of varying C_{PG} and C_{PS}, which implicitly assumes the first claim, we represent PLACE, GOAL, SOURCE as coordinates on a 1D line. Without loss of generality, we assign the coordinate of PLACE (C_P) as 0. We vary the coordinates of SOURCE (C_S) and GOAL (C_G) in the interval $[-5, 5]$ and calculate the confusion cost C_{ij} as $C_{ij} = |C_i - C_j|$, where $i, j \in \{P, G, S\}$. Regarding the first claim, in our formulation, if C_S and C_G have opposite signs, PLACE lies in the middle between SOURCE and GOAL ("place-centric" henceforth); otherwise, PLACE is said to lie outside SOURCE and GOAL ("place-marginal" henceforth). For the second claim, if $C_{PS} > C_{PG}$, the cost setting is favoring GOAL over SOURCE, and otherwise, it is favoring SOURCE over GOAL.

Meanwhile, under each permutation of {PLACE / GOAL / SOURCE}, we compute the average gNID among attested systems. In addition, we also repeat the same analysis with two additional types of need distributions: a uniform distribution, where each meaning has an equal need probability (coded as "uniform prior"), and a distribution where we keep the decay with the distal levels but even out the need distribution within each distal level (coded as "PLACE = PLACE = PLACE").

2.5.2 Results

The results for the major qualitative categories of parameter configurations are shown in Figure 2.6. Here, the black dots represent the average gNID among attested deictic systems of a potential arrangement of (C_G, C_S), categorized by which orientation is favored and whether it is place-centric or place-marginal. The horizontal axis shows the relative coordinates of PLACE (gray), GOAL (blue), and SOURCE (red) on a number line. Overall, the results show that the IB model best fits the typological data if 1) PLACE is in the middle of SOURCE and GOAL, and 2) the cost of confusing PLACE and GOAL is set to be lower than that of confusing PLACE and SOURCE—both of which patterns have been independently observed and motivated in the typological and psycholinguistic literature.

The results of permuting the need distribution are shown in Figure 2.7. Interestingly, the need distribution we encounter in reality, PLACE > GOAL > SOURCE, is the second worst in terms of optimality and only better than a uniform distribution.

2.5.3 Which matters more: need distribution or PLACE/GOAL/SOURCE costs?

To summarize, we found that 1) when the cost of confusing PLACE/GOAL/SOURCE with one another is $C_{GS} > C_{PS} > C_{PG}$, which is the configuration that reflects the cognitive bias towards GOAL found in the literature, spatial deictic demonstrative systems attested in real languages are closest to the optimal frontier; and 2) when the need distribution of PLACE/GOAL/SOURCE is PLACE > GOAL > SOURCE, which is the configuration demonstrated in text corpora, spatial deictic demonstrative systems attested in real languages are merely second to last closest to the optimal frontier, compared with other permutations of PLACE/GOAL/SOURCE need distributions. This leaves one question open: which factor affects the average distance to the optimal frontier more? To answer this question, we take the results from the grid search and perform a Bayesian random-effects regression, using the `brms` package [148] and the formula below:

$$\text{gNID} \sim (1 \mid \text{PLACE/GOAL/SOURCE cost}) + (1 \mid \text{need distribution permutation}). \quad [2.10]$$

We use a random-effects regression so that separate variances are fit to describe the effects of need distribution vs. cost configuration.[4] If the absolute value of random intercept for one factor is greater than that for another, we can say the former matters more compared with the latter, since the gNID is affected by the former factor to a greater extent.

The results are shown in Figure A.5. In each facet, a green dot represents the case where the random intercept of confusion cost is larger than the random intercept of need distribution orientation, under a combination of the two factors. There are a total of 192,000 data points in the figure, 154,403 of which are green, indicating that in most cases, the confusion cost affects the optimality of attested spatial demonstrative systems more than the need distribution. The relatively small effect of the need distribution mirrors the finding in [21] that color naming systems appear optimal under several choices of need distribution.

[4] In the model, we set the priors as weakly informative ones, 4 chains, and 2000 iterations per chain, including 1000 iterations for warming up. To this end, for each combination of the confusion costs and the need distribution permutation, we obtain 4000 sets of fitted random intercepts for each factor.

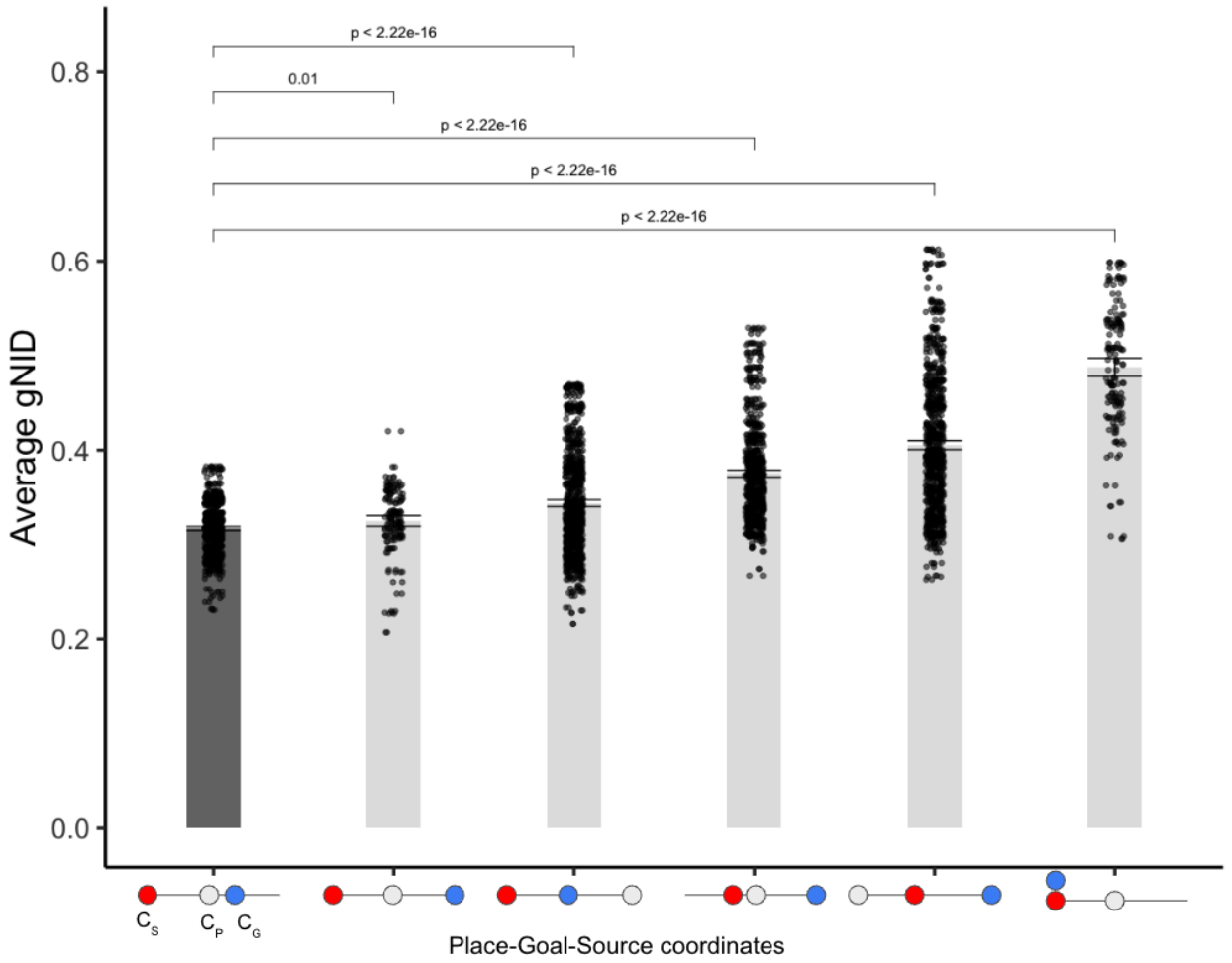


Figure 2.6: Each black point represents the average gNID of attested spatial deictic systems, in an increasing order, under a (C_G, C_S) combination, categorized by the relative position of C_G (blue dot), C_S (red dot), and C_P (gray dot) on a number line. The error bar represents 95% bootstrapped confidence interval. First column (shaded dark gray, actually observed in human languages): GOAL favoring, PLACE centric ($C_{GS} > C_{PS} > C_{PG}$); Second column: no favoring, PLACE centric ($C_{GS} > C_{PS} = C_{PG}$); Third column: GOAL favoring, PLACE marginal ($C_{PS} > C_{PG}, C_{PS} > C_{GS}$); Fourth column: SOURCE favoring, PLACE centric ($C_{GS} > C_{PG} > C_{PS}$); Fifth column: SOURCE favoring, PLACE marginal ($C_{PG} > C_{PS}, C_{PG} > C_{GS}$); Sixth column: no favoring, PLACE marginal ($C_{PG} = C_{PS} > C_{GS} = 0$). The results show that if PLACE is in the middle of SOURCE and GOAL, and the cost of confusing PLACE and GOAL is set to be lower than that of confusing PLACE and SOURCE (i.e. $C_{GS} > C_{PS} > C_{PG}$), attested systems tend to be closer to the optimal frontier. The number shows the p-value in a t-test comparing the mean for the configuration observed in human languages (dark gray) with others.

2.5.4 Discussion

Varying the parameters of the IB model, we found that the attested spatial demonstrative systems behave more similarly to optimal systems predicted by the model when the cost of

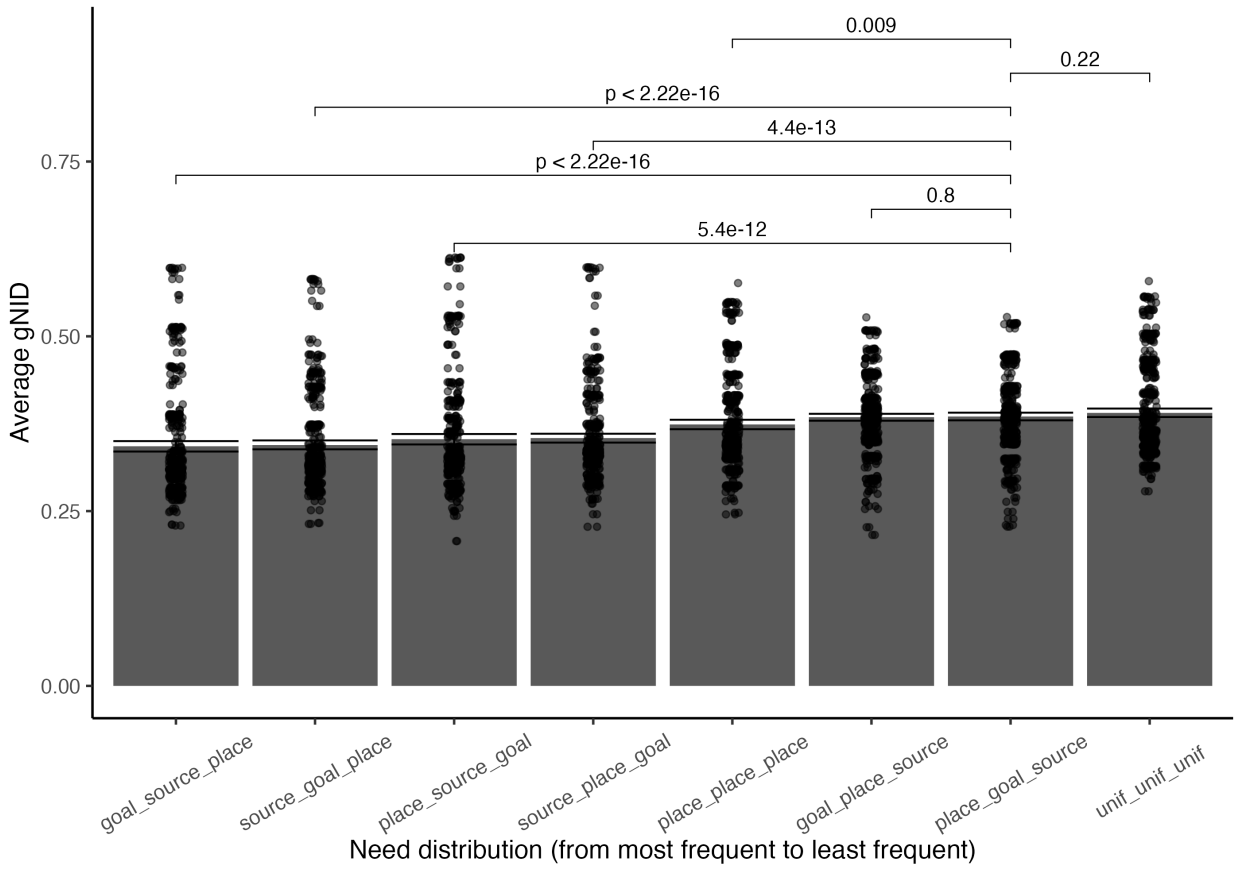


Figure 2.7: Each black point represents the average gNID of attested spatial deictic systems, in an increasing order, under each need distribution. The error bar represents 95% bootstrapped confidence interval. First column: GOAL > SOURCE > PLACE; Second column: SOURCE > GOAL > PLACE; Third column: PLACE > SOURCE > GOAL; Fourth column: SOURCE > PLACE > GOAL; Fifth column: PLACE = PLACE = PLACE; Sixth column: GOAL > PLACE > SOURCE; Seventh column: PLACE > GOAL > SOURCE (the actual need distribution); Eighth column: uniform need distribution. The numbers represent the p-value of a t-test comparing the mean of PLACE > GOAL > SOURCE with those in other configurations.

confusing SOURCE and GOAL is greater than that of confusing PLACE and SOURCE, which then is bigger than that of confusing PLACE and GOAL, a realistic constraint reported in the cognitive science literature. Although another realistic constraint on need distribution that PLACE is more frequently used than GOAL, which is more frequently used than SOURCE does not make the attested spatial demonstrative systems behave more similarly to predicted optimal systems, we demonstrate that the second constraint plays a minor role in affecting the optimality of attested systems, compared with the first constraint.

Our results also shed light on which model components are more important in terms of explaining deictic patterns. In particular, the results are relatively invariant to changes in the need probabilities (with the uniform prior performing better than the place > goal > source prior), whereas the model fit is dramatically worse when the orientation confusion costs are misspecified. This finding suggests that the reason for the goal–source asymmetry

in typology arises from asymmetries in the cost function and not from need probability.

2.6 Experiment 3: Information bottleneck with consistency

In all the simulations above, there often emerge optimal systems that are unlike real systems. These are not consistent: e.g., they have different PLACE/GOAL/SOURCE paradigms for one distal level as compared to another. Examples are shown in Figure 2.8.

There are no parameters in our model, nor in prior work on the information bottleneck, that can account for the fact that these inconsistent paradigms are dispreferred. In this section, we propose a new framework to account for this preference for consistency within the IB framework, through the addition of a new constraint.

2.6.1 Consistency

In this section, we examine whether the systems considered optimal under the information theory framework are the ones actually attested in real languages.

Although the demonstratives in different languages vary, how they are used to encode different meanings shows some commonality. For example, recall Table 2.1, which lists the spatial deictic demonstratives in English and Maltese, respectively. They partition the space in the same way: one word for proximal PLACE and GOAL, one word for distal PLACE and GOAL, and a separate set of words for conveying SOURCE information. We now call a particular strategy to partition the meaning space a **paradigm**. Each simulated system has only one paradigm, whereas many real systems might share the same paradigm. In fact, the 220 real systems in the database only utilize 34 out of the 21,146 possible paradigms. As we will demonstrate below, these real paradigms vary in their frequency.

The bottom panel Figure 2.8 shows all the simulated paradigms located on the optimal frontier. The top panel shows all 5 real paradigms closest to the optimal frontier. Among the 5 simulated paradigms on the optimal frontier, only three (21119, 21128 and 21145) are attested in our database. Simulated paradigm 21119 merges PLACE and GOAL in distal level D3 while keeping other meanings distinct, similar to Bengali and 3 other languages. Simulated paradigm 21145 gives every meaning its own word, which is the paradigm Fijian and 49 other languages adopt. Simulated paradigm 21128 gives every meaning its own word except merging SOURCE in D2 and D3. The other 2 simulated paradigms are not attested. This is to say, only a few of the possible optimal paradigms are actually adopted by real languages.

Meanwhile, if real languages are optimized in communicative efficiency, we would expect the paradigms closest to the optimal frontier to all be adopted by many languages. However, we see a clear disparity in terms of the number of languages adopting each paradigm. For instance, in the top panel of Figure 2.8, Fijian shares the same paradigm with 49 other languages, whereas Maybrat, Bengali, and other 4 languages are the only languages utilizing their respective paradigm, despite being very close to the optimal frontier. This indicates that information theory alone fails to predict why some paradigms are favored but not others, suggesting that some other factors might be affecting such preference.

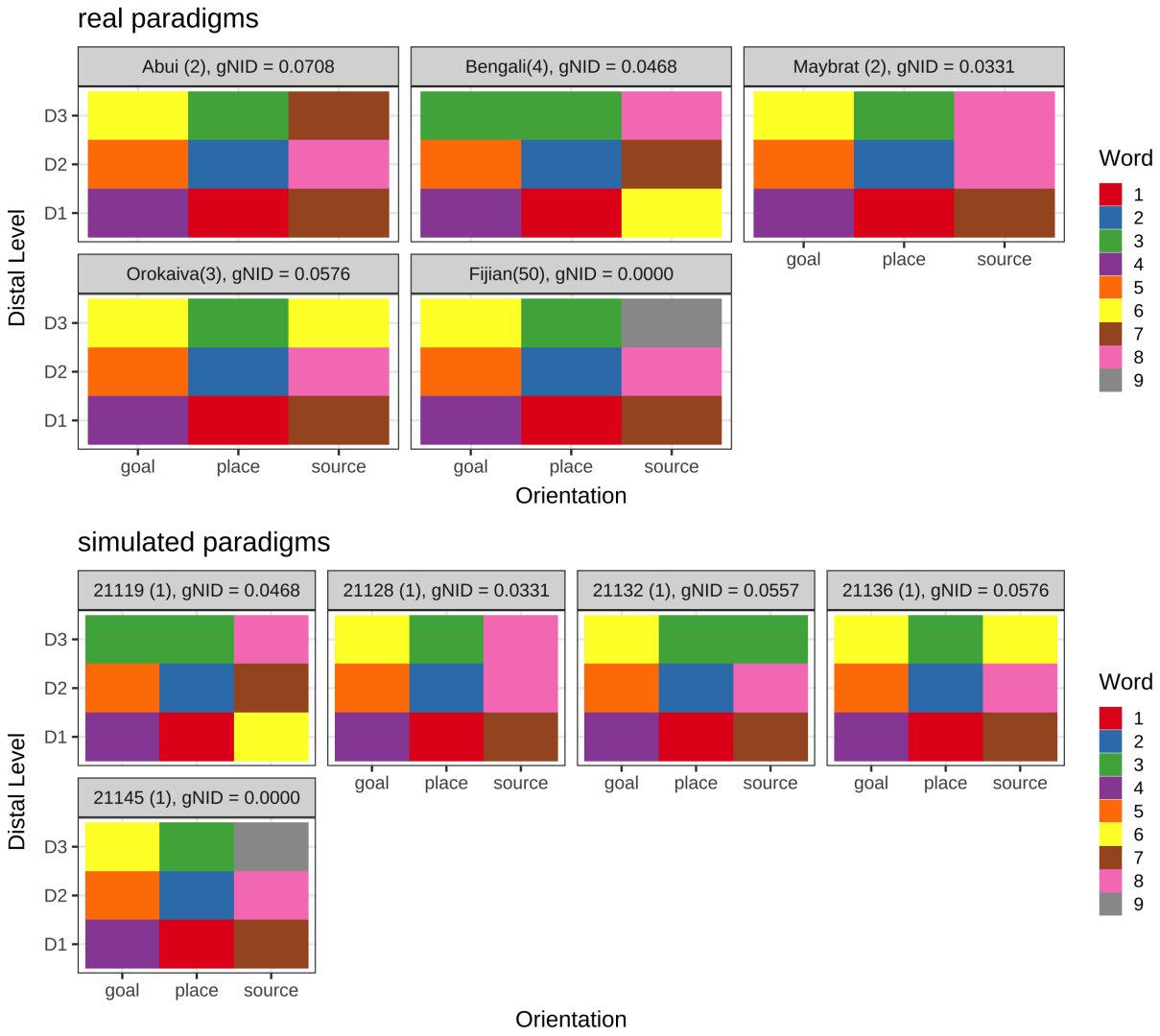


Figure 2.8: The 5 most efficient real paradigms (top) and the optimal simulated paradigms (bottom) from Experiment 1. The horizontal axis indicates the PLACE / SOURCE / GOAL distinctions, whereas the vertical axis denotes the distal level. The real paradigms are exemplified by their languages. The number in the parenthesis indicates the number of languages in a given paradigm. gNID denotes the distance to the frontier for each paradigm.

One of the additional factors appears to be consistency: the extent to which a paradigm is consistent. Consistent paradigms that are close to the optimal frontier tend to be adopted by more languages, compared with inconsistent paradigms. In this section we adopt the consistency score defined in Section 2.3.3.

2.6.2 Basic analysis

We calculate the consistency score of the real systems and the random systems generated in Experiment 1, using the same set of parameters. In Figure 2.9, we plot the consistency as well as the distance to the optimal frontier for each real paradigm and the simulated paradigm, faceted by the number of words used. The size indicates the number of languages in that paradigm. The most frequent real paradigms (shown as bigger-sized circles), utilized by a majority of the real languages in our database, fall into the bottom left corner in each facet, indicating that they tend to be consistent in addition to being efficient in balancing informativity and complexity. In contrast, infrequent paradigms (shown as smaller-sized circles) are generally located away from the lower-left corner in each facet, meaning that they are either not consistent or efficient. Meanwhile, simulated paradigms are scattered across the plot. The graph suggests that consistency, combined with the distance to the information-theoretic optimal frontier, is a good predictor of the typological frequency of real language deictic demonstrative paradigms.

2.6.3 Constructing optimal paradigms

Now we examine whether the optimal paradigms considered by both information-theoretic and consistency constraints would be utilized by most languages. To do so, we extend Equation 2.6 by adding a consistency term. The new optimality score is shown in Equation 2.11:

$$J_{DIB}[q] = \underbrace{\mathrm{H}[W]}_{\text{Complexity}} - \beta \cdot \underbrace{\mathrm{I}[W:U]}_{\text{Informativity}} + \gamma \cdot \underbrace{\mathrm{S}[W:M]}_{\text{Consistency}}, \qquad [2.11]$$

where $S[W:M]$ is the consistency score from Section 2.3.3.

Then, we search for the most efficient simulated system under each pair of (β, γ) values, where we let $(\beta, \gamma) \in [1, 10] \times [1, 10]$. The paradigms located on the new optimal frontier are shown in Figure 2.10. The optimal paradigms now resemble real systems much more closely. For example, the optimal paradigm under $(\beta, \gamma) = (5.012, 1)$, where GOAL and PLACE are merged in each distal level and the by-distal level distinction is kept, is shared by Irish and 15 other languages.

One difference between real systems and the optimal systems under Eq. 2.11 is that the optimal systems with fewer words show a clear preference in merging distal levels D1 and D2, instead of merging D2 and D3 as natural languages do. We believe this disparity stems from the way our data is coded: in particular we assumed that, if a language distinguishes fewer than 3 distal levels, the last distal level extends out to D3 (see Section 2.4.1). This can possibly be resolved in reformulating the world state, a possible direction for future research.

In Figure 2.10, we label optimal systems that differ from attested ones only by merging D1 and D2 instead of D2 and D3 with a suffix *-like*, and a total of 34 languages, including English,

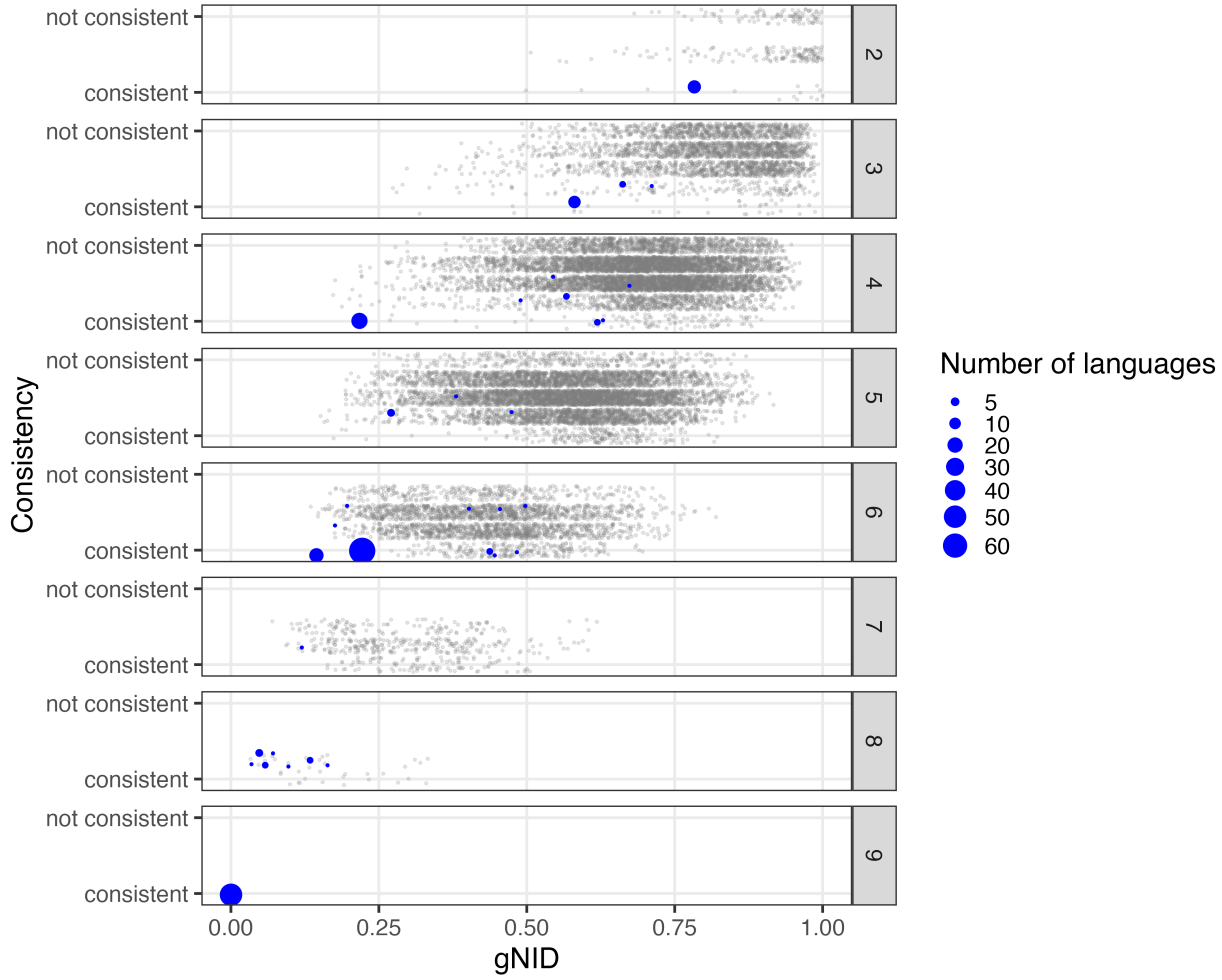


Figure 2.9: Paradigms utilized by most real languages tend to be both consistent and close to the optimal frontier. Each blue dot represents one of the 34 unique paradigms attested in the real languages, whereas each gray dot represents all the possible paradigms. The horizontal axis is its distance to the optimal frontier, and the vertical axis is its consistency (high to low). The size indicates the number of languages in each paradigm. The plot is faceted by the number of words used in the paradigm.

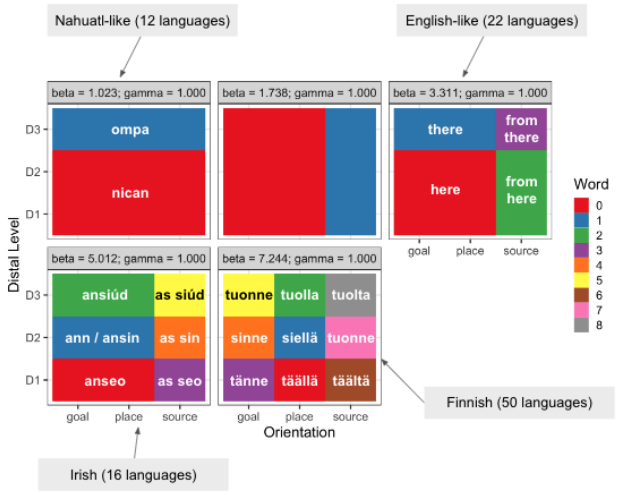


Figure 2.10: The most optimal and consistent simulated paradigms. The β and γ values on each facet indicate the smallest (β, γ) pair corresponding with that system. Gray boxes indicate an example language and the number of languages utilizing that paradigm. A "like" suffix is attached when the real paradigm only differs from the simulated paradigm by merging D1 and D2 instead of merging D2 and D3. About half of real languages fall into one of these patterns.

belong to this category. Interestingly, the paradigm under $(\beta, \gamma) = (1.738, 1)$ is not attested in any language, probably because this paradigm does not make distal level distinctions at all, while all the languages in [49] have at least 2 distal levels. This suggests that in real languages, distinguishing distal levels might be more prioritized than distinguishing orientations. Meanwhile, the most popular paradigm shared by 71 languages, where all 3 orientations are distinguished and D2/D3 are merged, is not among the optimal paradigms, since distinguishing all orientations adds to the paradigm's complexity.

2.7 Discussion

Spatial deictic demonstratives, a class of adverbs or adpositional phrases that encode spatial relations, vary both in forms and meanings they encode across languages in the world [51]. But there exist common patterns in how meanings are expressed by spatial deictic demonstratives [49]. Why are certain paradigms preferred over others? In this work, from an information-theoretic perspective [21,50,66,67], we have argued that spatial demonstrative systems in natural languages are communicatively optimized, relative to random statistical baselines. We have demonstrated that a deviation from appropriate parameter choices results in a less good fit to real languages, as measured by the residual complexity; the optimal choice of parameters reflects findings from past studies [92,93,103,104,107,111,112] that humans exhibit a strong bias towards GOAL compared with SOURCE. Then, we show that in addition to informativity and complexity within the information-theoretic framework, consistency is another constraint real languages tend to satisfy, in that real languages tend to be consistent,

in addition to balancing between informativity and complexity.

2.7.1 Why consistency?

As we have demonstrated in Section 2.6.1, most real languages have some degree of consistency in their paradigms, which our information-theoretic approach alone does not predict. In fact, languages could be more efficient by, e.g., having access to the additional degrees of freedom that come with being able to use different syncretism patterns at different distal levels.

There are several possible explanations for the consistency preference. From a learning perspective, consistent spatial deictic paradigms tend to have low Kolmogorov complexity: in other words, fewer words need to be used to describe the pattern in the paradigm [149], leading to relative ease and readiness of acquisition of paradigms in natural languages [150] and those in novel, artificial languages [53,133,134]. Broadly speaking, the tendency towards minimizing the description length of a paradigm is often reflected in historical linguistics, where, for example, in Indo-European historical linguistics, analogy seems to frequently play a role in shaping how ancient or archaic languages evolve into its modern descendants, in that irregular inflection / declension patterns are often replaced by regular ones [151]. Such tendencies have also shown in the lab: that structures gradually emerge as languages are passed down between generations [39].

2.7.2 Limitations of our consistency formulation

The consistency metric used in this work is rudimentary in that it only broadly classifies different paradigms in terms of the number of distinct distal level patterns and the number of distinct orientation patterns. However, this metric still lacks granularity, in that it fails to take the need distribution into account. For example, if a particular meaning is rarely used in everyday conversation, a lack of syncretism in this meaning should be considered more consistent than a lack of syncretism in a meaning that is frequently used. Hence, in future studies, a more frequency-based metric, preferably an information-theoretic one, should be developed to operationalize consistency.

2.7.3 What influences the evolution of spatial deictic demonstratives

Past studies have been focused on the evolution of communicative systems in semantic domains such as colors. Since it is impossible to conduct experiments on speakers from hundreds or even thousands of years ago, [152] instead investigate a rapidly-evolving language: Nafaara, spoken in Ghana and Côte d'Ivoire. They show that the language has acquired several new color terms while keeping the trade-off between informativity and complexity optimal. Spatial deictic demonstratives in languages today, albeit merely speculative, might have been through the same process of traversing along the optimal frontier. For instance, as mentioned in the beginning of this paper, English used to have 6 distinct spatial deictic words, distinguishing between 3 different orientations and 2 different distal levels. However, the GOAL demonstratives *hither* and *thither* merged with *here* and *there*, respectively, whereas the SOURCE demonstratives *hence* and *thence* were replaced with prepositional phrases *from here* and *from there*. Meanwhile, all the 3-way orientation distinctions in other Germanic

languages (such as Dutch, German, Danish, and Icelandic) are very much preserved. This is possibly because English has been extensively acquired as a second language, and the vast presence of L2 speakers leads to a simplification in the paradigm [153]. Both paradigms, as shown in Figure 2.5, are very close to the optimal frontier (the archaic English paradigm is the same as Dyirbal).

2.7.4 The continuous nature of distal levels

One limitation of our approach is that we assume that the space of distal levels is fixed and discrete, and that the mapping from meanings to words is deterministic. In reality, the space is likely continuous and word usage is likely to be stochastic. For instance, while objects that are extremely close to the deictic center are likely to be referred to by a D1 word (and never D2 or D3 words), it is likely objects that are somewhat farther away may be referred to using D1 or D2 in a way that is random or depends on the situation.

We believe that the Information Bottleneck approach could be used to make quantitative predictions about where boundaries between words would be found in this continuous, stochastic setting. Such a study would require data on the usage characteristics of these spatial words in a known spatial layout and situation, which is not currently available to the best of our knowledge.

2.8 Conclusion

Overall, we have shown that the information bottleneck can explain major patterns in the typology of spatial demonstratives. Supplementing the information bottleneck with a preference for consistent paradigms, a preference motivated by the human preference for regularity in learning and memorizing, improves the explanatory performance of the model.

Our analyses also show that there is value not just in asking whether observed features of languages can be explained by a drive towards efficiency, but by examining how model assumptions affect the ability of an efficiency-based model to fit the linguistic data. We found that using a cost function motivated by the observed source/goal asymmetry in cognition more generally led to real languages that looked more efficient. This match between cognitive-plausible model assumptions and fit to linguistic data not only provides further validation for the hypothesis that communicative efficiency drives linguistic behavior,it also suggests that efficiency-based approaches to linguistics can provide novel insight on underlying cognitive processes and can, as with our cost functions, provide evidence convergent with evidence from cognitive behavioral experiments.

2.9 Data Availability

All the materials can be found at https://github.com/mahowak/deictic_adverbs.

Chapter 3

Cross-Cultural Structures of Personal Name Systems Reflect General Communicative Principles

Note This chapter is adapted from the following publication (published under a CC-BY-4.0 license):

Ramscar, M., Chen, S., Futrell, R., & Mahowald, K. (2026). Cross-cultural structures of personal name systems reflect general communicative principles. Nature Communications.

Abstract The structure of personal names appears to differ widely across cultures. Using census records and historical datasets, we present an information-theoretic analysis of name systems that shows how the scope of this variation is more constrained than it might appear. We identify two constraints name systems must satisfy: encoding large numbers of identities, and ensuring these encodings are usable. We show that, historically, the world's languages satisfied these constraints using structurally similar, near-optimal codes. They did so by combining sets of name-specific words with existing vocabulary items, allowing unlimited numbers of identifiers to be created while keeping vocabulary sizes stable. Today, many natural name systems have been transformed into official codes based on hereditary patronyms. We show how, globally, these changes differentially altered the information structure of codes, leading to cross-cultural differences in the way names function as individuators that can have tangible effects in domains like scientific publishing.

3.1 Introduction

Personal names (which may comprise a word or set of words to form an identifier) provide a means for constructing identities and signaling information about individuals. For this latter function to be achieved, name systems must provide encodings that can allow individuals to be identified in communication, while also being organized to enable these encodings to be shared and used. An obvious strategy for achieving the first requirement would be to use a unique name word for each individual. However, although this approach might work for the 82

principal characters in "Game of Thrones" (who almost all have unique first names), it would fail the second requirement in any large-scale modern society, where it would require millions of pronounceable yet not already used words to be generated. Given the limited coding resources of natural languages, many names would end up being both very long and highly confusable, making the resulting system impossible to learn or use. Perhaps unsurprisingly, instead of employing unique name words, cultures satisfy the lexical demands of naming by creating combinatoric identifiers[42,154–156]. These combine small sets of lexical items that are often name specific with preexisting vocabulary elements to encode individual identities as phrases, as in the names of the badminton player Simon Archer and the basketball player Yao Ming 姚明 (Simon and Yao are names, whereas Archer is an English word for someone who shoots arrows with a bow, and Ming is Chinese word meaning brightness).

We present an information-theoretic analysis of personal names that reveals historical commonalities in the systems of combinatoric identifiers used in different cultures, following the approach in [154]. We use this analysis to examine the different ways in which the legalization of names – which occurs as societies become more organized and bureaucratic [43] – has changed these systems, and show how legalization can impede the communicative efficiency of name systems by imposing top-down structures on what had been organically evolved systems. Finally, taking international scientific publishing as an example, we show how the adoption of arbitrary standards that are blind to these differences can differentially impair the way the names of authors from different cultures function as individuators.

Since our focus will be on the coding function and information structure of name systems, we describe the parts of name-phrases that are spoken or written first as 'prefix-names', and the parts that follow them as 'bynames' (see Figure 3.1). Prefix-names can be specific lemmas (proprial lemmas [42]) or other vocabulary items marked by affixes [157]. The semantics of lemma prefix-names are typically opaque, such as John or Mary in English, or *Liú* 刘in Chinese [42], or else they are irrelevant in use, as in the English names Victor, Rose or Destiny. When affixing is used, for example, in Ngômbà, names are made by marking preexisting vocabulary items with prefixes, such as Mà for females and Tà for males [158]. Again, marking existing words as names changes their semantics, and often renders them irrelevant in use.

When it comes to bynames, across cultures name systems have largely either re-employed prefix-names for this purpose (e.g., Rhys ap Llywelyn – where ap means 'son of' – is a fairly typical patronym in the Welsh naming system [159]; see also Johnson in English, Jónsdóttir in Icelandic, etc.) or else re-deployed existing vocabulary items. For example, *Hóng* 红, Goch, Roth, and Červený, which have often been used as bynames in Mandarin, Welsh, German, and Czech, respectively, are simply forms or derivatives of the word for red. From a communicative perspective, this is notable because it enabled these systems to generate enormous numbers of individual identifiers while simultaneously avoiding the need to add huge numbers of extra words to linguistic inventories.

The relative benefits of redeploying old words versus adding new words in this way can be operationalized and measured using information theory [160], which provides a means for formalizing notions such as information and efficiency in communication, and has been fruitfully applied to a variety of questions in linguistics [18,21,23,161,162] by treating languages as information-theoretic codes. Formally, information (or entropy) is defined in terms of the uncertainty associated with an event or distribution of events. Efficient codes minimize the

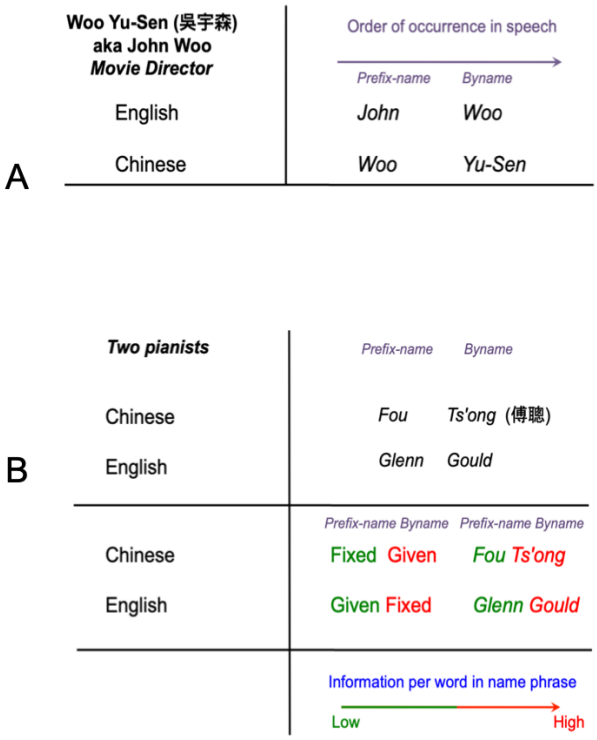


Figure 3.1: **Western names and East Asian names, despite differences in naming traditions, share a similar information structure.** (A) Two forms of the same name, movie director Woo Yu-Sen / John Woo. In Mandarin, the (inherited) prefix-name Woo comes first in speech. However, in its conventional Westernized form, Woo becomes the byname. This convention aligns the given and fixed (inherited) parts of EACS and Western names (because Western prefix-names are given and bynames inherited, whereas EACS bynames are given and prefix-names inherited), however it misaligns their information structures. In both Western and EACS name-phrases, prefix-names are systematically less informative than bynames, thus in (B): Chinese pianist Fou Ts'ong's inherited name is his prefix-name Fou, while Canadian pianist Glenn Gould's inherited name is his byname Gould; Fou's given name is his byname Ts'ong, whereas Gould's given name is his prefix-name Glenn.

complexity of codewords while maximizing their informativity about underlying messages [160]. Prefix-names seem to offer this kind of efficiency: If 33% of people are named John, and 1% Gerald, Gerald will be more informative than John (Gerald better discriminates individuals than John). Simultaneously, not only will John be easier to recognize and process [163], but the large number of Johns will reduce the total number of names, making the system as a whole easier to learn and use. Intuitively, a system where frequent prefix-names like John are distinguished by further optional codewords conditioned on them (i.e., John Adams, John Hancock...) may seem simpler and more efficient than a system in which everyone has a unique name. Information theory provides tools for assessing whether this is the case or not (and assessing the impact of changing the relative numbers of Johns and Geralds in systems).

Historically, across cultures, prefix-names served as the primary historical means for referring to individuals [43,164]. However, these names were not always fixed in the modern sense. Rather, the name by which an individual was known depended on the context in which it was used. This tendency is still evident in indigenous Australian prefix-names, which can change as people age or undergo major life events [165], or Yupik and Inuit prefix-names, which can change throughout a person's life (and have even been documented as changing after a serious illness [166]). It was also evident in the vernacular names of pre-modern Europe, which were similarly context-dependent [42]: the same "John" might have been "John (the) Smith" to distinguish him from "John (the) Baker" but "John Short" in other contexts [43]. In use, it even remains true for modern U.S. names: e.g., the 44th President of the U.S. is "Barack Hussein Obama" on his birth certificate, "Barry" (a childhood nickname) to intimates, "Barack" to others, "bo" on Twitter, etc.

All of which raises a question: why have the context-dependent vernacular name systems that characterize most of human history now been replaced by official, fixed name systems? The answer seems to lie in the way that the context-dependence of vernacular names conflicted with the needs of bureaucracies and governments which, as they developed, required that their citizenries be ever more "legible" [43] (i.e., that their activities be amenable to top-down observation and management). Fixed – context-free – official names offered bureaucracies an objective, tractable technology for managing citizens and their property [43], making tasks such as taxation, conscription, etc., far more manageable. Historically this conflict is most clearly visible in the laws that were used to fix the naming practices of particular populations, such as those that led to the enforcement of patronymic-bynames on Native Americans [167], Jewish people in Europe [43], and on the Inuit population in Canada (including attempts to assign each individual a unique number [168]). Meanwhile, the resolution of this conflict in favor of the needs of governments can be seen in the fact that today most states have legalized their citizens' names: across the globe, the vernacular naming practices that may be as old as languages themselves have now been replaced with official, fixed name codes. [168] summarizes it starkly: "Clashes [...] between different traditions and practices of naming, especially in the context of slavery and empire, illuminate with clarity the relevance of names as technologies of exclusion and belonging."

Critically, the legal imposition of fixed names (as hereditary patronyms) was implemented differently in the East Asian Cultural Sphere (EACS) as compared to the West. In particular, whereas laws in the EACS made prefix-names hereditary (and thus fixed), Western laws transformed bynames into hereditary, fixed patronyms (Figure 3.1). These differences provide a natural way of exploring whether name systems serve to communicate identities in the way we propose, because our information-theoretic analysis provides clear predictions about the way we should expect vernacular name systems to have been structured across cultures historically, and how they should subsequently have developed over time, given the changes that have been imposed.

Notably, from this perspective, it has been shown that the sets of prefix-names employed in China were historically quite small (in China, lǎobǎixìng 老百姓– "old hundred names" – is the colloquial term for the masses, such that historically 50% of the population shared just six of them [43,169]) with their overall distribution being exponential [170].

Not only does this provide an efficient way of organizing variable-length codewords [171], but it has also been shown that Korean prefix-name distributions had (and have) similar

properties [172,173]. It is thus notable that in every 50 year period between 1550–1800, records reveal that roughly 50% of the UK population shared 3 female prefix-names (Elizabeth, Anne and Mary) and 3 male prefix-names (John, William and Thomas) [174], and that the stocks of prefix-names in communities also typically approximated 100 [175]. Other historical prefix-name distributions across Europe were similar [174].

In China, people were assigned patronymic fixed hereditary prefix-names after population registration efforts in the Qin dynasty [176], and similar systems later developed in other EACS states. In contrast, when European nation states issued laws that fixed names from the 18th century onward, they turned bynames (already associated with property inheritance) into fixed hereditary patronyms (hereditary bynames) [43]. Importantly, because bynames rather than prefix-names were fixed in the West, it followed that if populations were to grow – as they did in response to industrialization – then the total number of unique identities encodable in European name-systems could only be increased by adding more prefix-names, the information structure of which appears to have been generally stable across Western cultures up to this point [173,174].

Here, we empirically test several hypotheses related to the communicative efficiency of names. In our first study, we collect and analyze the properties of historical name data from birth records, census records, and other historical sources. Our first key hypothesis is that, given that Western cultures largely retained vernacular name systems until the 19th century, we should expect pre-modern prefix-names distributions in these cultures to reflect this common system and show relatively low entropy on prefix-names. First, we establish that vernacular name systems across cultures show a similar low-information profile. Moreover, in countries where prefix-names were hereditary fixed patronyms – i.e., EACS countries where laws served to fix a small relatively invariant information system – we should expect distributions of prefix-names to largely resemble vernacular name systems, regardless of any subsequent population increases. Thus, empirically, we should expect modern distributions of EACS prefix-names to resemble those of pre-modern Western cultures. On the other hand, we expect that, once hereditary bynames were implemented in the West, as populations grew, Western prefix-name distributions would shift to communicate more information in proportion to this growth. In Study 1, we confirm these hypotheses using historical data.

In Study 2, to provide a more detailed test, we examine a dataset of church records from Finland (in the period roughly 1700 to 1900, a period during which the population grew from less than 500,000 people to over 2.5 million). These provide a particularly useful means for testing our proposal, first because hereditary fixed bynames were adopted during this time, and second because their adoption was asymmetrical across the country, occurring first in eastern Finland and then increasingly in the west of Finland.

Accordingly, given that this system enables us to observe the introduction and development of the use of hereditary fixed bynames in a name system, based on our first hypothesis, we predicted that we would observe a corresponding increase in the entropy of prefix-names across this period. Further, we predicted that if the increase in information in prefix-names is a result of the transformation of flexible bynames into hereditary fixed patronyms, prefix-name information entropy should increase as a function of the degree to which fixed hereditary bynames are used. To foreshadow our findings, we observe the predicted correlation with the entropy of Finnish prefix-names both over space (higher entropy first names in the east) and time (higher entropy first names later in time, as hereditary fixed bynames become more

prevalent).

Taken together, these results enable us to derive another main hypothesis – which we test in Study 3. We have argued that naming systems have evolved to yield name phrases that have low-information prefix-names and high-information bynames, and shown how different cultures have made different parts of these phrases into hereditary fixed patronyms. Accordingly, it follows that forcing names from different cultures into alignment according to their hereditary patronymics as opposed to their information structure (i.e. making names from EACS countries adopt the Western convention of putting the hereditary patronym at the end of a name phrase) could have the effect of seriously undermining the communicative function of some name-phrases as compared to others.

We show that this is the case in academic publishing in the Anglophone world by considering a sample of names from academic honor societies from 6 countries (China, Korea, Finland, France, Russia, and the U.S.). Under the Western convention, publications are often cited in papers by the hereditary patronym of the first author (where this is a high-information byname, such as Einstein et al.), whereas in China and Korea, publications are often cited by the full name of the first author (e.g., Li Zhaoping et al.). As predicted, we show that forcing names from China and Korea into the Anglophone model, in which only the hereditary patronym is given in full (which in both cases is a low-information prefix-name) and in which bynames are initialized (e.g. X. Li) induces unnecessary name ambiguity that can be resolved if other conventions are employed.

3.2 Results

3.2.1 Study 1: Name distributions across culture and time

Our first hypothesis predicted that the distribution of prefix-names in pre-modern Western populations (i.e., before bynames were turned into fixed patronyms) would have been similar, communicating relatively low levels of information (operationalized as Shannon entropy [160]); that they will resemble modern EACS prefix-names (which are themselves hereditary patronyms); and that modern Western prefix-names will have communicated increasingly higher levels of information as populations grew.

To test this, we examined the prefix-name distributions of birth-records in four pre-modern Scottish parishes: Earlston, Govan, Dingwall, Beith, between 1700-1800 [175], analyzed individually and collectively. We also examined the birth records from Finland from the same period, along with marriage registers in the counties of Northumberland and Durham in Northern England (again covering the same period [177]). Our EACS prefix-name distributions were taken from 2015 South Korean census data [178], 2018 Taiwanese census data [179], and Vietnamese-American population data extracted from the 2010 U.S. census [180]. It is worth noting that although prefix-names are patronyms in Taiwan, Korea, and Vietnam, the languages spoken in the three countries belong to different linguistic families: many of the major languages spoken in China and Taiwan are part of the Sino-Tibetan language family, [181], modern Korean is a language isolate [182], and Vietnamese is an Austroasiatic language, albeit with significant vocabulary borrowings from China [183].

For our modern Western analyses, we used post-1900 data taken from a larger genealogical

set of Finnish name records [184], covering the period after hereditary patronyms had become common in Finland [185], and data from the United States Social Security records between 1910 and 2010 [186], focusing on a representative large state, California, and a small state, Delaware. Table 3.1 lists the number of names and population in each dataset (see [154] for a smaller-scale analysis comparing just the US and Korea). No data were excluded from the analyses.

Locale	Name count	Cut-offs	Sample	LB	UB
California	17638	5	27517342	10.16	14.40
Delaware	1534	5	597042	8.57	12.72
All US, 1910–2010	28638	5	286605557	9.80	11.22
Finland Post-1900	1066	1	75814	7.20	7.20
Taiwan	500	135	23543563	5.68	5.71
Finland Pre-1800	1833	1	1748523	5.43	5.43
Korea	533	5	49705663	4.82	5.41
Northern England	225	1	25797	4.93	4.93
Dingwall	81	1	1685	4.87	4.87
Earlston	79	1	3121	4.85	4.85
Govan	121	1	12086	4.70	4.70
Scottish	361	1	23792	4.65	4.65
Vietnamese-American	62	100	1527654	4.35	4.59
Beith	80	1	6900	4.35	4.35

Table 3.1: Summary of several unique names and the entropy estimates of the prefix-name distribution in each dataset. The inclusion threshold, namely the least frequent prefix-name included, along with the total number of people sampled, is included. The lower bound entropy (LB in the table) reflects the observed sample (in bits) while the upper bound (UB in the table) reflects an estimate assuming that all unobserved names are unique (i.e., it assumes the maximum entropy for the unobserved portion of the distribution).

The information communicated by each prefix-name distribution was calculated as follows: if there are N prefix-names in a given distribution, each with a frequency of p_i, the entropy of the distribution is:

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i. \qquad [3.1]$$

Note that although there is a large amount of data on personal names and their distributions in existence, studying it is complicated by several related factors: name data tend to be the preserve of state governments, and responsible governments require that publicly released data relating to individuals be anonymized [187]. This last point necessarily conflicts with the *sui generis* function of names – and especially fixed, legally defined names – which make the identification of many individuals from full name data trivial. Accordingly, most of our data do not contain the prefix-name of every individual and contains a cut-off number, reflecting these concerns (see also Study 1 in the Methods Section). If the number of people sharing a prefix-name is lower than the cut-off, the prefix-name is not part of the data. Therefore, the entropy calculated from our data is only a lower bound of the true entropy of the population.

To estimate an upper bound, we assume an extreme (if unlikely) scenario, namely that everyone not included in the database has a different prefix-name. For example, the 2015 Korean census data do not contain the prefix-names of approximately 1.1 million people. To estimate an upper bound of Korean prefix-name entropy, we assume these 1.1 million people have 1.1 million different prefix-names and calculate the entropy of the population under this assumption.

Figure 3.2 plots the 100 most common prefix-names (dividing the frequency of each name by the frequency of the most common name) from each distribution. As predicted (Table 3.1), the different prefix-name distributions in our pre-modern Western samples communicated relatively small – and similar – amounts of information: Earlston, Govan, Dingwall, Beith have entropies of 4.85, 4.70, 4.87, and 4.35 bits, respectively (4.65 bits when aggregated together); the Northern English sample had an entropy of 4.93 bits; and the Finnish prefix-name distribution taken from the same period had an entropy of 5.43 bits.

Also as predicted, the entropies of our Vietnamese American and Korean samples (where 92% of the population share the 50 most common prefix-names) were in the same range (4.35 and 4.82 bits, respectively), as was the entropy of our sample of Taiwanese prefix-names (5.68 bits).

Critically, the US name distributions have by far the highest entropies (10.16 bits in California and 8.57 bits in Delaware), and the prefix-name entropy in our Finnish sample rises to 7.20 bits after 1800, approaching the range observed in our US data (this change is analyzed in detail below). It should be emphasized that higher entropies in these results do not simply reflect larger population sizes: the Korean sample is far bigger than the Delaware sample, but even the upper bound of the Korean name entropy is much lower than the lower bound of the name entropy for Delaware. In addition, results from a study on Korean family records [173] indicated that, despite a large growth in population between 1510 and 1990, and the inclusion of many prefix-names in this period, prefix-name entropy remains largely unchanged. Similarly, the entropy of our aggregated Scottish sample is lower than the individual entropies of 3 of the 4 parishes it comprises.

3.2.2 Study 2: Relationship between fixing patronyms and prefix-name distributions

Since industrialization, the world's population has grown considerably [189]. The population of 19th-century Britain grew rapidly [190] and as it did, the distribution of prefix-names changed: in 1800, $>$ 50% of children were given the most frequent 6 prefix-names; by 1900, it was $<$ 25% (Figure 3.2). This pattern is consistent with prefix-name information increasing to offset the development of hereditary fixed bynames. However, to make a more detailed assessment of these changes we examined a set of Finnish birth records covering the period 1600-1917 (from the same dataset examined in Study 1 [184]). Because of the late implementation of name laws in Finland (1929), and the fact that hereditary fixed bynames became dominant in eastern Finland earlier than western Finland [185], these data offer a unique opportunity to explore the association between information profiles in flexible byname systems as opposed to in hereditary patronymic-byname systems, for a system in transition. We divided the records, which comprise a baby's prefix-name, parish of birth and, where

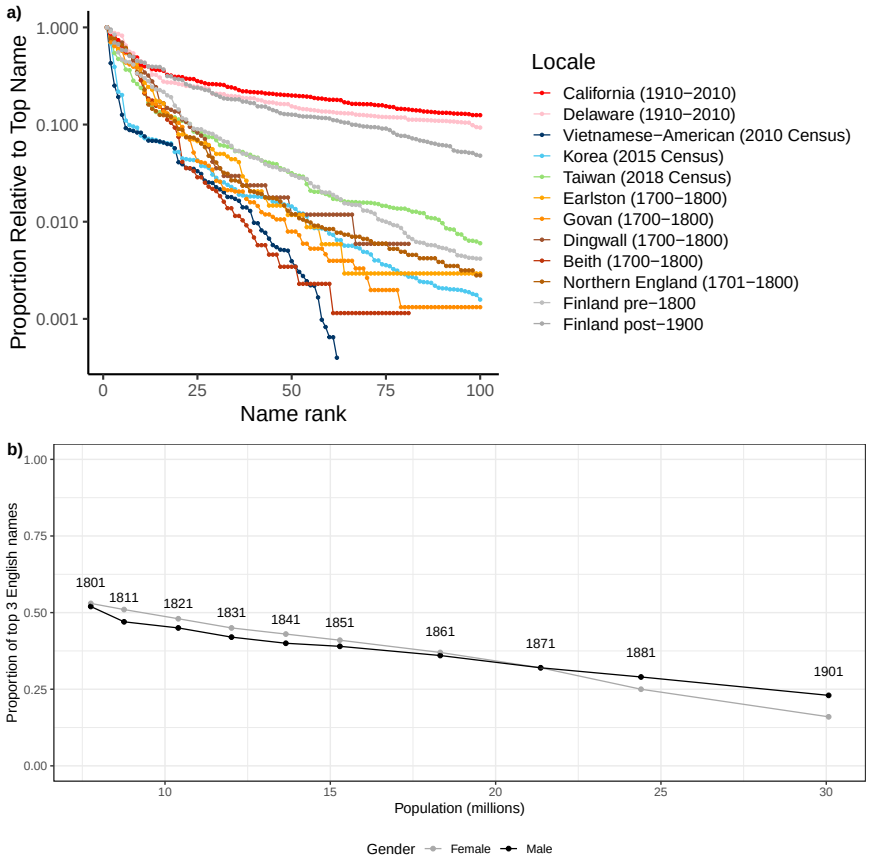


Figure 3.2: **Frequencies of most frequent names across places and time. a): The 100 most frequent prefix-names by rank in the prefix-name distributions analyzed in testing hypothesis 1** (plotted as a proportion of the most frequent name): four parishes in Scotland, 1701-1800 (Beith, Dingwall, Earlston, and Govan), two counties in Northern England (Durham and Northumberland) between 1701 and 1800, Finland pre 1800 and post 1900, South Korea (2015 census data), Vietnamese-American (reconstructed from US census 2010), Taiwanese (2018 census data), and US names between 1910 and 2010 from Delaware and California. The $y-$axis is logarithmic. Critically, before the introduction of naming laws, the prefix-name distributions of the Scottish parishes, Northern England, and pre-1800 Finland are highly skewed, and similar to those of Taiwanese, Korean, and Vietnamese-American prefix-names. By contrast, after laws required people in the West to inherit their bynames, these distributions began to change as the information conveyed by the prefix-name increased, as revealed by the longer, flatter tails of the modern prefix-name distributions in California, Delaware, and post-1900 Finland. **b) The frequency at which the top three male and female prefix-names were given in England by decade 1800–1880 and 1900**. As can be seen, these frequencies declined as the population grew, which is again consistent with our prediction that fixed bynames and population growth will be associated with flatter name distributions, and concomitant increases in the information conveyed by prefix-names (data from Table 1 in [188], plotted against the total population at each time point.)

available, a paternal byname (e.g., relating to the farm or village of the father) into 5 different year bins: pre-1730, 1730-1780, 1780-1830, 1830-1880, and post-1880. Due to the uneven number of records in each birth year bin (Table 3.6), from each birth year bin and parish we randomly sampled 50 births, from 473 parishes.

	Birth Year	East	West
1	up to 1730	72%	29%
2	1730–1780	67%	37%
3	1780–1830	71%	45%
4	1830–1880	81%	57%
5	post–1880	94%	79%

Table 3.2: Percentage of births registered with a paternal byname. Patronymic-byname use in the East starts high and increases from 72% to 94% in the period 1700–1900. By contrast, this increase is much sharper in the West, where it is 29% at the beginning of this period and increases to 79% by its end.

Taking the rates of occurrence of paternal bynames in records as a proxy for the use of hereditary fixed bynames, we first examined whether the transition from flexible to fixed bynames actually occurred in this period. Before 1730, 42% of birth records included a paternal byname. After 1880, this proportion had increased to 89%. We also observe a clear east/west gradient in these data, which changes over time, as shown in Table 3.2. To assess significance, we computed the proportion of the births for which a paternal byname was present, in each parish, in each time period. Then we fit a mixed-effect linear regression model predicting this proportion based on the longitude, binned birth year (coded 1 to 5), and their interaction. We included a random intercept for parish, with a random slope for birth year, following a maximum appropriate random effect structure from the nature of the dataset, as recommended in [191]. We found significant main effects of longitude (farther east means more paternal bynames; two-tailed $t(381.8) = 8.758$, $p < 0.001$, $\beta = 0.073$, 95% confidence interval=[0.057, 0.090]) and birth year (more paternal bynames overall as time progresses; two-tailed $t(350.4) = 5.505$, $p < 0.001$, $\beta = .259$, 95% confidence interval=[0.167, 0.351]), and an interaction between them (which we take to mean that the effect of longitude gets smaller as birth year increases; two-tailed $t(352.4) = -3.913$, $p < 0.001$, $\beta = -0.001$, 95% confidence interval=[-0.011, -0.004]). We also fitted the data with a null model without the interaction term between longitude and birth year, and we found the model with an interaction term explained the data significantly better than the model without one (one-tailed $\chi^2(1) = 15.059, p < 0.001$).

We then examined how the adoption of hereditary fixed bynames affected the informativity of prefix-names. We expected that prefix-name entropy would increase along an east/west gradient, initially increasing more in the east than the west. Figure 3.3 projects the entropy of prefix-name name distributions on a map of Finland. Each panel is a different time period and, moving from left to right, reveals an overall increase in prefix-name entropy modulated by an east/west cline that becomes less prominent over time.

Our results are not dependent on the particular sample presented here. In Appendix B.1, we repeat these random sampling procedures and analyses 500 times, and show that the

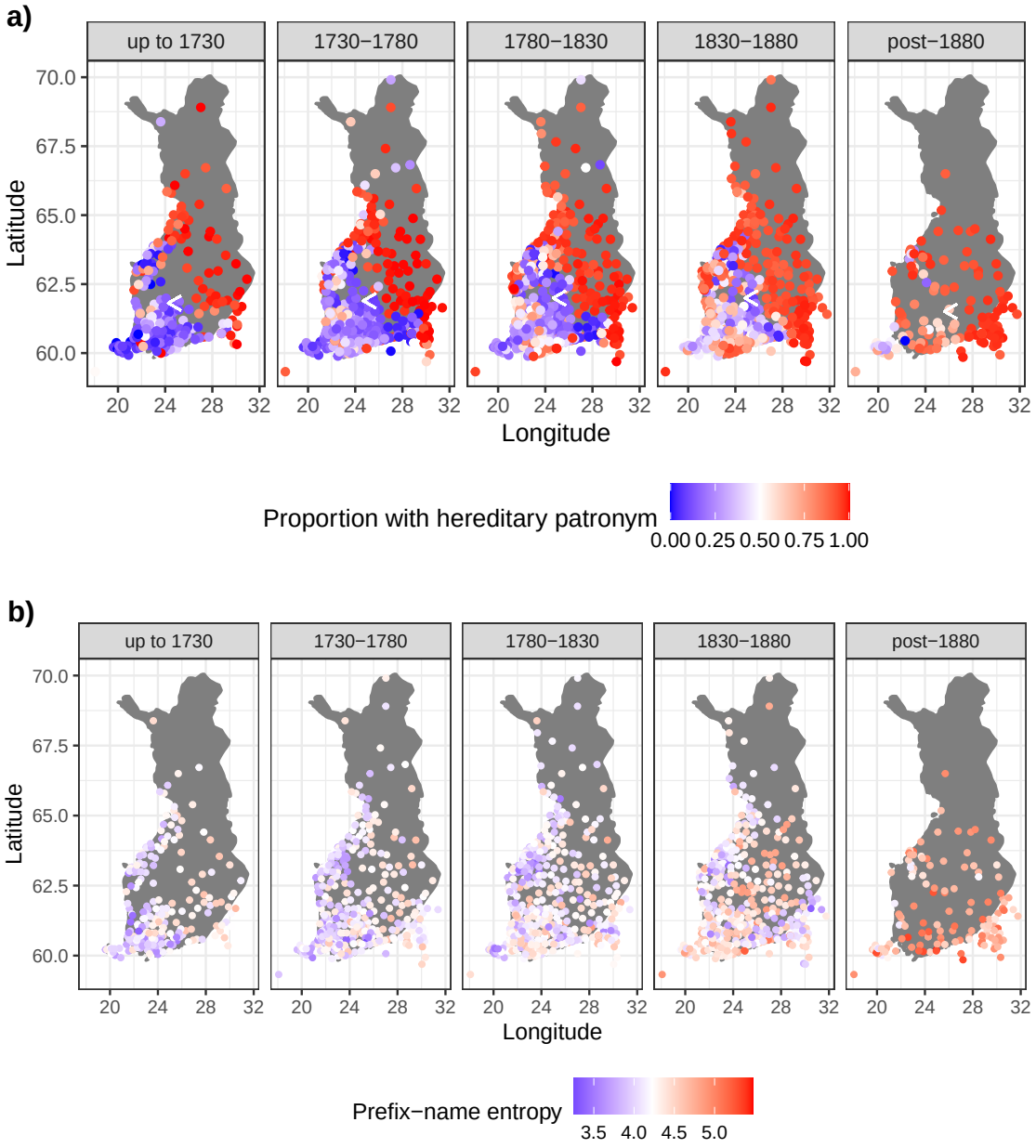


Figure 3.3: **In Finland, prefix-name entropy increased as the government required residents to have a hereditary patronym**. **a)**: Proportion of birth records in each Finnish parish at each time period containing a hereditary byname. **b)**: prefix-name entropy (for 50 randomly sampled names per parish) in each parish in each period. The plots indicate that Finnish prefix-name entropy increased as a proportion of the increased use of hereditary bynames as the government imposed name regulations, both of which are reflected in an east/west gradient of change: the proportion of hereditary bynames (a) and increases in prefix-name entropy (b) are initially evident more in the east than the west, before slowly spreading westwards across the country.

results are robust.

Having established the east/west patronymic-byname use effect, which starts large and decreases over time, we then asked whether prefix-name entropy shows a similar pattern. We ran a mixed-effect regression predicting prefix-name entropy, calculated by sampling 50 births per bin, as described above. The predictors included binned time period, longitude, and their interaction. We also included random intercepts for parish, and a random by-binned time period slope for parish. There were main effects of birth year (more prefix-name entropy over time; two-tailed $t(301.5) = 4.591$, $p < 0.001$, $\beta = 0.256$, 95% confidence interval=[0.146, 0.365]), longitude (more prefix-name entropy in the east than the west; two-tailed $t(280.2) = 6.475$, $p < 0.001$, $\beta = 0.044$, 95% confidence interval=[0.031, 0.058]), and crucially a significant negative interaction (two-tailed $t(301.5) = -2.860$, $p = 0.005$, $\beta = -.006$, 95% confidence interval=[-0.011, -0.002]; by a likelihood ratio test comparing the full model to an identical model without the fixed effect interaction term, one-tailed $\chi^2(1) = 8.053, p = 0.004$).

Birth Year	**Variable Pair**	**Spearman ρ**	**95% CIs**	**p-value**
up to 1730	Patronym:PrefixNameEnt	0.29	(0.16, 0.41)	< 0.001
	Patronym.:Longitude	0.54	(0.43, 0.65)	< 0.001
	PrefixNameEnt:Longitude	0.49	(0.38, 0.60)	< 0.001
1730–1780	Patronym:PrefixNameEnt	0.22	(0.11, 0.32)	< 0.001
	Patronym.:Longitude	0.35	(0.24, 0.46)	< 0.001
	PrefixNameEnt:Longitude	0.40	(0.31, 0.49)	< 0.001
1780–1830	Patronym:PrefixNameEnt	0.18	(0.08, 0.27)	< 0.001
	Patronym.:Longitude	0.38	(0.28, 0.47)	< 0.001
	PrefixNameEnt:Longitude	0.21	(0.11, 0.30)	< 0.001
1830–1880	Patronym:PrefixNameEnt	0.10	(0.003, 0.20)	0.039
	Patronym.:Longitude	0.42	(0.34, 0.51)	< 0.001
	PrefixNameEnt:Longitude	0.01	(-0.08, 0.11)	0.77
post-1880	Patronym:PrefixNameEnt	0.06	(-0.13, 0.24)	0.52
	Patronym.:Longitude	0.72	(0.63, 0.81)	< 0.001
	PrefixNameEnt:Longitude	0.16	(-0.01, 0.34)	0.058

Table 3.3: The Spearman rank correlation between proportion of patronymic-bynames present and the prefix-name entropy, between proportion of patronymic-bynames present and longitude, and between prefix-name entropy and longitude, along with their 95% confidence intervals (CIs) and p-values. The CIs are calculated using the SpearmanCI package [192] in R [145].

We also examined the Spearman rank correlation between the proportion of names in a parish with a paternal byname and the prefix-name entropy in that parish. In each time period, the correlation was positive, and it steadily decreased over time as shown in Table 3.3. These results suggest that the fatter-tailed distribution of modern prefix-names results from the changes in the information structures of names resulting from the introduction of hereditary fixed bynames.

3.2.3 Study 3: Mixing of naming systems with different conventions

The information in U.S. prefix-names grew exponentially in the 20th century [193], whereas EACS prefix-name distributions hardly changed [173]. The effect of this is easily quantified: the prefix-names of the 49.7 million South Korean citizens in the 2015 Census have an entropy of 4.82 bits; for the prefix-names of the 27 million Californians registering for Social Security between 1910-2010, 10.16 bits (Table 3.1). Since entropy is logarithmic, in theory this means that the information processing cost of Californian prefix-names is now about 40 times more ($\approx$ 5.3 bits more information) greater than Korean prefix-names, whereas our earlier analyses indicate that 200 years ago, Korean and English prefix-names contained roughly the same amount of information. This allows us to make detailed predictions about the cross-cultural impact of using Western naming conventions as a gold standard for indexing in science and publishing.

Before describing these predictions, it is important to stress how different ways EACS and Western names were legalized can make comparing them confusing (and lead to confused remarks like, "[in China] they put their last names first" [194]). This confusion arises because in the EACS, prefix-names are fixed and inherited, and bynames are 'given', whereas in the West, prefix-names are 'given', and bynames are fixed and inherited (see Figure 3.1). However, for current purposes, what is important is that both Western (fixed patronymic) bynames and EACS (given) bynames are far more diverse and more informative than US (given) prefix-names and EACS (fixed patronymic) prefix-names (See e.g. Table 3.4).

	Country	Prefix-name Entropy	Byname Entropy
1	China	6.15	8.64
2	Finland	6.94	8.54
3	France	7.03	8.68
4	Korea	4.72	8.71
5	Russia	5.27	8.58
6	US	7.43	8.65

Table 3.4: The prefix-name entropy and the byname entropy of matched samples taken from 2550 scientists from 6 different countries.

Because prefix-names convey less individuating information than bynames (and the individuating information in initials is similarly constrained), it follows: 1) Under current indexing conventions, which combine initialized 'given names' with full inherited names, EACS names will convey far less individuating information than Western names, making EACS names more confusable; 2) If indexing conventions were recast to reflect the information structure of name phrases, the confusability difference between Western names and EACS names could be drastically reduced. Note that in this study, we are not analyzing names under the categorization of prefix-names and bynames as in previous ones. Instead, we will follow the (Western) convention of name categorization and the indexing convention arising from it to show how such conventions lead to greater confusion if it is applied to EACS names. To avoid confusion, in this study, we will refer to the part of the name that is given as "**flexible given name**" and the part of the name that is inherited as "**fixed inherited name**."

To test the hypotheses, we obtained cross-cultural samples of scientific communities by scraping membership lists of academic/scientific societies in the United States, China, Korea, France, Finland, and Russia [195–200] (see Study 3 in methods). We explicitly chose these samples in order to examine the possible real-world consequences of current indexing conventions (and hence the transliteration of character-based Chinese and Korean names to the Latin alphabet) in a domain where the communication of identities is critical, namely academic publishing.

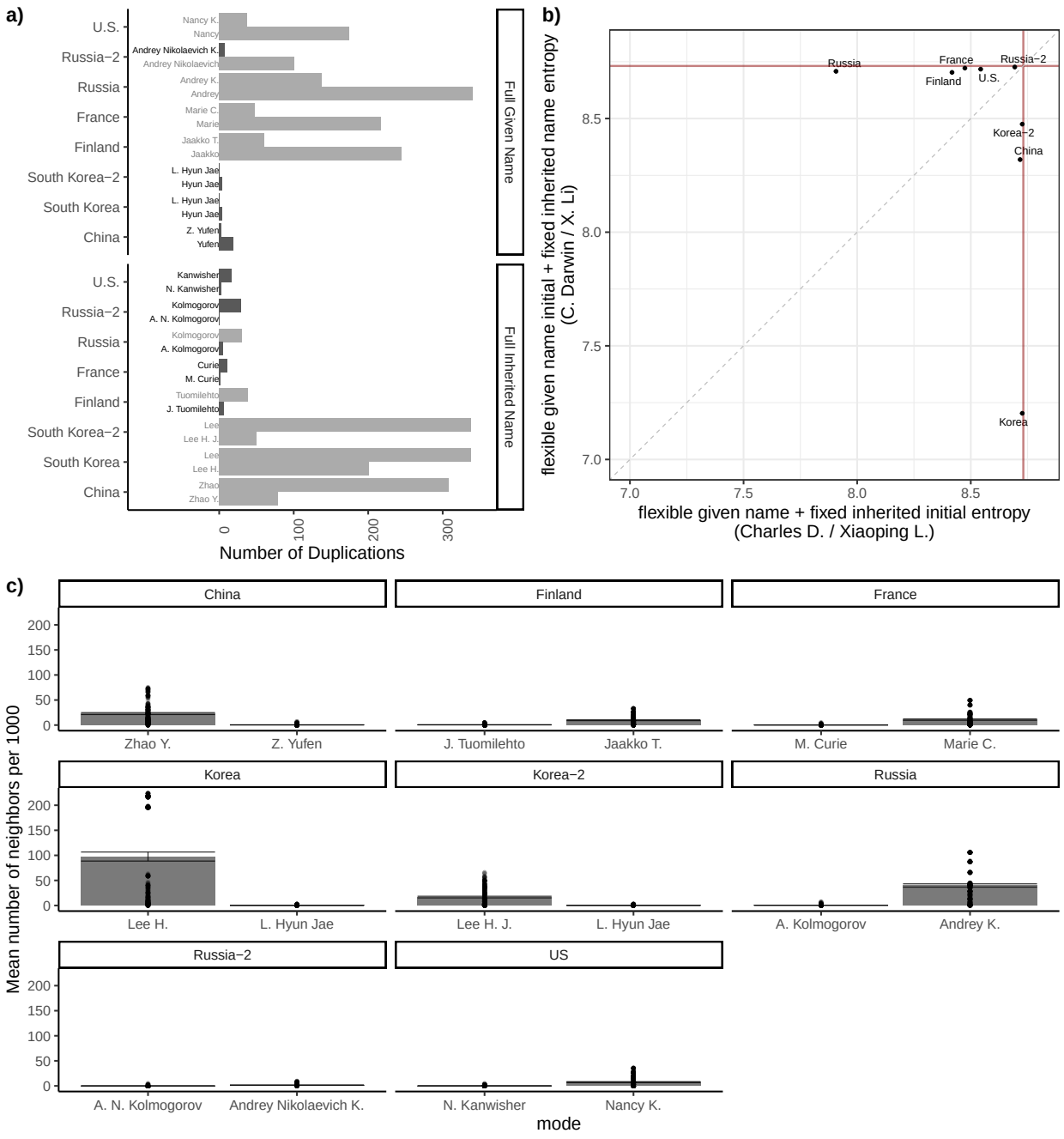
The lists of names were then spot-checked, and typos or extraction errors were fixed as needed. The exact list of names used is in the repository associated with this paper. To compare across datasets, we randomly downsampled all datasets to make them the same size as the Korean set (425 individuals), and then computed the number of repeated names: for example, if "Robert" occurs 4 times, its repeat count is 3. This was done for each: flexible given name, fixed inherited name, full flexible given name + fixed inherited name initial, and flexible given name initial + full fixed inherited name (see the top left panel in Figure 3.4). We then generated sets of possible shortened versions of the names by taking only the flexible given name, only the fixed inherited name, or the abbreviated initials of either the flexible given name or the fixed inherited name.

We then measured various properties of these distributions, which generated the following name styles: flexible given name initialized (e.g., C. Darwin / Z. Li), fixed inherited name initialized (e.g., Charles D. / Zhaoping L.), flexible given name alone (e.g., Charles / Zhaoping), fixed inherited name alone (e.g., Darwin / Li).

We also considered the potential influence of two further naming conventions present in these samples. The first is the presence of Korean generational names, which (in South Korea at least) typically comprise the first syllable of a flexible given name, which is shared across siblings, such that the second syllable is unique to the individual, e.g., Korean soccer-playing brothers Son Heung-min and Son Heung-yun. Accordingly, in Korean, names are typically indexed by both syllables, such that Son Heung-min's name is rendered as "H M Son". However, in our basic analysis the given name of Hyun Jae in Figure 3.4 is simply rendered as H. To better reflect the specific structure of Korean names, we also explored an alternative rendering that treated each syllable in a Korean given name independently, such that the given name initial of Hyun Jae was represented as H. J. instead of H. We denote this alternative convention as "Korea-2" in Figure 3.4.

The second alternative consideration relates to patronyms in Russian names. Unlike most Western societies, a full Russian name also includes a third component in addition to flexible given name(s) and a fixed inherited name: a true patronym, derived from a parental given name (a practice also common in other East Slavic and ex-Soviet societies). For instance, the late Russian linguist Vadim Kasevich's full name is Vadim Borisovich Kasevich, reflecting the fact that his father was Boris, and the late linguist Alexandra Kamynina's full name is Alexandra Alekseevna Kamynina, since her father's name was Aleksey. Russian publications often cite individuals by their flexible given name initial, their patronymic initial, and their full inherited name, rendering these two linguists as V. B. Kasevich and A. A. Kamynina, respectively. This alternative convention is denoted as "Russia-2" in Figure 3.4.

Our analysis also makes clear that current academic publishing conventions (flexible given name initial + full fixed inherited name) lead to more ambiguity for Chinese and Korean names than for U.S. names, even though almost every full name phrase was unique in every



dataset. Figure 3.4 plots the entropy of the flexible given name + the fixed inherited name initial (i.e., where Charles Darwin, Charles Dickens, and Charles Dodgson are all indexed as Charles D.) against the flexible given name initial plus the fixed inherited name (C. Darwin, etc.), and shows that when EACS names are *not* initialized according to Western conventions they are far less ambiguous. In addition, indexing every syllable in a Korean flexible given name (e.g., H. M. Son instead of H. Son) increases the entropy of the index distribution dramatically, and initialized Russian inherited names become less confusing if the initials for both the full flexible given name and the full patronyms are utilized.

In our analysis so far, we have treated each name atomically, but it cannot capture

Figure 3.4: **Imposing the indexing convention from one culture to another makes name indices less informative and more confusable. a)**: Number of repeated names, for each of flexible given name, fixed inherited name, and the two initialism conventions we consider (i.e., Russia-2 and South Korea-2). The darker bars show naming conventions which are mostly unambiguous on the scale of scientific communities as measured in our sample. Russia-2 and Korea-2 refer to forms using the Russian patronymic and Korean generational name as part of the initials (e.g. A.N. Kolmogorov and Lee H. J.). **b)**: Entropy of names of the form C. Darwin vs. Charles D., across cultures. The dark red line represents the maximal attainable entropy from 425 names. **c)**: The likelihood of confusion, as measured by the mean number of neighbors per 1000 for each name style (left: flexible given initial + fixed inherited name; right: flexible given name + fixed inherited name initial) among scientist names in each country. Error bars represent 95% bootstrapped confidence interval. Two-tailed t-tests suggested that EACS names are more confusable than Western names when they are indexed by flexible given initial + fixed inherited name ($t = 24.014$, $p < 0.001$), and Western names are more confusable than EACS names when they are indexed by flexible given name + fixed inherited name initial ($t = -28.531$, $p < 0.001$).

a critical factor in the empirical functioning of name systems: psychological confusability. Words are harder to process in the presence of competitors, and this effect increases as discriminability decreases [201]. To estimate the effect of this factor, we computed the number of orthographic neighbors in our datasets, defining neighborhoods in terms of the number of 0-edit and 1-edit neighbors of each name under the possible initial + name sequences, namely flexible given name initial + full fixed inherited name, and full flexible given name + fixed inherited name initial. The results are illustrated in Figure 3.4. EACS names have many more 1-edit orthographic neighbors when indexed in the Western convention (flexible given name initial + full fixed inherited name; two-tailed $t(1274.3) = 24.014$, $p < 0.001$, $\beta = 19.309$, 95% confidence interval=[17.732, 20.888]): Chinese names have an average 22.79 neighbors per 1000 names, and Korean names have an average of 97.5 neighbors per 1000 names, compared with Western countries where the numbers are lower than 0.5 per 1000 names. In addition, similar to the results in the entropy analysis, the number of 1-edit neighbors is sharply reduced for Korean names under the Western indexing convention when every syllable in a Korean flexible given name is indexed (16.7 per 1000 names, compared to 97.5 if only the first syllable is indexed). On the other hand, EACS names have many fewer 1-edit neighbors than Western names when the names are indexed by full flexible given name + fixed inherited name initial (two-tailed $t(2134.1) = -28.531$, $p < 0.001$, $\beta = -5.813$, 95% confidence interval=[-6.212, -5.413]). In our sample, "H Kim" is the index most likely to be confused, with 95 1-edit neighbors such as W Kim, B Kim, and H Lim. "Vladimir M" is the most confusable index in non-EACS names, having 45 1-edit neighbors such as Vladimir T and Vladimir D. But when including the full patronym, Russian name indices become much less likely to be confused.

3.3 Discussion

In the sample examined above, Western naming conventions resulted in far more confusable and ambiguous renderings of EACS names. To put these results in a wider perspective, in 2010, neuroscientist Li Zhaoping described how the conventional way of publishing her name in U.S. academia (Z. Li) made it ambiguous [202]. Given the importance of name recognition and name memorability in academia [203–206], Zhaoping went on to describe how the confusability that results from the way their names are rendered in western publishing continuously puts EACS scientists at a disadvantage, and announced that while by convention her name ought to written as 'Z. Li' for scientific publishing, she had decided to use L. Zhaoping instead.

Our analysis provides quantitative support for this position by showing that when the names of all of our EACS scientists are encoded Zhaoping's way, duplication rates in the EACS and Western datasets become similarly low. These results can enable other scientists to make decisions about their authorial identities in the context of a broader theoretical and historical framework that provides a more complete account of the functionality and effectiveness of names.

We suggest that increased popular awareness of these issues can play an important role in increasing cross-cultural understanding. The common misconceptions currently associated with name codes can make the world's name systems appear arbitrary. Researchers examining Chinese given names emphasize how "individuality is expressed by the [Chinese] given name" and note how "the opportunities for creativity and originality in the [Chinese] given name are much greater than in English name-giving tradition" [194]. Meanwhile, Western-focused narratives make exactly the 'opposite' point when it comes to the expressiveness of inherited names: "how tedious it would be if British surnames had the same neatness – and terseness: Park is a longer than average name in Korea. Koreans don't have names which ramble away into the distance like Featherstonehaugh, Haythornthwaite, or McGillicuddy" [207]. We have shown how, when seen from an information-theoretic perspective, the world's name systems share a common structure. In the languages examined here, a relatively small set of largely meaningless prefix-names was used to mark other lexical items as bynames to yield systems that, at least historically, enabled large sets of identifiers to be created without inflating the lexical inventories of languages. These vernacular name systems allowed identities to be communicated in ways that were smooth and efficient.

We have also shown how modern differences in these systems arose out of the different ways in which vernacular names were repurposed as fixed identifiers. And we showed how the lack of attention to the false equivalences that these changes have created in name systems can lead to unforeseen negative outcomes—along with some simple ways in which, once the common information structure of names is understood, these negative outcomes could be reduced or even eliminated. In particular, future research could explore how naming conventions could be improved for scientific publishing—e.g., using scholars' full names and/or offering scholars flexibility in how they choose to be identified—so that the various requirements of publishers, libraries, databases etc. when it comes to recording and indexing the output of scientific authors can be reconciled with the need of each author to have their specific outputs associated with their own authorial identity.

If we extrapolate from historical data anywhere in the premodern world, 12-15 of the main

characters in *Game of Thrones* ought to have shared the same most common prefix-name. Instead, the prefix-name of almost almost every character was unique! What our results indicate is that, while this kind of unique name system might seem plausible in the realm of fantasy, historical name systems were very different and featured large amounts of duplication. Information theory not only provides an explanation for that duplication, it can help explain why name systems across languages shared a common structure in the past. It also explains how important it is to understand this structure when it comes to manipulating these systems, whether this be for the purpose of creating official identifiers, or establishing standards in publishing. While accepting that the requirements of indexers must also be considered, these results indicate that a reversion to something closer to the historical norm – allowing individuals more flexibility in choosing how their names are presented – not only has strong historical and information-theoretic precedent, it could both help individuals reduce nominal ambiguity and do so more efficiently in the future.

3.4 Methods

This study was reviewed and declared exempt by the Institutional Review Board at the University of Texas at Austin as STUDY00004122.

3.4.1 Study 1

To estimate and compare prefix-name entropy across cultures and historical time, we drew name data from various sources. In what follows, we describe these datasets in further detail, and characterize how we compared them. Our usage of these datasets are compatible with their respective terms and conditions (See B.2 for more details).

Materials

US Name Data were taken from birth name records made available by the Social Security Administration [186]. The dataset includes prefix-names of babies born between 1910 and 2018 in all 50 states in the US. Since 1910 was the first year when the dataset was available, and 2010 was the last year for which census name data were available, the data between 1910 and 2010 were analyzed. Only prefix-names that are shared by more than 5 people in each state in each year are included in the data, with entries being restricted to cases where an individual's birth year, sex, and state of birth are on record, and where their given name is at least 2 characters long. The names in the records are then pre-processed to remove hyphens and spaces, so that Julie-Anne, Julie Anne, and Julieanne all count towards the same name's frequency.

For our detailed analyses, we chose a representative small state (Delaware) and a representative large state (California); however, further analyses showed that the name data for other states pattern similarly.

The majority of the **Vietnamese-Americans** sampled in the 2010 census [180] are naturalized citizens who were born—and named—in Vietnam [208]. This made it possible to

recreate a distribution of Vietnamese prefix-names (family) from US family name data to examine whether previous findings regarding Korean and Chinese prefix-names (which are exponentially distributed [170,172]) generalized to other EACS languages (see Figure 3.4a). To make this estimate, two lists of frequent Vietnamese prefix-names [209,210] were combined with a Vietnamese names list from Wikipedia. This yielded a total of 62 prefix-names (family names), the frequencies of which were then extracted from a dataset containing counts of the family names that occurred most frequently in the US 2010 census published by the US Census Bureau (source: [180]). The US census data contains family names shared by at least 100 people, along with the number of people bearing each last name, and the percentage of each name's distribution across 6 Race and/or Hispanic Origin subgroups.

US census entries use only the 26 letters of the English alphabet, which means that when many Vietnamese names are entered into it they become confusable with Western family names (e.g. the Vietnamese family name "Bạch" is entered as "Bach", a typically German family name). To improve the accuracy of our estimate we multiplied the number of people having each last name by the percentage of people with that last name in the Asian/Pacific Islanders (API) subgroup.

Name data for **South Korea** were obtained from the Korea National Statistic Office website, which published a dataset containing a list of the prefix-names shared by more than 5 people in 2015 census, along with the number of people with each name [178]. The prefix-names are listed in both Hangul (a phonetic writing system) and Hanja (a logographic writing system based on Chinese characters). Since multiple prefix-names written in Hanja sometimes share the same Hangul (e.g. the prefix-names in Hanja "到","度", "桃", "", "道", "都", and "陶" are all written as "도(Do)" in Hangul), we use Hanja characters as our unit of analysis. There are in total 533 Hanja prefix-names in the census, covering 49.7 million people in total.

Taiwanese Name Data was taken from a list published as part of a report by the Taiwanese Ministry of Interior in 2018 (source: [179], Table 57, pp.282-304). It contained 1832 prefix-names listed using traditional Chinese characters, along with the number of people sharing each name. We manually logged the 500 most popular prefix-names, covering a population of 23.5 million. The 500th most popular prefix-name is shared by 135 people.

Finnish Name Data were extracted for the period post 1900 from a larger set of genealogical Finnish name records collected and organized by [184].

The Scottish data, was extracted from the National Records of Scotland and published as an appendix in [175], and comprises the prefix-names recorded in four parishes (Earlston, Govan, Dingwall, Beith,) from 1700 to 1800. The English name data were extracted from George Bell's parish marriage register transcriptions for Northumberland and Durham between 1701 and 1800 [177].

The English records had been pre-processed to take account of alternate spellings of the same name (e.g., Ann / Anne / Hanna) and common abbreviations (e.g., Thomas being abbreviated to Thos.). Accordingly, we applied the same process to the Scottish data, as well as additionally removing obvious bynames that had been mistranscribed as prefix-names (e.g., where McVey is given as a prefix name). To estimate the potential impact of this kind of pre-processing, test comparisons between the processed and unprocessed distributions were then conducted. The results of these comparisons indicate that the effects of pre-processing on comparisons between name distributions are marginal (all pointwise $R^2 > .98$), suggesting

that the pre-processing that many of the other datasets we analyzed were subject to would have been unlikely to significantly affect the results of our analyses. We also calculated the prefix-name entropy of the four Scottish parishes using the original data (results presented in Table 3.5), and we again found the influence of pre-processing on entropy estimation to be minimal.

Parish	Entropy before pre-processing (bits)	Entropy after pre-processing (bits)
Beith	4.42	4.35
Dingwall	5.00	4.87
Earlston	4.85	4.85
Govan	4.84	4.70

Table 3.5: The prefix-name entropy of four Scottish parishes calculated using the original data and the pre-processed data (where we account for alternate spellings of the same name, name abbreviations, mistranscription, etc.). The result from the pre-processed data was the one shown in Table 3.1

The premodern Finnish name data are taken from records made before 1800 in the same dataset from which the modern Finnish name data were taken from [184]. The pre-1800 dataset comprises 1748523 birth records in total, with 1833 unique names, and again, spelling variations have been standardized as per [184].

3.4.2 Procedure

For each location, we calculated the entropy of the prefix-name distribution as a measure of the information conveyed by the distribution. For privacy reasons our modern datasets do not report name data below a certain population threshold. To account for these missing data, we estimated a lower bound and an upper bound of the true prefix-name entropy in each location.

As was described in the Results section, the entropy calculated from the datasets serves as a lower estimate of the true entropy. If a dataset contains N names from M_0 people, and the occurrence frequency of name i is f_i, then $M_0 = \sum_{i=1}^{N} f_i$, and the lower bound entropy of the dataset can be estimated by Equation 3.2 below:

$$H_{\text{low}} = -\sum_{i=1}^{N} \frac{f_i}{M_0} \log_2 \frac{f_i}{M_0}. \qquad [3.2]$$

Unlike our modern samples, the pre-modern Western datasets for Finland, Northern England, and Scotland are comprehensive, such that the entropy from Equation 3.2 is the true entropy of the population.

For the other datasets, an upper bound estimate is calculated as follows. First, we obtain the population of each locale from census data [211]. The population of the entire US, California, Delaware, and Vietnamese-Americans is extracted from the 2010 US census data [212–215]. The population of Taiwan is taken from data published by the Taiwan Department of Household Registration in January 2018 [216]. The population of South Korea is extracted

from the 2015 Korean census [217]. If the population of a locale is M, then there are $M - M_0$ persons whose prefix-name data are left out in our datasets. We assume each of these $M - M_0$ persons gets a unique prefix-name, which makes the total number of unique prefix-names $N + M - M_0$, and $f_i = 1$ for $i \in [N + 1, N + M - M_0]$. Therefore, the upper bound entropy can be calculated as

$$H_{\text{up}} = -\sum_{i=1}^{N+M-M_0} \frac{f_i}{M} \log_2 \frac{f_i}{M}. \quad [3.3]$$

The discrepancy between H_{low} and H_{high} will generally be higher when the data are more incomplete.

3.4.3 Study 2

Materials

As noted above, the combination of well-preserved genealogical sources and the fact that the transition from flexible to hereditary fixed bynames proceeded asymmetrically in Finland (hereditary bynames became dominant in eastern Finland earlier than western Finland [185]) mean that the Finnish name records described above [184] provide a unique opportunity for exploring the consequences of changing previously flexible bynames into the systems of hereditary fixed patronymic-bynames widespread today.

For the period 1600-1917, the dataset comprises 4908687 birth records in total, with 5377 unique prefix-names. Each birth record comprises the prefix-name of the baby, the parish where the baby was born and recorded, and, when available, a paternal byname. We divided the records into 5 different year bins: pre-1730, 1730-1780, 1780-1830, 1830-1880, and post-1880 (no data were excluded from this analysis).

Procedure

We focused on the following variables for each birth:

- **parish**: The church parish where the baby was born and recorded.
- **latitude**: The latitude of the parish.
- **longitude**: The longitude of the parish.
- **baby first name (`child_first_nameN` in the original data)**: The first/given/prefix-name of the baby.
- **father byname (`dad_last_nameN` in the original)** The byname, when available, of the father. This serves as a critical variable in that we use the choice to include or exclude the father's byname as reflective of the usage of fixed, patronymic-bynames in the parish. When the father's patronym (father's name, a separate variable in the original data) is reported, we do not include this.

birth year bin	count
up to 1730	214617
1730-1780	950686
1780-1830	1745407
1830-1880	1549536
post-1880	448441

Table 3.6: Birth year categories for Finnish data, with counts

We filtered to include birth years from 1600 on. That left us with data from 482 parishes, with birth years from 1600 to 1917. The birth years were broken into discrete categories, as shown in Table 3.6.

To compare equal-sized amounts of data, for each parish and for each birth year bin, we sampled 50 births.

From this data, we computed, for each parish and birth year group:

- the entropy over first names, computed as $\sum_i -p_i \cdot \log_2(p_i)$ (as per Study 1).
- the proportion of births that included the father's surname.

3.4.4 Study 3

Materials

The list for **U.S. scientist names** were extracted from the United States National Academy of Sciences [199], an honor society for academics in the sciences, broadly construed. The list contains living and deceased members. We extracted the part of the name that comes first as the flexible given name and the part that comes last as the fixed inherited name. We treated the name that appears between as a middle name, but did not consider it since typically (although not always) middle names are not used in the U.S.

The list for **Chinese scientist names** were extracted from the Chinese Academy of Sciences [195], a scientific honor society in China, containing living and deceased members. The list contains members elected as early as 1955. Names on the Wikipedia site appeared transliterated (as pinyin, Latinized script) and were already displayed in the Western name order, e.g., Zhiqin Xu. Most names consisted of two parts. We assume the part of the name listed first is the flexible given name (the byname in Chinese), and the second name listed is the fixed inherited name.

The list for **French scientist names** were taken from the French Academy of the Sciences, a scientific honor society in France. The list [197] contains "full members, adjunct members, corresponding members, and foreign associated members, of all times, since its foundation until today." We treated the part of the name that comes first as the flexible given name and the part that comes last as the fixed inherited name.

The list for **Finnish scientist names** were taken from the Finnish Academy of Science and Letters contains. We copied the list of "domestic members" from the site given [196]. The list of members is made publicly available by the Academy. There were typically two

elements to each Finnish name, and we treated the part that comes first as the flexible given name and the part that comes second as the fixed inherited name.

The list for **Russian scientist names** were extracted from the Russian Academy of Sciences, which contains names of the electees on its website [200]. The names are made publicly available by the Russian Academy of Sciences. We extracted the names "full members" from the site, which provides them in Cyrillic script. All subsequent analyses were conducted using this script. Names for Russian individuals typically included three elements: a flexible given name, a patronymic, and a fixed inherited name. We extracted the list into these three elements when possible. We consider names with and without patronymics in our analyses.

The list for **Korean scientist names** were extracted from the Korean National Academy of Science, which includes members from the humanities, social sciences, and natural sciences. We copied the publicly available names from the Academy's website [198]. Korean names often have a fixed inherited name and a flexible given name, where the flexible given name has two components that are separated by a space or hyphen when transliterated. In our dataset, the names are transliterated ahead of time. When the flexible given name is separated by a space or hyphen, we conduct two different analyses: treating it as a single given name or as two components. We make this decision since some Korean writers publishing under Western conventions use a double-barreled initialism with a full fixed inherited name (e.g., K.W. Chang for Chang Ki Won).

Three of our samples were obtained from Wikipedia, copied in a way consistent with Wikipedia's terms of service. The other 3 samples (for Korea, Finland, and Russia) were copied from the websites of their respective academies. In all cases, the names of members are intentionally provided to the public as public information and were accessed as intended.

Procedure

We used these datasets to obtain cross-cultural samples of names from the relevant communities. These were then processed and downsampled as described in Study 3 Results.

3.5 Data Availability

The American, Korean, and Taiwanese name census data, as well as the birth records of parishes in Scotland and northern England, have been deposited on Zenodo (https://doi.org/10.5281/zenodo.17644337) [218]. The Finnish birth records are available from [184]. The scientist name data are not available due to data privacy concerns, but can be accessed on request by contacting the corresponding author.

3.6 Code Availability

Code for reproducing analyses is available on Zenodo (https://doi.org/10.5281/zenodo.17644337).

3.7 Acknowledgements

Some parts of this manuscript are based on work previously presented at the 35th Annual Conference of the Cognitive Science Society and published in its Proceedings. We thank Amalia Spyromilio, Stella Schneider, Holly Jenkins, Evelina Fedorenko, Ted Gibson, and Damian Blasi for helpful comments. M.R.'s contribution was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG 547529231).

Chapter 4

The conceptual structure of laws leads to legalese

Note This chapter is adapted from a manuscript to be submitted for review:

Chen, S., Brown, T., Mollica, F., Martínez, E., and Gibson, E. (in prep), The conceptual structure of laws leads to legalese In preparation.

Abstract Legal texts are notoriously difficult to understand, and recent work has identified complex sentence structures as the key culprit. One explanation for the complexity of legalese is that of a historical artifact: an early version of laws was written in a convoluted way and then just edited in subsequent editions, preserving the complexity of the original, perhaps because the style was successful in achieving its goals. Another explanation is that the meanings that laws need to convey naturally lead to complex grammar. To evaluate these possibilities, drawing on corpora from the 7th century Tang Code, we examined linguistic complexity of legal texts in ancient China, which were created independently of the laws in Western civilizations. We found that classical Chinese legalese exhibits substantially greater grammatical complexity (longer inter-word dependencies) than non-legalese texts from the same time period, which suggests that the complexity may stem from the complexity of the intended meanings. However, our analysis of modern Chinese legalese texts show that grammatical complexity is much reduced compared to both the Tang Code and to modern American English legalese. This reduced complexity is likely due to modern-day Chinese lawmakers adopting alternative drafting strategies aimed at reducing syntactic complexity while conveying the same meaning, by breaking down single sentences into multiple simpler clauses. In summary, although the initial tendency of lawmakers may be to compose a rule (law) using a single sentence, which often leads to long "if ... then" dependencies, it is possible to express the same meanings in simpler ways.

Significance Statement American legalese heavily relies on complex grammar, especially long-distance inter-word dependencies. Here we show that the legal texts in ancient China, a civilization with distinct legal origins and traditions from the US, also use similarly complex sentence structures. In contrast, modern Chinese laws, which overlap in the kinds of meanings with the classical Tang Code, use much simpler sentences, demonstrating that meanings of

laws do not necessitate complex grammar. We hope these findings contribute to ongoing efforts to simplify legal texts in the US to make them more understandable for both lawyers and laypeople.

Keywords Legal language, dependency grammar, corpus studies, sentence processing

4.1 Introduction

Laws are formalized rules telling people what they can and cannot do, as well as the corresponding punishments if such rules are violated. In many countries, in order for these rules to be enforceable, they need to be clearly stated, so that laypersons are able to understand what exactly are prohibited. In the United States, this principle is stipulated in the fair notice doctrine [219] and the vagueness doctrine [220]. However, this seems to be not the case in the US in practice, as American legalese texts are found to be notoriously difficult to understand for laypeople [45,48]. Although there have been continuous efforts to simplify legal language from the top of the US Government [47,221], on the contrary, it is found that legal texts remain difficult to understand [222].

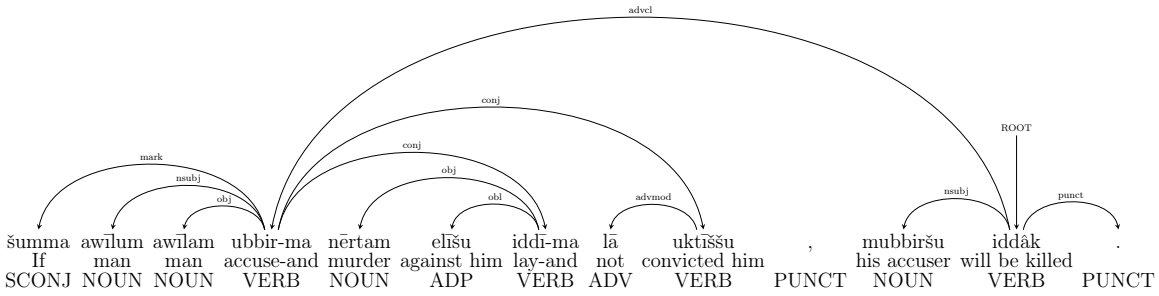
In [45], a large-scale corpus analysis showed that compared to non-legalese texts, such as academic papers and newspaper articles, legalese texts have a higher concentration of ALL CAPS usage, lower average word frequency, a higher concentration of passive voices, and a higher concentration of center-embedded syntactic structures [a clause embedded in the middle of another clause, such as "the dog {which chased the cat} ran away", [28]). A behavioral experiment in [45] suggested that among these features, center-embedding sentences hindered participant comprehension and recall to the most extent. This finding is consistent with previous studies which suggest that structures with long-distance syntactic dependencies, such as center-embedding, lead to comprehension difficulty [28–30,32,223]. This leaves us to wonder: what is the origin of such complexity in legal language, and why does it persist?

[224] evaluated several hypotheses for why the complexity of American legalese persists. Their results showed that lawyers preferred plain-English versions of legal texts and, like laypeople, found legalese harder to understand than equivalent non-legalese texts [225]. These effects provide evidence against several hypotheses: (a) that legalese helps lawyers justify fees by making legal work seem inaccessible [226]; (b) that legalese improves precision or enforceability [46]; or (c) that legalese becomes easier to process with legal training. [227] then examined whether legalese is reproduced even by people without legal training. They found that laypeople produced more convoluted language when asked to write official laws than when asked to write unofficial legal texts of equivalent conceptual complexity, suggesting that legalese persists both through copying from traditional templates and because its convoluted style signals authority.

Whereas prior work has sought to explain why the complexity of legalese persists, it remains unclear why it appeared in the first place. Here we introduce the **Content-of-Law Hypothesis** as a possible explanation. The concepts underlying laws are predominantly higher order predicates that specify relations among sets of events. One such relation is conditionality: a law consists of details of offenses (complex events) as preconditions, and

the associated punishment (further complex events) if the preconditions are met. A way to write laws that is isomorphic to the conditional conceptual structure of law is to convey such structure in a single sentence that leads with a conditional (e.g. "if...then..."). Such construction induces long-distance syntactic dependencies. For example, the structure is seen in the Code of Hammurabi [228], one of the oldest legal texts: in (4.1), Article 1 of the Code, the situation is outlined in a conditional clause led by a subordinate conjunction šumma (if), followed by the punishment. A long-distance dependency is created to connect the verb of the conditional clause *ubbir* with the main verb *iddâk*.

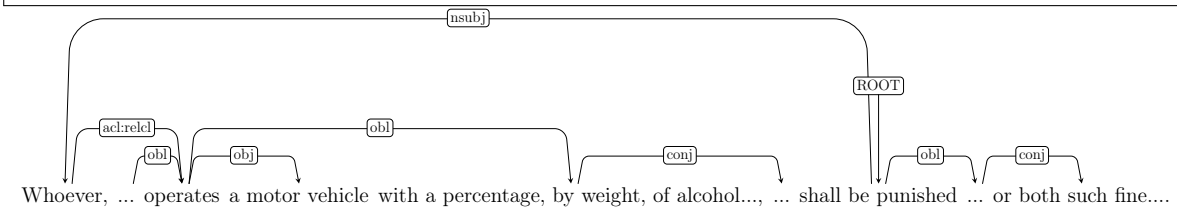
Source: The Code of Hammurabi, Article 1
Sentence [229]: **šumma** awīlum awīlam ubbir-ma nērtam elīšu iddī-ma lā uktīššu, mubbiršu iddâk.
Translation [229]: **If** a man accused a man and laid (a charge of) murder against him but has not convicted him, his accuser will be executed.



[4.1]

The dependency will be even longer if both the offense and the punishment are complex events, resulting in convoluted laws like the DUI provision in the Massachusetts General Laws (4.2). Here the condition includes details about the location of the offense, the substance involved, and the physiological state of the driver, which is inserted as clausal modifier of the subject There are also three potential options for the punishment. Long-distance dependencies are created to connect the subject with the main verb of the sentence and to connect all the details about the condition with each other.

Source: Massachusetts General Laws, Chapter 90, Section 24
Sentence: **Whoever,** upon any way or in any place to which the public has a right of access, or upon any way or in any place to which members of the public have access as invitees or licensees, operates a motor vehicle with a percentage, by weight, of alcohol in their blood of eight one-hundredths or greater, or while under the influence of intoxicating liquor, or of marijuana, narcotic drugs, depressants or stimulant substances, all as defined in section one of chapter ninety-four C, or while under the influence from smelling or inhaling the fumes of any substance having the property of releasing toxic vapors as defined in section 18 of chapter 270 **shall be punished by** a fine of not less than five hundred nor more than five thousand dollars or by imprisonment for not more than two and one-half years, or both such fine and imprisonment.



[4.2]

The above examples from the Code of Hammurabi and from the Massachusetts General Laws seem to be consistent with the Content-of-Law Hypothesis. However, they are also consistent with a hypothesis proposed in [227]: the **Copy-and-paste Hypothesis**, which says that new laws were drafted by copying and pasting from existing templates [230]. American legal tradition, or more generally, the Western legal tradition, date back to the New Canon Law in 11th century CE, which draws inspirations from legal traditions and thoughts in ancient Rome, ancient Greece, and ancient Israel [231]. Other sources also pointed out the influence of ancient Egypt and ancient Mesopotamia [232], where the Code of Hammurabi was the law. According to the Copy-and-paste Hypothesis, the complexity of American legalese texts may be a historical accident, as American laws may have inherited the template from the Code of Hammurabi, which was written in a royal dialect of Akkadian [228].

To evaluate the two hypotheses, we examine the earliest legalese texts in a different civilization to see if they shared the same way of writing. If they do, it will provide support to the Content-of-Law Hypothesis, and if they do not, it will support the Copy-and-paste Hypothesis.

Specifically, in this study, we analyze legalese and non-legalese texts from ancient China, a civilization with a distinct history and tradition. Western law's predecessors mostly had a divine origin, or in other words, these laws were said to have been given to mankind by a god or gods [233]. For example, in Plato's book *Laws* (Νόμοι), when a Cretan was asked by his companion from Athens on whether he would attribute the origin of their legal system to a god or to a man, he replied "to a god" (Θεός) and further elaborated that the god is Zeus in Crete and Apollo in Sparta [234]. In contrast, Chinese legal tradition seems to have an independent, secular origin: *Lü Xing* (Punishments of Lü, 呂刑), the legal code in early Zhou Dynasty and likely the earliest explanation of the origin of Chinese legal code [233], states

that during the reign of King Mu (*circa.* 950 BCE), frictions between different parts of society became severe, and the ruling class had to create written legal codes to keep the society in order. According to *Shang Jün Shu* (the Book of Lord Shang, 商君書), a foundational book on Chinese legalism, "as the people were numerous and wickedness and depravity arose among them, they [the sages] therefore established laws and controls and created weights and measures, in order thereby to prevent these things" [235]. *Huai Nan Zi* (The Master of Huai-nan, 淮南子) explicitly refutes a divine origin of Chinese law, stating that "Law is not a gift of Heaven, not a product of Earth. It was devised by humankind but conversely is used to rectify themselves" [236].

We compare the average dependency length per sentence between *the Tang Code* (Chinese: 唐律疏議), which is the oldest comprehensively preserved legal code in ancient China, and non-legalese texts from the same time period (the Tang Dynasty, 618-907 CE). The **Content-of-Law Hypothesis** predicts laws in the Tang Code exhibit longer dependencies than non-legalese texts. It will also predict that similar to the Code of Hammurabi, laws in the Tang Code also organize complex concepts into single sentences, because such a way of writing is isomorphic to the conceptual structure of law.

Laws being difficult to understand is hardly a new problem. Many legislatures have carried out reforms to make their laws more accessible to laypeople [44,221]. China, however, went through a more fundamental change in legal tradition: following the fall of the Qing Dynasty, the traditional legal system was abandoned, and a drastically different modern legal system was established [237–239]. While new laws were being drafted, the Chinese government was carrying out a reform to simplify the Chinese language in order to increase the literacy rate [240].

How effective was the reform? In the second part of the study, we analyze legalese and non-legalese texts in mainland China and compare their dependency lengths. If the reform is effective, we would expect no much difference in dependency length between modern Chinese legalese and non-legalese.

Mainland China is not the only society legislating in Chinese: so are Hong Kong and Taiwan. Although the three societies share a same language, legal traditions and practices in Hong Kong and Taiwan are heavily influenced by the United Kingdom and Japan, respectively [241,242]. Can foreign influence affect the complexity of legalese texts in Chinese by introducing syntactic features of other languages into Chinese? In the third part of the study, we analyze legalese texts in Hong Kong and Taiwan, and we compare the dependency length between these texts and non-legalese texts in Chinese.

The paper is organized as follows. We first analyzed legalese and non-legalese texts in classical Chinese. For classical Chinese, we analyzed texts in *the Tang Code* and 40 articles from the same time period as non-legalese controls. We found that legalese texts have longer dependencies than their non-legalese counterparts. Next, we analyzed legalese and non-legalese texts in modern Chinese, drawing from a law database and several other contemporary Chinese corpora such as Wikipedia and news. Here we found that legalese texts have longer dependencies than their non-legalese counterparts, but the difference is much smaller than what was observed in a previous study on American English [222]. We found the same for Hong Kong and Taiwanese legalese texts. We then discuss the implications of our findings.

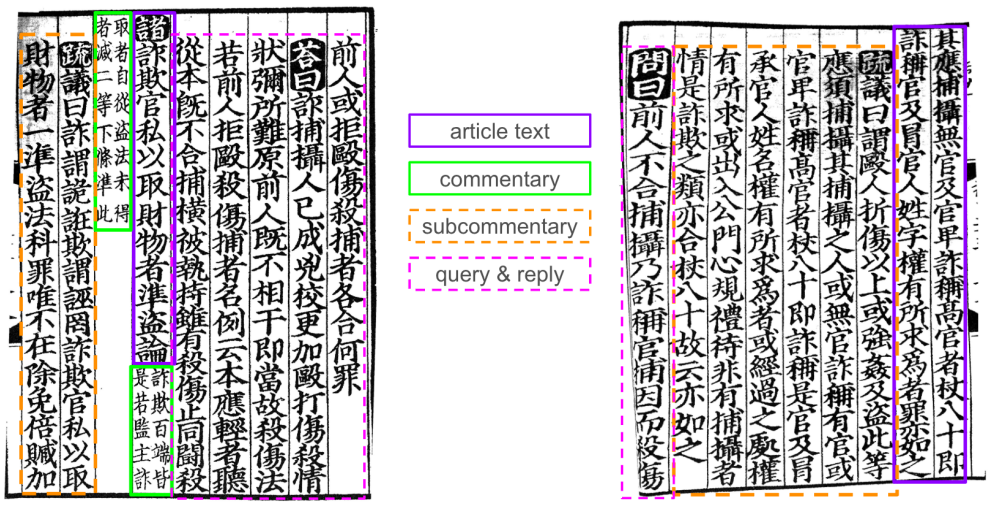


Figure 4.1: **An illustration of different texts in the Tang Code**. **Purple**: the article text, laying out the offenses and the punishments associated with each offense. **Green**: commentary, providing essential information to the article text. **Orange**: subcommentary, providing additional background information to the article text. **Pink**: query and reply, clarifying certain points in the article text or providing hypothetical scenarios. We only analyzed article texts and commentaries in this study. The images are scanned from a Song Dynasty (10th-13th Century) copy of the Tang Code, available on Wikisource under a CC BY-SA 4.0 License [245].

4.2 Dependency length comparison between legalese and non-legalese in classical Chinese

We analyzed legalese and non-legalese texts in classical Chinese. The legalese texts are taken from the Tang Code (*Tang Lü Shu Yi* 唐律疏議), the earliest legal code in ancient China that is comprehensively preserved [243], dating from the Tang Dynasty (618-907 CE). It also has profound influence on criminal investigation in China until the Qing Dynasty (1616-1911 CE) as later dynasties often adopted a considerable amount of the Tang Code as their own codes with little or no modifications [244]. An example of the Tang Code can be seen in Figure 4.1, with each aforementioned text highlighted. We randomly sampled 150 articles from the Tang Code (30% of the articles) for the analysis. The non-legalese texts are a collection of 40 political commentaries, narrative fictions, and biographical texts, all of which were composed during the Tang Dynasty (618-907 CE). Tables S1 and S2 in the Supplementary Information lists the name, the author, and the genre of the 40 non-legalese texts we selected. We randomly sampled 10 sentences from each article for the analysis.

We manually annotated and parsed all the sampled texts, largely according to conventions in Universal Dependencies [246]. A detailed description of our annotating and parsing convention can be found in Supplementary Information 1. The main variable of interest is

Table 4.1: Number of sentences and mean dependency length in each text type.

Genre	# sentences	Avg. dep. length
Legalese	614	2.62
Non-legalese	485	2.06

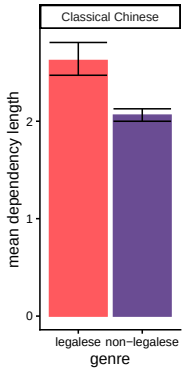


Figure 4.2: **Legalese texts in classical Chinese exhibit longer dependencies than non-legalese texts from the same period**. Mean dependency length per sentence in legalese and non-legalese classical Chinese texts.

the average dependency length per sentence, as dependency length is regarded as a proxy for comprehension difficulty [28]. In addition to average dependency length, we also considered average word frequency as another metric for comprehension difficulty [45]. The methods and results are shown in Supplementary Information 2. We did not find a significant difference in word frequency between legalese and non-legalese texts.

Table 4.1 lists the number of sentences and the mean dependency length in each type of texts in our analysis. Figure 4.2 shows the average dependency length for legalese and non-legalese texts in classical Chinese.

In classical Chinese, legalese texts on average have longer dependencies, compared to their contemporary non-legalese counterpart. We ran a mixed-effects linear regression using the lme4 package [247] in R [145], where the genre (legalese vs. non-legalese) were coded as fixed effects, with random intercepts by article. We found a significant main effect of genre ($\beta = -0.62$, SE=0.091, $t = -6.8$, $p < 0.001$).

What might have contributed to longer dependencies in legalese texts? Here we looked into the average length of each dependency, and how frequently they occur.

The results are presented in Figure 4.3. Among classical Chinese legalese texts, the longest dependencies are **det** (determiner), **advcl** (adverbial clause), **acl** (clausal modifier of nouns), and **conj** (conjunct). Within these dependencies, in non-legalese texts, **advcl** also occurs at a similar frequency. **conj** is used to a less extent, whereas **acl** and **det** are rarely used. Such disparities likely explain the difference in dependency length between the two genres.

Is there a common sentence structure in the Tang Code? When analyzing the text, we observed that apart from the first few articles that describe the general principles, virtually every article follows the same conditional structure: "諸...者", where the situation of the

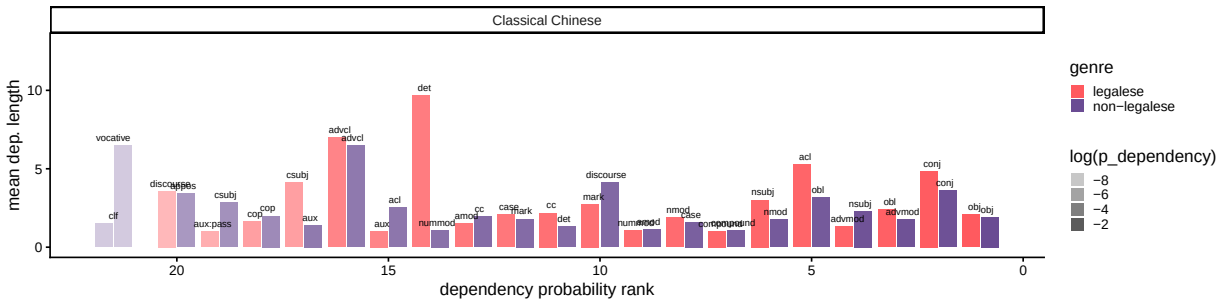


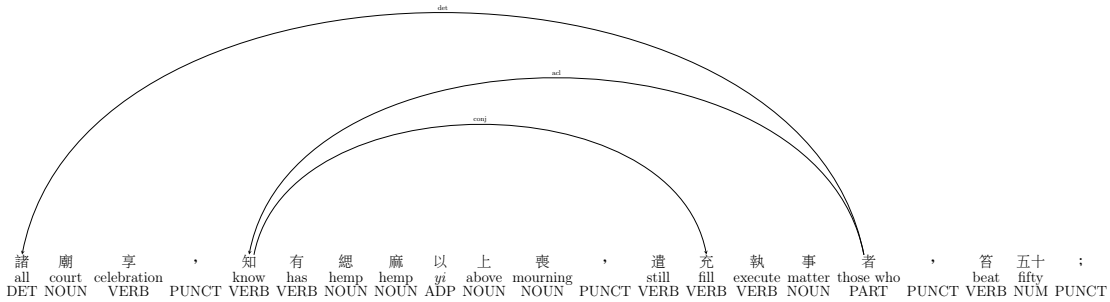
Figure 4.3: **Dependencies related to conditional relations contribute to the longer dependencies in classical Chinese legalese**. The average dependency length by dependency relation in classical Chinese. The dependencies are arranged by their relative frequency in each genre (left: less frequent, right: more frequent).

crime is described between the determiner "諸" (all) and the particle "者" (those who..., cases where...), followed by the associated punishment. This situation-punishment structure is consistent with the ones we have mentioned in the Code of Hammurabi (4.1) and the Massachusetts General Law (4.2).

In fact, this "諸...者" structure is likely the main contributor to the longer dependency length in the Tang Code: the particle "者" (cases) is the head of the clause describing the situation; "諸" (all) is the determiner modifying the head noun; and more importantly, all the details of the offense are inserted between "諸" and "者" as clausal modifiers. For example, in the sentence below (4.3), the details of the offense (during a court celebration, putting a person in charge of it while knowing the person is in a mourning period for a close relative) is presented as the clausal modifier of the particle "者". This insertion introduces three long-distance dependencies: **acl** (dep. length=12), which connects the top node of the clausal modifier "知" to the particle "者", **det** (dep. length=16), which connects the determiner "諸" to the particle, and **conj** (dep length=9), which connects the top node of the second coordinated element "充" with the top node of the first element "知". In contrast, in non-legalese texts, clausal modifiers do not appear as often as in the *Tang Code*, and determiners usually appear next to the noun they modify. In fact, in classical Chinese, clauses are rarely connected together via function words, as the conditions between them are either usually clear either in context or from their relative order [248]. For example, in (4.4), the four clauses describe a rebellion and its aftermath in a chronological order.

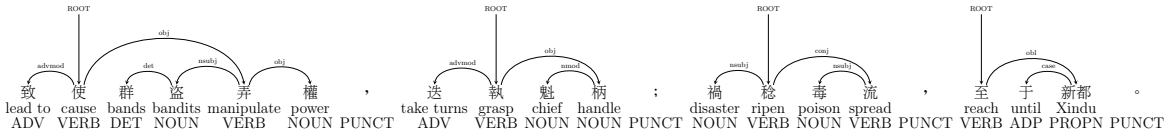
The results from classical Chinese seem to support the *Content-of-Law Hypothesis*: lawmakers in ancient times across cultures, from the Code of Hammurabi to the Tang Code, express complex conditional relations in single sentences. The results also imply that the complexity in American legalese texts did not originate from a historical accident by inheriting a peculiar writing style from the ancient times.

ID: 9:101:1:1
Sentence: 諸廟享，知有緦麻以上喪，遣充執事者，笞五十；
Translation [244]: All cases of putting a person in charge of a court celebration while knowing that he is in mourning of a relative within the fifth degree of mourning are punished by fifty blows with the light stick.



[4.3]

ID: liang-han-bian-wang:2:11
Sentence: 致使群盜弄權，迭執魁柄；禍稔毒流，至于新都。
Translation: It causes bands of rebels to seize power, repeatedly taking control of the country. Disorders accumulated, and evil spread across the whole country, extending even to when the Marquess of Xindu taking control.



[4.4]

4.3 Dependency length comparisons between legalese and non-legalese texts in modern mainland China

Law being difficult to understand is hardly a new problem, as many have complained about it. As King Edward VI of England (1537-1553) once said, "I wish that the superfluous and tedious statutes were brought into one sum together, and made more plain and short, to the intent that men might better understand them." [249] Legislatures in many countries have carried out reforms to make laws easier to understand. For example, in 1875, a committee was established in the House of Commons in the British Parliament to explore ways to improve the language used in statutes [44]. Since 1975, the US Federal Government also started the Plain English Movement to advocate the usage of plain English [221]. However, not all of these reformations are successful. In the US, despite the top-down effort to simplify legal texts, they remain difficult to understand [227].

Language	Genre	# sentences	Avg. dep. length
Chinese	Legalese	1556022	3.16
	Non-legalese	13876111	2.97
English	Legalese	1308424	3.21
	Non-legalese	6074672	2.72

Table 4.2: Number of sentences and mean dependency length in each text type. Results from American English are pulled from [222], as a comparison of the effect size.

In China, however, there was a more fundamental change in the legal system. The traditional legal system, exemplified by the Tang Code, was practiced until the end of the 19th Century [239], when the Qing government struggled to modernize China following the invasions of Western power. Then, the traditional legal system was abandoned as the Qing Dynasty fell in 1911 [237]. The subsequent legal systems implemented by the Chinese Communist Party since 1949 were drastically different from the traditional one [238,239]. The Communist government placed great emphasis on the simplicity of language. For instance, the government carried out a campaign to simplify Chinese characters in order to boost the literacy rate among people [240]. The same seems to be true in legal language too: during the drafting stage of the Constitution, two linguists (Ye Shengtao and Lü Shuxiang) served as consultants on language [250], and later in 2007, the Legislative Affairs Commission in the National People's Congress established a committee consisting of linguists to proofread draft bills to be voted on by the Congress [251].

How effective is the simplification effort in Chinese law? In this study, we analyzed legalese and non-legalese texts in mainland China and compared the average dependency length across the two genres. If laws are successfully made less convoluted, we should expect similar dependency length between the two genres.

The legalese texts are taken from an online database maintained by the National People's Congress (全國人民代表大會, NPC) of the People's Republic of China[1]. We included both statute laws (those passed by the NPC) and sub-statutory documents (e.g. administrative regulations). We only analyzed legal documents that are currently in effect. The non-legalese texts in our analysis consist of multiple genres and are taken from various corpora, such as Chinese Wikipedia, news, and WeChat articles.

Texts in modern Chinese are annotated and parsed by Stanza (Version 1.10.1) [252], one of the state-of-the-art part-of-speech taggers and parsers for Chinese texts. The dependent variables here are identical to those in the classical Chinese analysis.

Table 4.2 lists the number of sentences and the mean dependency length in each type of texts in our analysis. Figure 4.4A shows the average dependency length for legalese and non-legalese texts in classical Chinese. Legalese texts in modern Chinese on average also appear to have longer dependencies, compared to non-legalese texts. To account for the uneven sample size in legalese and non-legalese data, we randomly sampled from the non-legalese corpus the same number of sentences as in the legalese corpus. We then combined the sampled non-legalese sentences with all legalese sentences and fitted a linear regression

[1]國家法律法規數據庫, https://flk.npc.gov.cn/

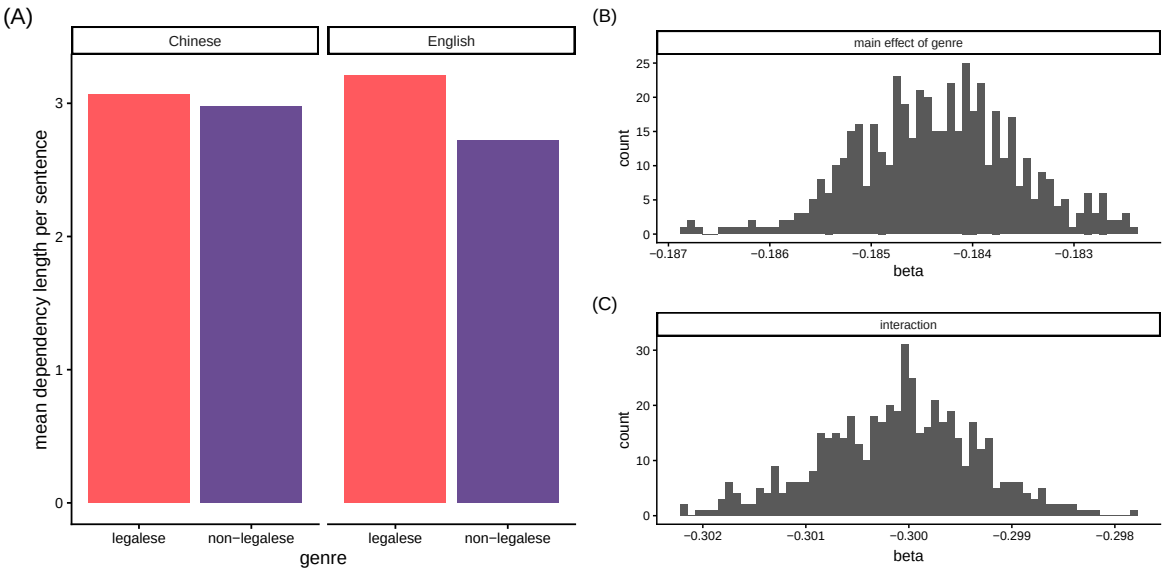


Figure 4.4: **Modern Chinese legalese texts from mainland China exhibit longer dependencies than non-legalese texts, but with a smaller effect than in American English**. (A) Mean dependency length per sentence for modern Chinese. Results from American English are pulled from [222], as a comparison of the effect size. (B-C) Results from the sub-sampling analysis: (B) Distribution of genre effect estimates (beta) obtained from 500 comparisons between the legalese data and the size-matched subsamples of the non-legalese data. (C) Distribution of interaction between genre and language, obtained from 500 comparisons between the legalese data and the size-match subsamples of the non-legalese data, along with legalese and non-legalese data from a previous study on American English [222].

predicting dependency length from genre. We set legalese as the reference level. The effect size (β) was recorded, and the entire procedure was repeated 500 times. The results are shown in Figure 4.4B: the effect of genre is consistently lower than zero and is distributed around -0.184. Hence, we can say that the dependency length in modern Chinese legalese is significantly higher than in modern Chinese non-legalese.

However, since our database contains tens of millions of sentences, the statistically significant difference in dependency length we found might not be meaningful. To assess whether the effect size is simply an artifact of our large sample size, we drew legalese and non-legalese texts in American English from [222] as a comparison. American legalese texts have been found to have longer dependencies than their non-legalese counterparts [227] and to be harder to understand [45]. We therefore compare the mean dependency length across genres as well as across languages. If there is a main effect of genre but no interactions between language and genre, then the difference we found in our analysis reflects a general pattern and is therefore more likely to be meaningful. On the other hand, if there is an interaction between language and genre, especially if the interaction effect size is much larger than the main effect size, we cannot conclude that such difference in dependency length will necessarily imply that modern Chinese legalese texts are more difficult to comprehend than their non-legalese counterparts.

The American legalese texts were taken from [222] and consisted of every public law, private law, concurrent resolution, and proclamation issue by the U.S. Federal Government between 1951 and 2009. The American English non-legalese texts were taken from the Corpus of Historical American English [253], which includes fiction, nonfiction, news, and magazine articles published between 1951 and 2009. We filtered out sentences that are fewer than 10 words in length, in order to remove sentence fragments such as titles. The dependency length was calculated in the same way as in the analysis of modern Chinese. Table 4.2 showed the number of sentences in each genre.

The American English data is presented in the right panel of Figure 4.4A. We conducted another sub-sampling analysis as before, but here we also added data from the American English study. We set Chinese and legalese as the reference levels. We repeated the procedure 500 times and plotted the distribution of the interaction. The results are shown in Figure 4.4C: the interaction is consistently lower than zero and is distributed around -0.3. This result suggests that although modern Chinese legalese texts have significantly longer dependencies than modern Chinese non-legalese texts, the difference is much smaller compared to that observed in American English.

In addition, results related to the average word frequency are presented in Supplementary Information 2: contrary to American English [45], we instead found that the average word frequency in modern Chinese legalese texts are instead higher than in non-legalese texts, indicating that laws written in Chinese are less likely to contain low-frequency legal jargon.

Notably, when manually inspecting laws in mainland China, we found that not all the laws follow the "situation-punishment" structure. When the factual conditions become lengthy, some provisions first state the punishment framework and then place the covered acts in a trailing list, as in (4.5).

Taken together, our results seem to suggest the *Content-of-Law Hypothesis*. Law drafters in ancient China, just like those in ancient Babylon, followed an intuitive logical structure to draft laws: describing the situations of an offense, and then the associated punishment.

However, lawmakers in mainland China possibly recognized the comprehensibility problem of organizing information with such structure and started to explore other ways to express the same information in order to facilitate readability.

Genre: legalese - statute laws in mainland China
Source: 中華人民共和國刑法
Sentence: 違反藥品管理法規，有下列情形之一，足以嚴重危害人體健康的，處三年以下有期徒刑或者拘役，並處或者單處罰金；對人體健康造成嚴重危害或者有其他嚴重情節的，處三年以上七年以下有期徒刑，並處罰金：
(一）生產、銷售國務院藥品監督管理部門禁止使用的藥品的；
(二）未取得藥品相關批准證明文件生產、進口藥品或者明知是上述藥品而銷售的；
(三）藥品申請註冊中提供虛假的證明、數據、資料、樣品或者採取其他欺騙手段的；
(四）編造生產、檢驗記錄的。

Translation: Whoever violates pharmaceutical administration regulations and commits any of the following acts, thereby sufficiently endangering human health, shall be sentenced to fixed-term imprisonment of not more than three years or criminal detention, and shall also, or alternatively, be fined; where serious harm to human health is caused or other serious circumstances exist, the offender shall be sentenced to fixed-term imprisonment of not less than three years but not more than seven years, and shall also be fined:
(1) producing or selling drugs prohibited for use by the drug regulatory department of the State Council;
(2) producing or importing drugs without obtaining the relevant approval documents for pharmaceuticals, or knowingly selling such drugs;
(3) providing false certifications, data, materials, samples, or using other deceptive means in the course of pharmaceutical registration applications;
(4) fabricating production or inspection records.

[4.5]

4.4 Dependency length comparisons between Hong Kong/-Taiwan legalese and Mainland China non-legalese texts

So far, we have only analyzed legalese texts in mainland China, but mainland China is not the only legislatures that draft laws in modern Chinese: legislatures in Hong Kong and Taiwan also do. Although they share the same language, legal traditions and practices in Hong Kong and Taiwan are different from those in mainland China, as the system in Hong Kong has strong British influence [241] and that in Taiwan has strong Japanese influence [242], while mainland China was under Soviet influence [237]. Moreover, laws in Hong Kong are drafted in both English and Chinese, and texts in both languages are regarded as conveying the

Language	Genre	# sentences	Avg. dep. length
Modern Chinese	Laws - mainland China	1556022	3.16
	Laws - HK	177363	3.19
	Laws - Taiwan	652191	2.82
	Non-legalese	13876111	2.97
English	law	1308424	3.21
	Non-legalese	6074672	2.72

Table 4.3: Number of sentences and the mean dependency length in each type of text. Acronyms: HK - Hong Kong, TW - Taiwan.

same meaning [2]. It is possible that to adapt to English and to concepts in a common law system, the Chinese version could be influenced by syntactic features in English [254] and potentially result in convoluted sentences. Given these differences in legal traditions and drafting environments, we ask whether dependency length also differs across legal Chinese in these three societies.

We analyzed all the laws passed by the Hong Kong and Taiwanese legislature, as well as all the sub-statutory regulations from mainland China. The legalese texts from Hong Kong are taken from an online database maintained by the Hong Kong government[3]. The database consists of the Basic law (基本法) and other law texts drafted by the Department of Justice (律政司) or members of the Legislative Council (立法會成員), passed by the Legislative Council (立法會), and signed into law by the Chief Executive Officer (行政長官). The legalese texts from Taiwan are taken from an online database maintained by the Taiwanese government[4]. The database consists of the Constitution and other law texts proposed by the 5 government branches, passed by the Legislative Yuan (立法院) and signed into law by the President. We only included texts that are currently in effect.

Table 4.3 shows the number of sentences, and Figure 4.5A shows the mean dependency length by genre in legalese and non-legalese texts. We compared the effect sizes between legalese and non-legalese texts across different types of modern Chinese legal texts, using the difference observed in American English as a reference point. For each comparison, we followed the same sub-sampling approach as before: we combined legalese texts, a size-matched subsample of non-legalese texts, and data from [222]. We set the legalese and Chinese as the reference level in genre and language, respectively. We ran a linear regression and recorded the interaction between genre and language. We then repeated the whole process 500 times.

Results for Hong Kong legalese texts are presented in Figure 4.5B. We found a significant main effect of genre: the β values are consistently below zero and are centered around -0.22. More importantly, there is also a significant interaction between language and genre, as the β values are consistently below zero and centered around -0.264. Figure 4.5C shows the results for Taiwanese legalese texts. We found a significant main effect of genre going to

[2]See https://www.elegislation.gov.hk/interpretbileg
[3]電子版香港法例, https://www.elegislation.gov.hk/
[4]全國法律資料庫, https://law.moj.gov.tw/

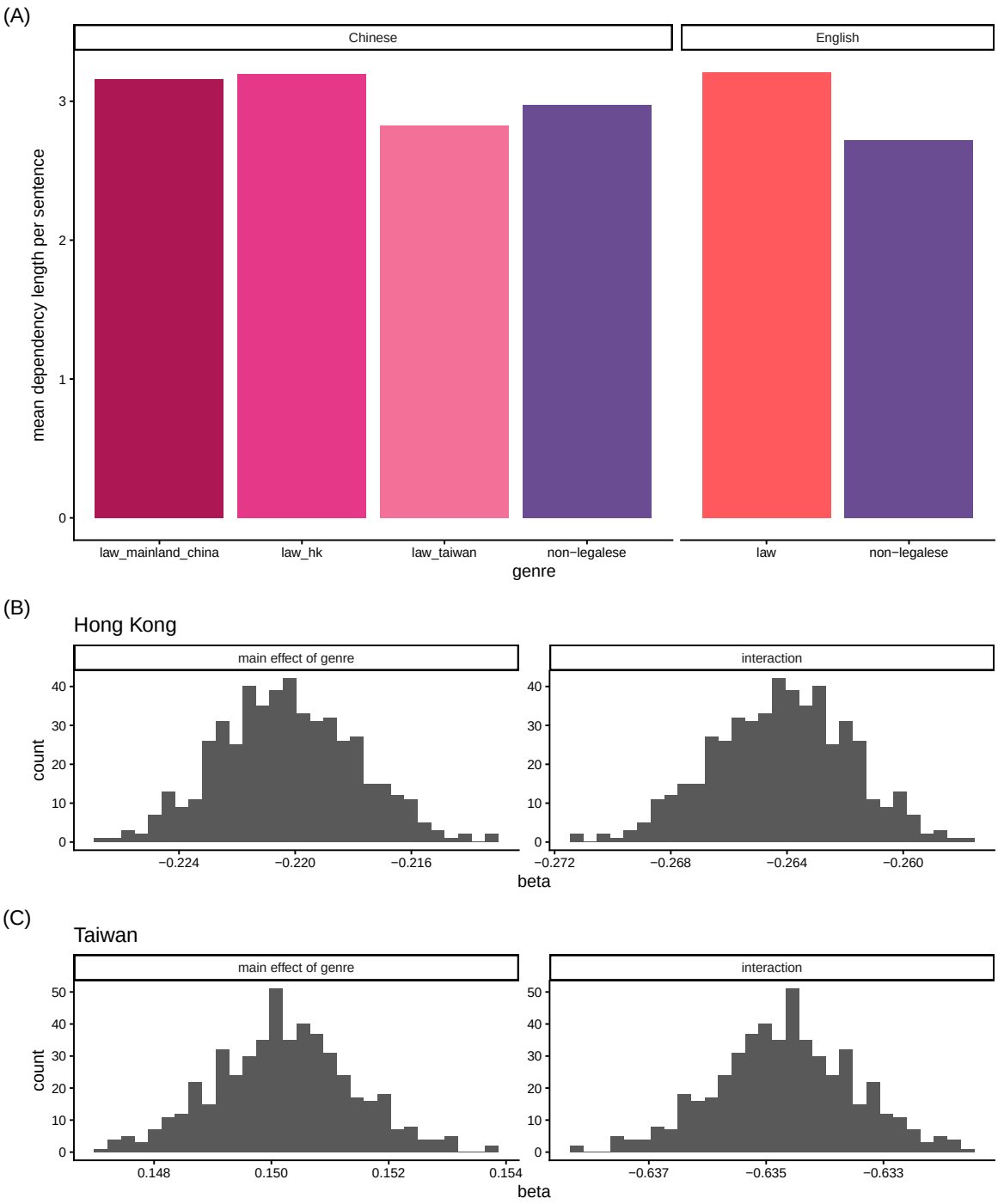


Figure 4.5: **Legalese texts from Hong Kong and Taiwan exhibit longer dependencies than non-legalese texts, but also with a smaller effect than in American English**. (A) Mean dependency length per sentence for legalese texts from Hong Kong and Taiwan, along with non-legalese texts in modern Chinese. Results from American English are pulled from [222], as a comparison of the effect size. Acronyms: HK - Hong Kong. (B) Distribution of (left) main effect of genre and (right) interaction between genre and language, obtained from 500 comparisons between the Hong Kong legalese data and the size-match subsamples of the non-legalese data, along with legalese and non-legalese data from a previous study on American English [222]. (C) Same as (B) but the legalese data is from Taiwan.

the opposite direction: the β values are consistently above zero and are centered around 0.15, which suggests that Taiwanese legalese texts on average exhibit shorter dependencies than non-legalese texts in modern Chinese. In addition, we found a negative interaction between language and genre: the β values are consistently below zero and are centered around -0.63. This indicates that British influence and specificity of law are likely to influence the dependency length, but the resulting difference between legalese and non-legalese texts is still smaller than that in American English. On the other hand, interestingly, legalese texts in Taiwan have shorter dependencies than Chinese non-legalese texts. From Appendix C.3, similar to the findings above, all types of legalese texts show higher word frequency than non-legalese ones.

4.5 Discussion

American legalese texts are more difficult to understand compared to their non-legalese counterparts [46,48] because of long-distance syntactic dependencies [45,222]. Previous studies [227] suggest that such long-distance dependencies may be a result of copying and editing from old templates, or it may be a way to signal authority of law, but it is still unclear why laws are written in this way in the first place. Here we offered an explanation for the origin of long-distance dependencies in legal documents, termed *Content-of-law Hypothesis*: laws are predominantly conditional, and long-distance dependencies are a result of lawmakers writing conditional statements in single sentences. In this study, to evaluate these hypotheses, we analyzed legalese and non-legalese texts in classical and modern Chinese, as Chinese legalese traditions had an independent origin from the Western ones, and China has different legal traditions and practices. We found that legalese texts on average show longer dependencies than non-legalese texts in classical Chinese. The same was also found in legalese texts in mainland China, but the difference was much smaller than in American English [222]. By also analyzing sub-statutory regulations in mainland China and laws in Hong Kong and Taiwan, we found that foreign influences and specificity of law may affect the dependency length, but their effects are still small to explain the effect size in American legalese.

Our results seem to be consistent with the *Content-of-law Hypothesis*. Concepts in laws are largely conditional in nature: to write a law, one needs to outline the situation of an offense, and the associated punishment if the offense is committed. A syntactically transparent way to convey this information is to use conditional such as "if...else...". Lawmakers in ancient China and ancient Babylon likely adopted such a structure because it provides a structurally isomorphic way to link offenses to their corresponding punishments. However, conditional constructions create long-distance dependencies between the offense and the punishment, and these dependencies become even longer as the description of the offense becomes more complex. This in turn could explain why the Tang Code exhibits longer average dependency than non-legalese texts in classical Chinese. It is possible that in the 20th Century, the Chinese government placed emphasis on simplifying the Chinese language to increase the literacy rate, and under this political climate, lawmakers in mainland China recognized the comprehensibility problem of fitting complex information into a simple structure and started to explore alternative ways to express the same information in order to facilitate readability. One such alternatives is to describe the situations of an offense at the end of the article in

bullet points [45].

What makes the simplification effort successful in some legislatures but not in others? Why are dependency lengths of mainland Chinese and Hong Kong laws similar to those in non-legalese texts, whereas dependency lengths in US laws are substantially higher than in US non-legal texts? One possible explanation is that not all legislatures have identified long-distance dependency as the primary source of complexity in legal language. Some legislatures have issued drafting guidelines aiming at reducing the dependency length. For example, the guidance from the British Parliament[5] contains instructions such as "avoid inserting words between the subject and the main verb (p.3)", avoid using complex if-clauses before the main clause (p.25), and break conditions and exceptions into separate propositions, instead of one (p.28). The guidance from the Hong Kong Legislative Council[6] also instructs bill drafters to "keep related words as close together as possible (p.89)" in English and "相關字詞應盡量靠近 (p.83)" in Chinese. In contrast, others have not recognized the crux of the problem: although the US House of Representatives also has a guidance on law drafting, which also advocates using plain language[7], the guidance simply contains general instructions to use simple sentences, without more detailed and more actionable ones like in the UK and Hong Kong. It is possible to make American legalese texts easier to understand once there are instructions targeted at reducing the dependency length.

It could be observed by comparing Figure 4.2 and Figure 4.4 that modern Chinese has overall longer average dependency length than classical Chinese. We speculate that this could be due to the following reasons. First, on a grammatical level, conversion (using a word in a different part of speech) is prevalent in classical Chinese, but in modern Chinese, additional words are needed to convey the same meaning. For example, in the sentence "公奇之", the word "奇 (strange)" is normally used as an adjective but is used here as a verb, meaning "regard something as strange". However, such usage is no longer present in modern Chinese, so to convey the same meaning, a verb is needed to form the construction "認為...奇怪". Second, on a technical level, Stanza [252] may over-segment lexicalized compounds, treating words like "雞蛋 (egg)" as two separate words rather than a single one. Both boosted the dependency length in modern Chinese by introducing additional words. Third, Stanza connects syntactically independent clauses together in the modern Chinese analysis, while we did not in the classical Chinese analysis. This resulted in the dependency lengths in modern Chinese being inflated (See Appendix C.2.3 and C.4 for details). In our study, we attribute the small observed difference in the dependency lengths in modern Chinese to the simplification of legalese texts. However, it can also partially be due to the complexification of non-legalese texts. In the 1910s, there was a movement to introduce Western technology and culture into China. Amid the movement, many scholars and writers such as Lu Xün and Hu Shih advocated Europeanizing modern Chinese grammar [248], hoping that doing so would boost the ability of Chinese to express new ideas and thoughts [255]. Some features introduced in this period include adding aspect markers such as "了" after verbs to signal completeness of events, inserting additional information parenthetically after a word, using "的" construction to introduce a clause modifying a noun, and explicitly combining clauses

[5] https://www.gov.uk/government/publications/drafting-bills-for-parliament

[6] https://www.doj.gov.hk/en/publications/pub20030011.html

[7] https://legcounsel.house.gov/sites/evo-subsites/legcounsel-evo.house.gov/files/documents/ManualDraftStyle_2022.pdf

with conjunctions. All these features could potentially lengthened the dependency in modern Chinese. Future studies could evaluate the relative contribution of these two forces.

In conclusion, in our study, we analyzed legalese and non-legalese texts in both classical Chinese and modern Chinese. We found that legalese texts in classical Chinese are more convoluted than non-legalese texts, while the difference in modern Chinese is much smaller compared to what was observed in American English. Our findings seem to indicate that the convoluted structures in legalese texts are likely a consequence of expressing complex relations in single sentences. However, such a convoluted structure is neither easy to understand nor required to establish authority of law, which may explain why countries have attempted to simplify legal language. We speculate that efforts of simplification are most likely to succeed when lawmakers correctly identify the source of the complexity.

4.6 Materials and Methods

4.6.1 Classical Chinese analysis

Materials

First, we analyzed legalese and non-legalese texts in classical Chinese. The legalese texts are drawn from the Tang Code. It consists of 30 chapters (*Juan* 卷). A total of 502 articles (*Tiao* 條) are grouped into these 30 chapters, each covering the punishment of a specific offense within the chapter's topic. In additional to the article text itself, each article may contain commentaries (*Zhu* 注), subcommentaries (*Shu* 疏), or queries and replies (*Wen da* 問答). Commentaries serve as appendices to the article texts, often providing essential information to the articles, such as specifying the amount of copper for a person to be bailed out when the article texts only mention the punishment. It has been suspected that commentaries had already been an integral part of the text in the Sui Dynasty legal code, instead of an addition during the Tang Dynasty [244,256]. Subcommentaries provide further information on the article texts by giving definitions on the key terms, relevant background on a specific article, or providing references to other articles within the Tang Code. Queries and replies clarify certain points in the article texts or provide hypothetical cases not anticipated in the article texts [244]. Since commentaries are written inline with the article texts and provide essential information to the law, whereas subcommentaries and queries and replies appear in a separate paragraph and provide additional information, in this study, we excluded subcommentaries and queries and replies from our analysis and only focused on the article texts and commentaries. We randomly sampled 150 articles from the Tang Code (30% of the articles) for the analysis.

The non-legalese texts are a collection of 40 political commentaries, narrative fictions, and biographical texts, all of which were composed during the Tang Dynasty (618 - 907 CE). Tables S1 and S2 in the Supplementary Information lists the name, the author, and the genre of the 40 non-legalese texts we selected. We randomly sampled 10 sentences from each article for the analysis. For articles shorter than 10 sentences, we annotated and parsed the entire article.

Annotation and parsing

We manually annotated and parsed all the sampled texts, largely according to conventions in Universal Dependencies [246]. Specifically, we first segmented the texts into words. In most cases, one word corresponds to one character, except numbers and proper nouns such as "二十三 (23)" and "新都 (Xin Du)" are treated as one word. Then, we annotated each word's universal part-of-speech tag (e.g. whether it is a noun, a verb, or an adposition, etc.). In the parsing stage, we identified each word's head and its universal dependency relation to the head. The detailed conventions are described in Appendix C.2.

Dependent variables

The main variable of interest is the average dependency length per sentence, as dependency length is regarded as a proxy for comprehension difficulty [28]. Dependency length is defined as the absolute difference between the ID of a word and the ID of its head. For example, if the head of the 6th word of a sentence is the 19th word, then its dependency length will be 13. Following [33], two dependency relations, **ROOT** and **punct**, are removed from the calculation. In addition to average dependency length, we also considered average word frequency as another metric for comprehension difficulty [45]. The methods and results are shown in Supplementary Information 2. We did not find a significant difference in word frequency between legalese and non-legalese texts.

4.6.2 Modern Chinese analysis

Materials

The legalese texts are taken from an online database maintained by the National People's Congress (全國人民代表大會, NPC) of the People's Republic of China[8]. This database includes all the legal documents drafted, passed, and enacted by the Chinese government since 1949. In this analysis, we analyzed the **Constitution of the People's Republic of China** (中華人民共和國憲法), along with other **statute laws** (法律), defined as documents proposed and passed by the NPC and signed into law by the President. We also analyzed **sub-statutory documents**, including 1) Administrative regulations, local laws and regulations supervisory regulations (行政法規, 地方性法規, 監察法規律), which contain documents drafted by the State Council (國務院) and signed into law by the Premier (總理), documents drafted and passed by regional legislative bodies (地方人民代表大會), and documents drafted by the State Supervisory Council (國家監察委員會); 2) Judicial and statutory interpretations (司法解釋, 法律解釋), which are interpretations issued by the Supreme People's Court (最高人民法院) or by the Supreme People's Procuratorate (最高人民檢察院); and 3) Decisions to amend or repeat laws (修改、廢止的決定), and Decisions on Issues concerning Law and Major Issues (有關法律問題和重大問題的決定) - decisions to amend or repeal laws or on certain major issues, often issued by the NPC or the State Council. We only analyzed laws that are currently in effect.

[8]國家法律法規數據庫, https://flk.npc.gov.cn/

We included both statute laws and sub-statutory documents because mainland Chinese statute laws are drafted in a general and abstract form [257], whereas their detailed implementations are outlined in sub-statutory texts such as administrative regulations, regional laws, and interpretations [238,257]. For example, drunk driving is an offense stipulated in the Criminal Law (刑法), but the text merely states that those who drunk drive will be imprisoned and fined, without elaborating what constitutes as drunk driving and the length of the imprisonment. Such information is instead outlined in an interpretation document jointly issued by the executive and judicial branches of the government. Including sub-statutory documents ensures that the content coverage of our analysis on Chinese law is similar to that on American law [45,222].

The non-legalese texts in our analysis consist of multiple genres and are taken from various corpora: Chinese Wikipedia[9], WeChat articles[10], newspaper articles[11], hotel reviews[12], novels[13], and speeches[14]. For both legalese and non-legalese texts in modern Chinese, following [227], we removed sentences that are shorter than 10 characters, in order to exclude sentence fragments such as titles and credits. In addition, for each genre of non-legalese texts, we removed sentences that are above 95-percentile in length, in order to filter out junks such as references, hyperlinks, and list of names.

Annotation and parsing

Texts in modern Chinese are annotated and parsed by Stanza (Version 1.10.1) [252], which is among the state-of-the-art part-of-speech taggers and parsers for Chinese texts. To assess the reliability for Stanza, we randomly selected 191 sentences (6850 dependencies) from our dataset. We then analyzed their dependency relations by hand, and compared the results with Stanza. We found that although the results generated by Stanza only matched our results 60% of the time, the average dependency length produced by the two sources were largely similar. Additional details are provided in Supplementary Information 3.

Dependent variables

The dependent variables in this study are identical to those in the classical Chinese analysis. To calculate the average dependency length, in addition to **ROOT** and **punct**, we also excluded **parataxis** and **flat:foreign** from the calculation, as the former tends to link syntactically independent sentences, thus inflating the average dependency length, and the latter falsely attaches all subsequent non-Chinese words to the first non-Chinese word. We also removed **advcl** from our calculation if it is not licensed by a subordinate conjunction

[9]Wikipedia corpus - Chinese, https://huggingface.co/datasets/wikimedia/wikipedia/viewer/20231101.zh

[10]WeChat public corpus, https://github.com/nonamestreet/weixin_public_corpus

[11]BAAI Industry Corpus - News media, https://data.baai.ac.cn/datadetail/IndustryCorpus2_news_media

[12]Chinese sentiment corpus - hotel reviews, https://www.kaggle.com/datasets/marquis03/chnsenticorp-htl-all

[13]We drew texts from a web novel corpus: Guofeng Webnovel, https://github.com/longyuewangdcu/GuoFeng-Webnovel, as well as from 3 famous Chinese novels: *Dawn Blossoms Plucked at Dusk* (朝花夕拾) by Lu Xun (魯迅, 1881-1936), *Fortress Besieged* (圍城) by Qian Zhongshu (錢鐘書, 1910-1998), and *In the Name of the People* (人民的名義) by Zhou Meisen (周梅森, 1956-)

[14]see Table S3 in the Supplementary Information

(e.g. “如果/if”) or by particle “的”, which appears often in modern Chinese legalese texts in clauses describing situations [258].

4.7 Data Availability

All the data, scripts, and figures are available online at https://github.com/cshnican/chinese_law.

4.8 Acknowledgments

This work is supported by the MIT Human Insight Collaborative (Award Number: 2412846). We are grateful to Rachel Zhu, Zhenyi Zhou, Yuyang Zhang, Xuewen Zhang, Deeraj Pothapragada, Yijun Lyu, Waiwa Kou, Annie Guo, Nicole Ding, Jonathan Dase, Iris Chen, and Hanxiao Bo for annotating and parsing texts in classical Chinese.

Chapter 5

Conclusion

In this thesis, we investigated how the pressure towards communicative efficiency can potentially be interfered by top-down institutional constraints, with three case studies: spatial demonstratives (Chapter 2), personal names (Chapter 3), and legal language (Chapter 4).

In Chapter 2, we studied a baseline case where efficient communication seems to be the dominant pressure. Our results suggest that languages from diverse families and locations can describe locations of different entities with relatively high informativity, with relatively few spatial demonstratives. While there seems to be no institutional constraints on how spatial demonstratives are used, there may still be other factors at play. For example, the surrounding environment of language users can possible shape spatial demonstratives in different languages to adapt to the communicative need: if describing spaces in more details is more useful, a language will have more complex spatial demonstrative system, and vice versa. The varying communicative need in different languages about space is likely affected by the physical and cultural environment. For example, Tzeltal uses an "uphill-downhill" system to describe space, instead of systems like "here" and "there" [259], because the hill is a salient component in the lives of its speakers, as they live on an inclined hill. However, it still remains unclear what affects a language's relatively weight of informativity and simplicity when describing space. In color naming, for example, it would probably be the prevalence of objects that differ only in color [16]. Future studies could investigate what might affect the trade-off between informativity and simplicity in spatial demonstratives. In addition, we show that not all the communicatively efficient spatial demonstrative systems are attested in our datasets, and the attested ones are predominantly consistent, suggesting that factors such as ease of learning are also at play [127,260,261]. Interestingly, the extent to which the structure of semantic systems can be explained by communicative efficiency and consistency seems to vary across domains [262]. Future studies could also examine why certain semantic domains value consistency over communicative efficiency more than others.

In Chapter 3, we presented a case where an efficient information structure of personal names was altered by name regulations in various countries. Names across cultures historically shared a similar structure, possibly as a result of efficiently distinguishing a large number of individuals with a small number of unique words. Later, regulations on personal names were implemented for governments to efficiently collect taxes and track properties [43]. As a result, we show that the information structure in personal names has been altered in the West: the entropy of prefix-names increased in order to keep names distinctive in a growing population.

This may have made names harder to use. Future studies could experimentally compare the difficulty of remembering and retrieving names under a historical and current information structure. In addition, our study only examined the change in information structure among Western names due to data availability, but the same prediction holds for East Asian Cultural Sphere (EACS) names: we would expect an increase in byname entropy after the regulation on names was implemented in the Qin Dynasty (3rd Century BCE). Future studies may potentially test this prediction by examining prefix-name and byname distributions in tax records from the Zhou Dynasty (11th - 3rd Century BCE) and Qin Dynasty (3rd Century BCE). It seems to be the case in our study that many cultures across the world adopt a similar information structure in personal names where the prefix-name has less entropy than the byname, but our study did not explain the prevalence of such information structure. One possibility is that it facilitates the usage of personal names. Since lower frequency items are found to be more difficult to retrieve than higher frequency ones [163], and entropy can be seen as correlated with average frequency, we can infer that the higher entropy of the names, the more difficult it is on average to retrieve a specific name from the distribution. As prefix-name is the first item to be retrieved, keeping its entropy relatively low facilitates its retrieval. Once the prefix-name is retrieved, some of the uncertainty about the byname has already been resolved, so the byname can carry at least as much information in isolation as it does conditioned on the prefix-name. Future studies may test this hypothesis by examining the entropy of prefix-names and bynames in large-scale databases across societies.

In Chapter 4, we presented a case where the reduction of processing effort in legal language is enforced by institutional pressures. Efficient communication pressures language to minimize the processing effort while conveying a meaning. This seems to be the case in many languages [33], but a rare exception is American legalese [29,45]. In this chapter, we first offered an explanation on why lawyers write it this way. We hypothesize that laws are written in a complex way as a result of lawmakers trying to express complex legal concepts transparently into complex sentences. While it may be a transparent way to write laws, it inevitably increases the comprehension difficulty by introducing long syntactic dependencies. Our results on classical Chinese legalese seem to be consistent with the hypothesis. Then, we examined laws in modern China, which went through a drastic change in legal traditions from imperial China. Meanwhile, the government in mainland China started a campaign to make Chinese accessible to laypeople. Such campaign has likely simplified modern Chinese laws, as our results suggest the difference in syntactic dependency length between legalese and non-legalese in modern Chinese is much smaller than previously found in American English, possibly because lawmakers have adopted alternative drafting strategies. Future studies could behaviorally compare the comprehensibility of Chinese legalese texts between laws including everything in one sentence and a simplified version where they are written into multiple sentences. Moreover, in this study, instead of simply the pressure from communication, regulators have to step in to reduce the processing cost of legalese texts. Why is this the case? This may reflect a tension between two competing pressures: ease of production and ease of comprehension in language [263,264]. Speakers may generate utterances that are easy for them to produce but result in comprehension difficulty for the listener [265]. Usually, the two pressures jointly influence human languages, but in the case of legalese, the pressure of production seems to prevail over the pressure of comprehension. Future studies could investigate why this seems to be the case.

Taken together, these studies indicate that the languages can sometimes still achieve efficient communication despite the presence of other pressures. This seems to be the case for spatial demonstratives and legal language. For spatial demonstratives, communication pressures spatial demonstratives towards an efficient frontier, and factors such as environment selects which system on the optimal frontier a language adopts. For legal language, communication and institutional pressure can jointly push legal language to be easy to understand. However, this is not the case for personal names, where the institution pressure pushes personal name systems in Western societies away from an efficient information structure. Why does the difference arise? One possibility is that the communicative pressure prevails only when other factors do not directly act on the variables optimized for communication. For example, in spatial demonstratives, communication first selects the systems that are on the optimal frontier, and then among these systems, different languages settle on an efficient system that best satisfies their communicative needs. In legalese, there are ways to express a meaning and achieve the same legal objectives without long dependencies [224]. The communicative pressure and the institutional pressure have the potential to jointly select the structures that are easiest to understand for readers. On the other hand, in personal name systems, the variables are already highly constrained by communicative pressure in order to maintain an efficient system capable of generating a unique name label for a large number of people, in the form of a composite name system. However, name regulations in Western societies directly capped the amount of information that can be carried in bynames by making them hereditary. This forced prefix-names to carry more information and has potentially made the resulting name systems less efficient to use.

Our studies suggest that the bottom-up mechanism towards efficient communication could be vulnerable to top-down institutional pressures, which is best showcased by our study on personal names, where regulations could alter the historical information structure. This is possibly because regulations have an explicit enforcement mechanism: governments explicitly state how a policy will be implemented, as well as potential consequences of non-compliance. When the objective of the institutional pressure is not aligned with the objective of the bottom-up mechanism for efficient communication, the latter may need to give in to comply with relevant regulations. On the other hand, if applied properly, institutional pressures can also be a positive driver towards communicative efficiency in language, such as making legal language more understandable for laypersons. Our results indicate that regulations can be a double-edged sword on communicative efficiency, illustrating the importance of considering the linguistic consequences when designing regulations that may impact language use. Besides policy makers, other groups may also seek to promote particular ways of speaking, such as the advocacy for gender-neutral occupational terminologies in English. What makes certain proposals more likely to be conventionalized than others? Without an explicit enforcement mechanism like regulations do, it is possible that proposals that do not hinder communicative efficiency are more likely to succeed. Future research should investigate how effective different forms of top-down interventions, from government regulations to non-governmental advocacy, are in shaping language and their impact on communicative efficiency.

Appendix A

Supplementary information for: An information-theoretic approach to the typology of spatial demonstratives

A.1 The decay parameter

Here we test the effect of changing the decay parameter on the conditional distribution of world states given a meaning $p(u|m)$. From Equation (2.5), a low μ indicates $p(u|m)$ decreases very quickly as the number of mismatches between world state u and meaning m increases. Therefore, in this situation, the cost for a person to confuse different distal levels and orientation will be very high, since such a cost is proportional to $-\ln \mu$. On the other hand, a high μ suggests a relatively low cost for confusing distal levels and orientation. In this analysis, we kept other parameters constant, let $\mu \in [0.05, 0.99]$, and compute the average gNID among attested spatial deictic systems.

The results in Figure A.1 show that as long as $\mu < 0.75$, the decay parameter has a minor effect on the optimality of attested spatial deictic systems. In other words, unless we assign a very low cost to confusing between distal levels and orientatons, attested spatial demonstrative systems tend to stay close to the optimal frontier.

A.2 Alternative complexity measures

In this study, similar to [21], we defined complexity as the mutual information between word W and meaning M (which was reduced to $\mathrm{H}[W]$ due to determinism). Meanwhile, a reviewer pointed out that our formulations of complexity and consistency are closely related, and they suggested two alternative ways to operationalize complexity: first, to use the log number of words $\log |W|$ instead of the entropy $\mathrm{H}[W]$; and second, to combine complexity and consistency into one single metric, namely, the minimal description length (MDL) of a spatial deictic demonstrative paradigm. Here we discuss these two alternative metrics and their implementation.

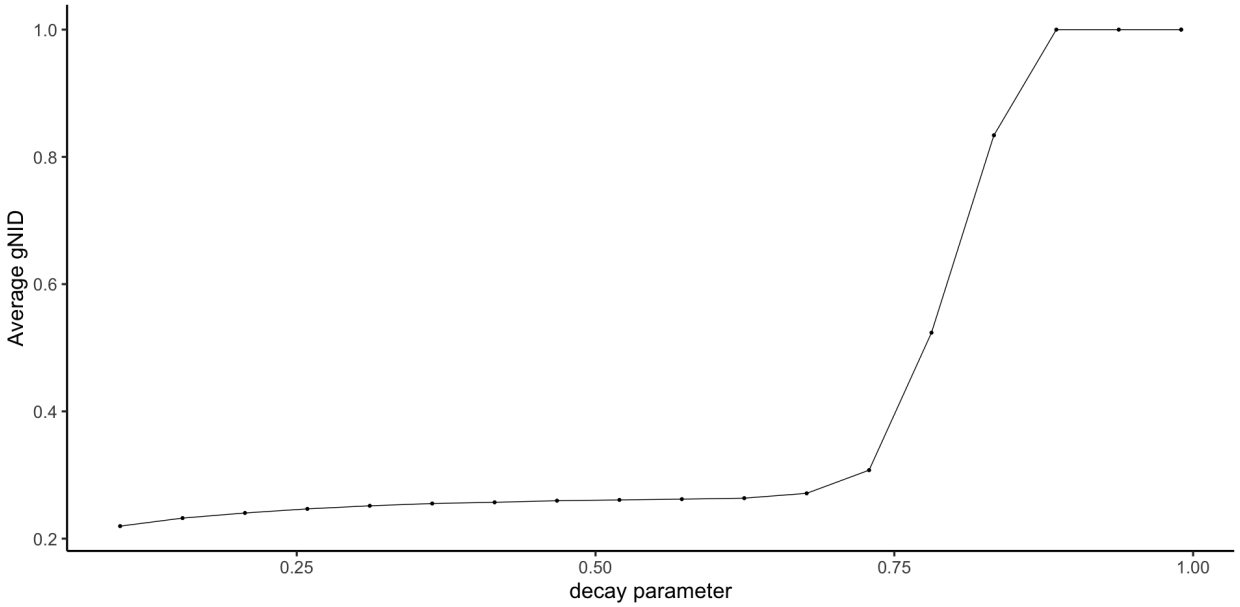


Figure A.1: The average gNID for different values of μ.

Replacing H[W] with number of words One reviewer was suggesting using the number of words as the complexity metric, in order to get around the issue that complexity and consistency are closely related. For instance, if the number of words is 9, the complexity is maximized, and in this case, the consistency in our definition can only take the value of 2. We don't see much improvement other than making both complexity and consistency count-based, since the number of words and consistency are also closely related (See Figure A.2).

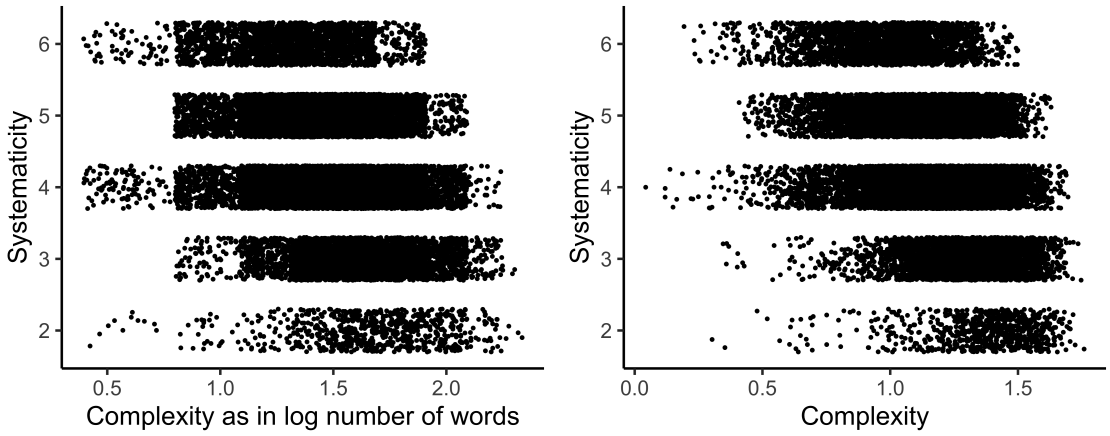


Figure A.2: Plotting consistency against different formulations of complexity. Left: consistency vs. complexity defined as the log number of demonstratives in a system; Right: consistency vs. complexity defined as the mutual information between words and meanings, as in the main text. It can be seen that both metrics are related to consistency, which is also supported by statistics ($R^2 \approx 0.18$ in both cases)

Incorporating complexity and consistency into one single metric by taking an MDL approach Simple operationalizations of MDL are not sufficient to create a unified measure of complexity and consistency. We adopted a similar approach to [113]: counting the minimal number of distal levels and orientations needed to describe a demonstrative, along with logical operations such as AND and OR. For instance, consider the spatial demonstrative system for English (Table A.1a):

Table A.1: English (a) and Fake English (b) spatial demonstratives

	goal	**place**	**source**
D3	there	there	from there
D2	there	there	from there
D1	here	here	from here

	goal	**place**	**source**
D3	there	there	from there
D2	here	here	from there
D1	here	here	from here

For each demonstrative, let's consider the minimal number of distal levels and orientations needed to fully describe it:

- *here*: $(G \cup P) \cup D_1 \rightarrow 3$ features (This means the full definition of the demonstrative "here" is a one that describes GOAL and PLACE at distal level D_1)
- *there*: $(G \cup P) \cup (D_2 \cup D_3) \rightarrow 4$ features
- *from here*: $S \cup D_1 \rightarrow 2$ features
- *from there*: $S \cup (D_2 \cup D_3) \rightarrow 3$ features

Therefore, when we sum them up, under this formulation, English has a complexity of 12. Similarly, Finnish has a complexity of 18 since every word will have 2 features.

However, let's consider the case of Fake English (Table A.1b), where we simply replace the demonstratives at (D_2, PLACE) and (D_2, GOAL) from "there" to "here". Fake English is an example of inconsistent system, as it does not show syncretism of distal levels at different orientations. It is also not attested in the database in [49]. However, let's consider the MDL formulation:

- *here*: $(G \cup P) \cup (D_1 \cup D_2) \rightarrow 4$ features (This means the full definition of the demonstrative "here" is a one that describes GOAL and PLACE at distal level D_1)
- *there*: $(G \cup P) \cup D_3 \rightarrow 3$ features
- *from here*: $S \cup D_1 \rightarrow 2$ features
- *from there*: $S \cup (D_2 \cup D_3) \rightarrow 3$ features

The MDL complexity is still 12. In other words, although Fake English is clearly less consistent than English, their MDL complexity is the same. Hence, MDL is probably not a metric that can incorporate both complexity and consistency. As a result, in the main text, we kept our current metrics for complexity and consistency.

Although the MDL theory above does not capture systematicity in our sense, more sophisticated description languages may do so. Another challenge for linking MDL with the IB sense of complexity is the fact that mutual information as complexity metric depends on the quantitative real-valued probabilities of words and meanings, while MDL approaches typically model discrete phenomena.

A.3 Alternative need distribution sources

In the main text (See Section 2.5), we approximated the need distribution of different meanings $p(m)$ by the word frequency distribution of Finnish spatial deictic demonstratives, since in Finnish, each meaning has its own, unique spatial demonstrative. We drew the Finnish word frequency data from Lexiteria, which contains the word frequency data on the internet between 2009 and 2011. However, a disadvantage of Lexiteria is that this database is not open-source. Therefore, here we present analysis from two other databases: WorldLex [266], a database of Twitter and blog word frequencies, and OpenSubtitles [267], a database of movie and TV subtitles. For each spatial demonstrative in the Worldlex corpus, we add its Twitter frequency and blog frequency together.

	Lexiteria	**WorldLex**	**Opensubtitles**
D1, place, tänne	232,946	10,913	297,313
D1, goal, täällä	94,887	3,931	121,392
D1, source, täältä	38,402	1,926	45,204
D2, place, sinne	109,576	11,724	139,150
D2, goal, siellä	42,923	6,431	54,970
D2, source, sieltä	10,006	4,233	111,80
D3, place, tuonne	43,016	2,245	49,141
D3, goal, tuolla	17,587	448	21,193
D3, source, tuolta	3,850	480	4,928
Average gNID for attested systems	**0.239**	**0.271**	**0.237**

Table A.2: Finnish spatial deictic demonstrative frequency from three different corpora: Lexiteria, WorldLex [266], and OpenSubtitles [267], as a proxy for the need distribution $p(m)$, and the average gNID among real spatial demonstrative systems under each frequency distribution.

The results are shown in Table A.2: the last row shows the average generalized normalized information distance (gNID, [21]) among real deictic systems, under each corpus. As also explained in Section 2.5, the lower the average gNID is, the closer the real deictic systems are to the optimal frontier. The average gNID from Lexiteria and OpenSubtitles are very close to each other (0.239 vs. 0.237, respectively), while they are relatively far from the gNID calculated from WorldLex (0.271), probably because the frequency distribution in WorldLex does not decay with respect to distal level, like those in Lexiteria and Opensubtitles. However,

the discrepency is still relatively small compared to the variation in average gNID under different PLACE/GOAL/SOURCE costs (See Section 2.5).

A.4 Results by language family

The dataset used in this study [49] sampled approximately 50 languages from each of the 5 geographic regions around the globe. However, they did not control for language relatedness, as many languages from the same region are closely related. For example, 15 out of the 50 languages in Europe are Slavic. As a result, our finding in Figure 2.5 did not rule out a possibility that the efficiency of spatial deictic systems are mainly pushed by some large language families. For example, it could be possible that the spatial deictic systems in Indo-European languages are efficient due to chance, and since Indo-European languages constitute a large portion of the database, they can easily inflate the results shown in Figure 2.5.

To address this potential confound of language relatedness, in Figure A.3, we present the information plane in the same way as Figure 2.5, but instead of plotting all 220 languages that we investigate, we randomly sample 1 language per language family. All attested spatial deictic systems shown in the plot as colored points do not share a common ancestor with each other. Since a large portion of them are still located very close to the optimal frontier, we can say that language relatedness is an unlikely confound in our study.

A.5 Results by merging distal levels D1 and D2

In Section 2.4.1, we treat every attested language as having three distal levels. However, there exist languages (e.g. English) that only distinguish two distal levels. Hence, in that section, we adopt a convention that if the spatial deictic system in a language only distinguishes two distal levels, we assume the second distal level extends out, encompassing both D2 and D3. As one anonymous reviewer points out, there is an arbitrary decision made by us, with no theoretical motivation. In this analysis, we repeat Experiment 1 by assuming if a language has only two distal levels, the first distal level includes both D1 and D2, instead of just D1. The results are shown in Figure A.4, suggesting that there are no qualitative differences in our results that depend on our choice in distal level merging.

A.6 Effects of need distribution permutation and place/-goal/source cost on optimality

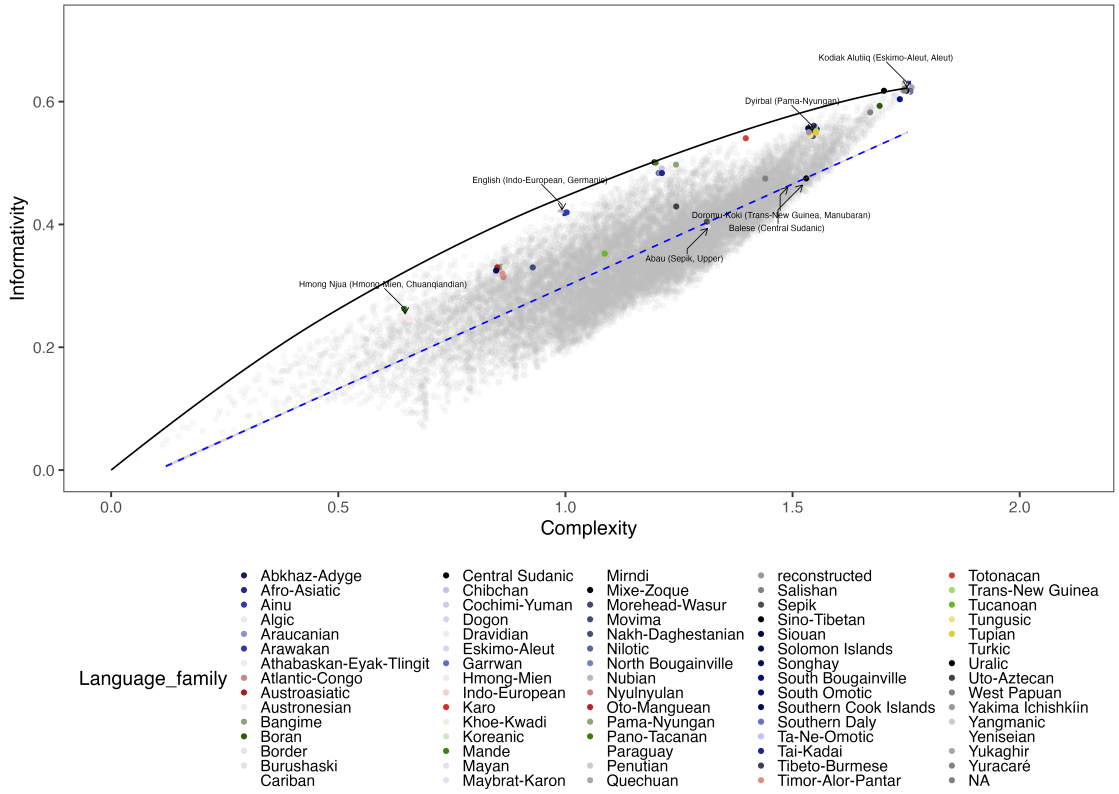


Figure A.3: Similar plot as Figure 2.5, except we only sampled and presented 1 language per language family.

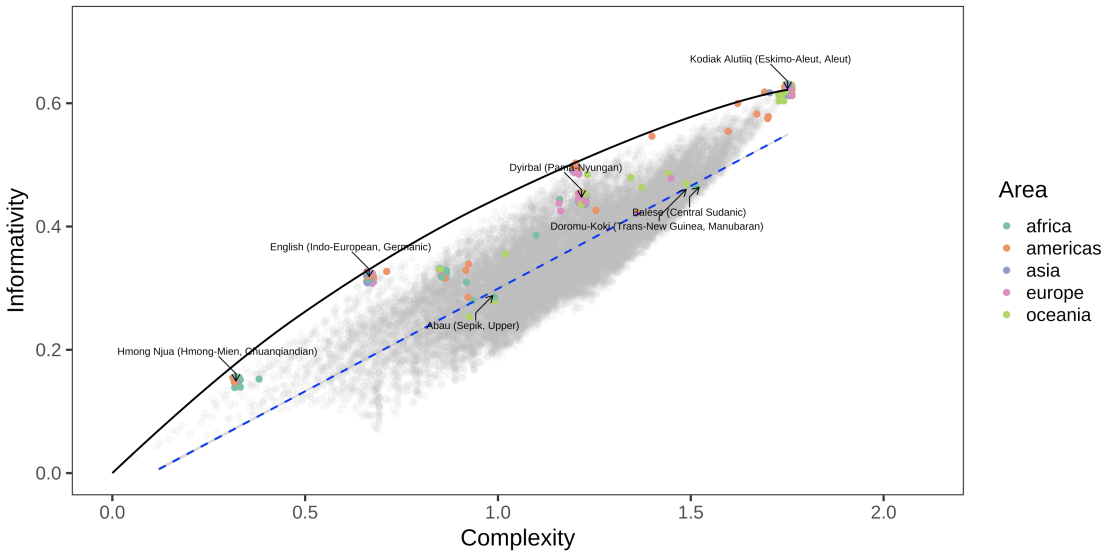


Figure A.4: Similar plot as Figure 2.5, except we assume the first distal level encompasses D1 and D2, instead of just D1.

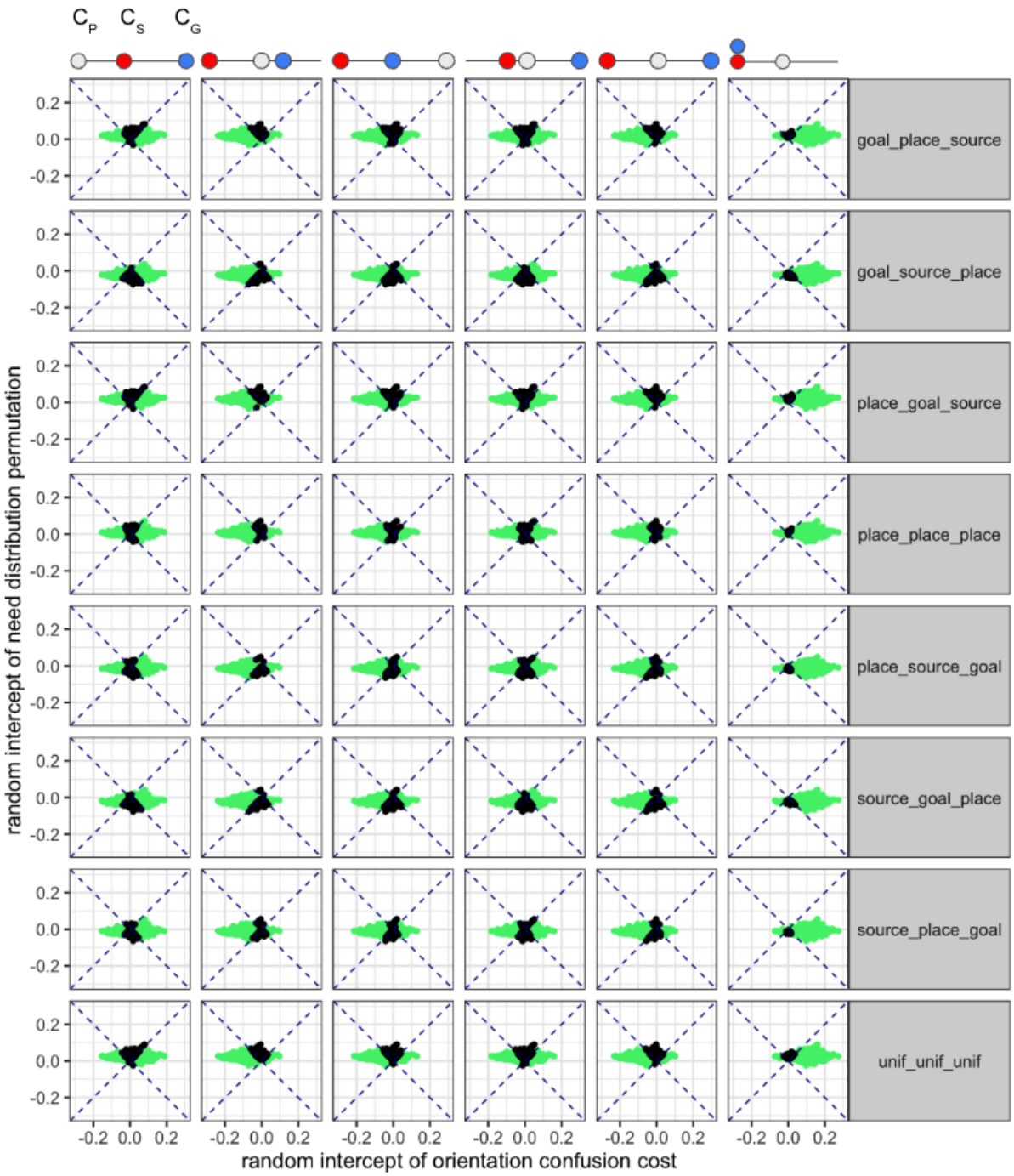


Figure A.5: The fitted random intercept per sample for PLACE/GOAL/SOURCE confusion costs (x-axis) and for need distribution permutations (y-axis), under each combination of these two factors (shown in each individual facet), in the regression of Eq. 2.10. Color code: green - the random intercept for the confusion cost has a higher absolute value than that for the need distribution permutation; black - the random intercept for the confusion cost has a lower absolute value than that for the need distribution permutation. Most of the samples (154403 out of 192000) are colored green.

Appendix B

Supplementary information for: Cross-Cultural Structures of Personal Name Systems Reflect General Communicative Principles

B.1 Bootstrapping the correlations analysis in Finnish data

In Experiment 2, we calculated the Finnish prefix-name entropy of each birth year bin and each parish from a sample of 50 births, as the number of births in each parish varied. We found a main effect of birth year on prefix-name entropy, a main effect of longitude on prefix-name entropy, and an interaction between birth year and longitude on prefix-name entropy. Our results in Table 3 showed positive correlations between prefix-name entropy and the percentage of people having patronyms, as well as a positive correlation between prefix-name entropy and longitude, and such positive correlations decreased over time.

An anonymous reviewer raised an issue on the reliability of such a subsampling approach in calculating entropy. To address the reliability issue, we repeated the aforementioned analyses 500 times with different randomly sampled names and visualized the related statistical quantities as histograms in Figure B.1. First, the effects of birth year and longitude, as well as their interaction, remained robust, since among the 500 analyses, all of them showed a significant effect trending in the same direction (Figure B.1a) as was shown in Experiment 2. Similarly, the correlations among patronym percentage, longitude, and prefix-name entropy remained robust (Figure B.1b), since among the 500 analyses, all of them showed similar trend as was shown in Experiment 2. We also repeated the analysis by sampling 100 names each time and obtained similar results (Figure B.2).

Therefore, the results we showed in Experiment 2 were unlikely to be caused by artifacts introduced when we subsampled our data for entropy calculation.

B.2 Licensing information of datasets used in this study

In this study, we mainly used data from publicly available sources. Below is a compilation of permissions from our data sources. Our use of datasets fully complies with the terms and conditions set by each data provider.

Social Security Administration The Social Security Administration allows data to be used by "researchers interested in naming trends".[1]

U.S. Census Bureau The U.S. Census Bureau explicitly states that their datasets are made available for use in research.[2]

Wikipedia Data from Wikipedia is available for re-use under a CC-BY-SA license, [3] which allows its data to be shared and adapted as long as appropriate credits are given.

Taiwanese Ministry of Interior The Taiwanese goverment allows their published data to be freely used for noncommercial purposes.[4]

Korean Statistical Information Service One function of the Korean National Statistic Office is to make statistical data available for researchers.[5]

Population Data UK Population Data UK is a site "dedicated to providing information about the population of the United Kingdom".[6]

Data from Douglas A. Galbi One purpose of such dataset, according to Galbi, is "to spur further analysis of given names".[7]

National Records of Scotland All contents in the National Records of Scotland operates are available under the Open Government License v3.0, which allows its data to be copied, published, distributed, and transmitted as long as the source is acknowledged.[8]

HisKi database Permission to use the HisKi database has been given to several researchers. Several papers have been published using information in this database, such as [184,268].

[1] https://www.ssa.gov/oact/babynames/limits.html
[2] https://www.census.gov/data/developers/about/terms-of-service.html
[3] https://en.wikipedia.org/wiki/Wikipedia:Text_of_the_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License
[4] https://data.gov.tw/license
[5] https://kostat.go.kr/menu.es?mid=a20602030000. See the Dissemination of Statistical Information paragraph.
[6] populationdata.org.uk
[7] https://www.galbithink.org/names/
[8] https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

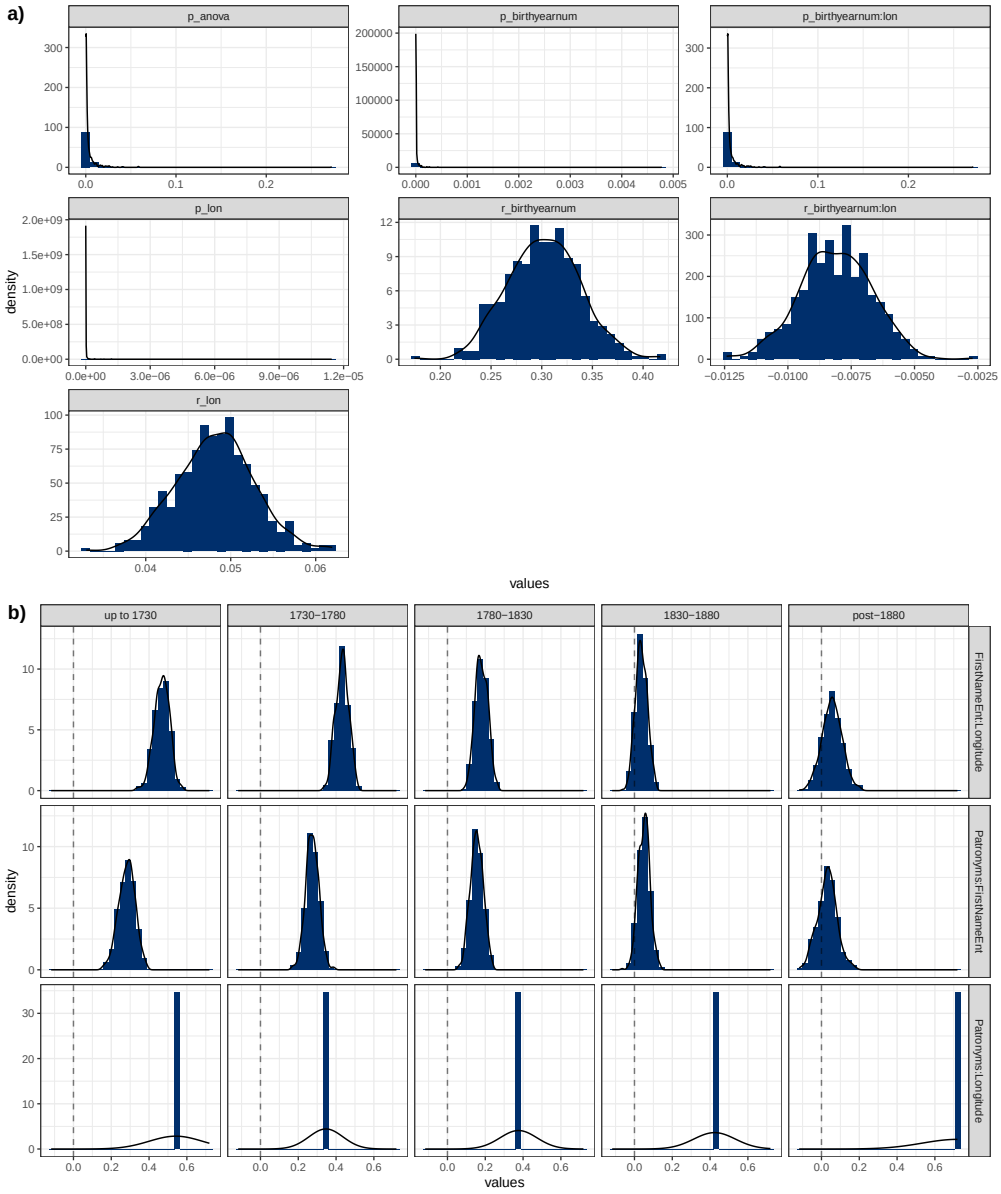


Figure B.1: **Results in Experiment 2 were independent of sampling**. Results from the bootstrapping analysis, where we sampled 50 births in each Finnish parish in each birth year bin, conducted the analyses in Experiment 2, and repeated the whole process for 500 times. **(a)** The distribution of two-sided p_values of model comparison (an interaction model vs. an additive model), two-sided p_values and the main effect of birth year on prefix-name entropy, two-sided p_values and the interaction between birth year and longitude, along with two-sided p_values and the main effect of longitude on prefix-name entropy. **(b)** The distribution of correlations between prefix-name entropy and longitude (first row), percentage patronym and prefix-name entropy (second row), and the percentage patrobym and longitude (third row), in each birth year bin (columns). The black curve in each graph represents the kernel density generated by `geom_smooth` function in R [145]

.

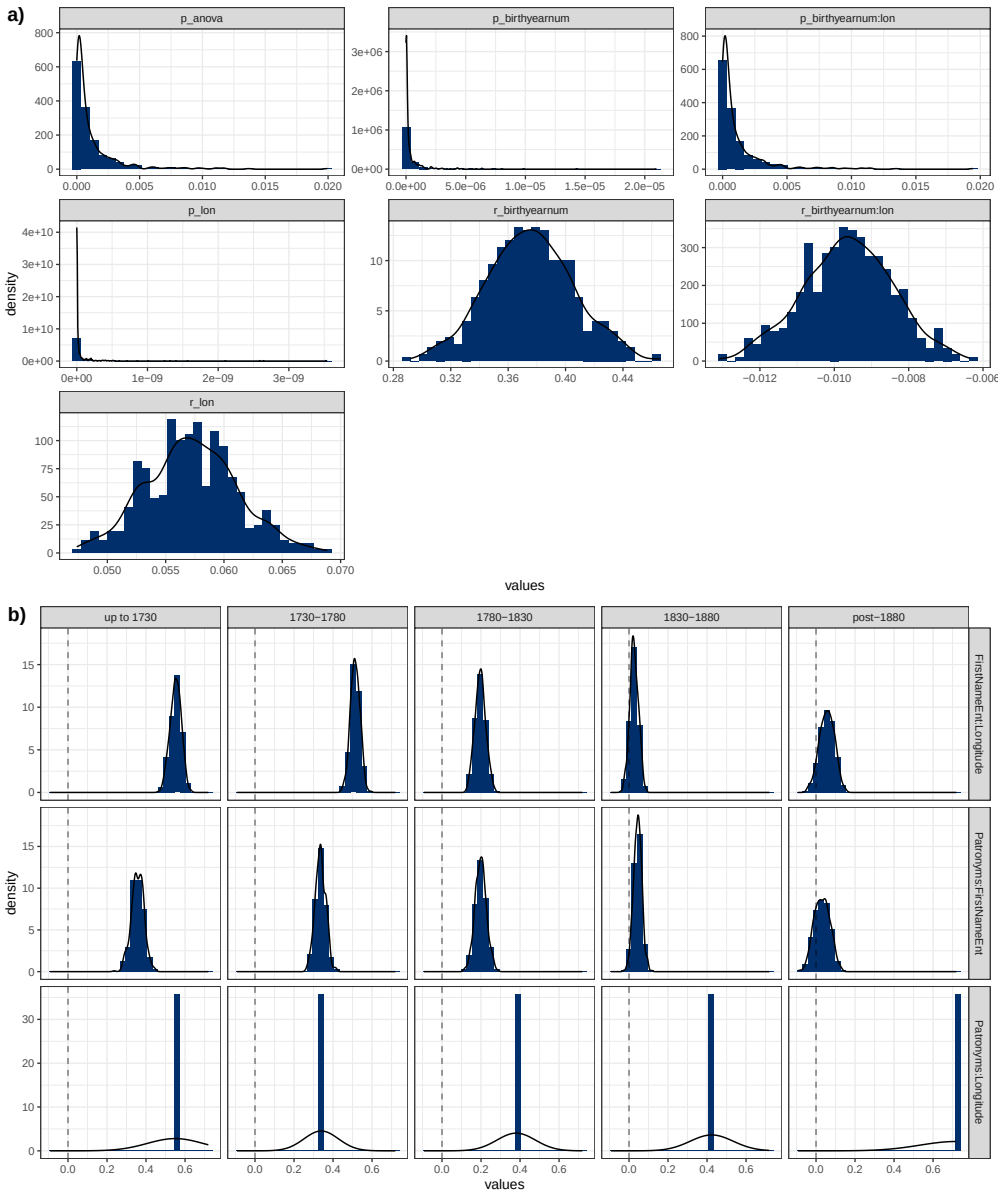


Figure B.2: **Results in Experiment 2 were independent of sampling**. Results from the bootstrapping analysis in which we sampled 100 births instead of 50 (as in Figure B.1) in each Finnish parish in each birth year bin. The black curve in each graph represents the kernel density generated by `geom_smooth` function in R [145]

.

Appendix C

Supplementary information for: The conceptual structure of laws leads to legalese

C.1 Supplementary Tables

C.2 Annotation and parsing conventions for classical Chinese texts

C.2.1 Definitions

Clause A clause is a string of characters ending in comma, period, question mark, or exclamation mark. For example, Both "Yesterday morning" and "I was at the gym" in "Yesterday morning, I was at the gym." are clauses. As another example, the sentence "When I was young, I used to go to the gym a lot" contains two clauses, "when I was young" and "I used to go to the gym a lot".

Independent clause An independent clause is a clause that can stand alone as a sentence (i.e. with a root). Examples of independent clauses include "Run!", "He is a singer", and "I was young.". Examples of **dependent clauses** include "Although he is a singer", "'Yesterday morning", and "When I was young".

Sentence A sentence is a collection of one or more clauses that end in period, question mark, or exclamation mark. A sentence in Chinese can contain multiple independent clauses, unlike in English. Sentences that contain multiple independent clauses are referred to as **run-on** sentences [269].

C.2.2 Indexing conventions for classical Chinese texts

All texts in the Tang Code are uniquely identifiable via an ID, such as 9:101:1:1 for the sentence in 4.3 in the main text. The first digit is the chapter number. The second digit is

Table C.1: The genre, article name, and author of the 40 non-legalese texts in this study.

Genre	Article	Author
Short Story	捕蛇者説 (Tale of a snake catcher)	柳宗元 (Liu Zongyuan, 773-819 CE)
Commentary	原道 (Essentials of the moral way)	韓愈 (Han Yu, 768-824 CE)
Diary	廬山草堂記 (Record in the thatched cottage in Lu Shan)	白居易 (Bai Jüyi, 772-846 CE)
Fiction	柳氏傳 (Biography of Ms. Liu)	許堯佐 (Xu Yaozuo, 9th century)
Commentary	讀司馬法 (Reading the Sima Law)	皮日休 (Pi Rixiu, 834-883 CE)
Commentary	南柯太守傳 (Biography of the Nanke Governor)	李公佐 (Li Gongzuo, 778-848 CE)
Commentary	諫太宗十思疏 (Commentary with 10 thoughts for the Emperor)	魏徵 (Wei Zheng, 580-643 CE)
Fiction	鶯鶯傳 (Biography of Yingying)	元稹 (Yuan Zhen, 779-831 CE)
Letter	與韓荊州書 (Letter to Han Jinzhou)	李白 (Li Bai, 701-762 CE)
Diary	機汲記 (An account of collecting water with machines)	劉禹錫 (Liu Yuxi, 772-842 CE)
Commentary	書褒城驛壁 (Writing on the wall of the Bao City post station)	孫樵 (Sun Qiao, mid 9th century)
Biography	李賀小傳 (A mini biography of Li He)	李商隱 (Li Shangyin, 813-858 CE)
Commentary	復仇議 (Comments on revenge)	陳子昂 (Chen Zi'ang, 661-702 CE)
Commentary	野廟碑 (Scriptures on a random temple)	陸龜蒙 (Lu Guimeng, ?-881 CE)
Fiction	柳毅傳 (Biography of Liu Yi)	李朝威 (Li Zhaowei, 766-820 CE)
Commentary	敘事 (Storytelling)	劉知幾 (Liu Zhiji, 661-721 CE)
Diary	養貍述 (An account on feeding a wildcat)	舒元輿 (Shu Yuanyu, 791-835 CE)
Commentary	杭州新造南亭子記 (Record on the newly built Southern Pavillion in Hangzhou)	杜牧 (Du Mu, 803-852 CE)
Commentary	兩漢辨亡論 (Commentary on causes of end of the two Han dynasties)	權德輿 (Quan Deyu, 759-818 CE)
Commentary	上韓昌黎書 (Appeal to Han Changli)	張籍 (Zhang Ji, 766-830 CE)

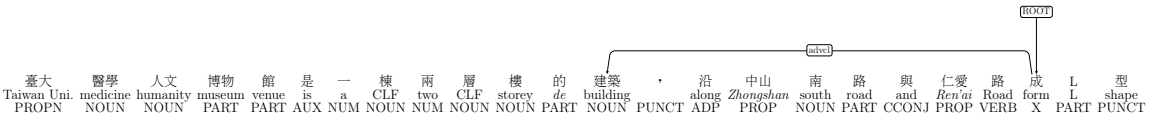
the article number. The third digit is the paragraph number within an article, as displayed in Wikisource [270]. The fourth digit is the sentence number within a paragraph. A sentence is defined as a sequence of characters delimited by periods or paragraph breaks. Commentaries follow the template of <chapter number>:<article number>:<paragraph number>:<sentence number>c<commentary sentence number>, such as "25:368:1:1c1". Occasionally subcommentaries were included as example sentences in the Supplementary Information. They are indexed with a prefix "s", such as "29:469:1:s5".

All items in the control texts are also uniquely identifiable via their ID, such as liang-han-bian-wang for the sentence in [4] in the main text. The text string is the pinyin of the first four characters of the title. The first digit is the paragraph number. The second digit is the sentence number within a paragraph. If there are multiple sentences enclose by a pair of quotations, all the texts enclosed share the same ID and are further differentiated by suffixes "a", "b", "c", and so on.

C.2.3 How are clauses connected together?

The original Tang Code texts are not punctuated, as punctuations were only introduced in Chinese in the 20th Century. The punctuations in Tang Code used in our analysis were added to the original text afterwards, possibly following the usage of modern Chinese punctuations. However, unlike English, in Chinese, independent clauses are commonly connected with commas in a sentence without any conjunctions [269,271–273]. An example of run-on sentence is shown below in (C.1): both [A] and [B] are independent clauses, and they are connected with a comma.

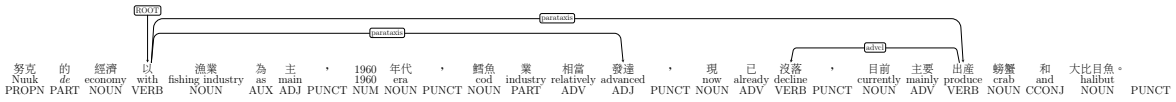
ID: zhgs/test4
Sentence: [A]臺大醫學人文博物館是一棟兩層樓的建築，[B]沿中山南路與仁愛路成L型。
Translation: [1]The Museum of Medical Humanities of National Taiwan University is a two-storey building, [2]it forms an L-shape along Zhongshan South Road and Ren'ai road.



[C.1]

In Universal Dependencies, the practice seems to be that each sentence can only have one root. In the case of run-on sentences, the roots of each clauses are usually connected together via **parataxis** or **advcl**, even if such relations are not licensed by conjunctions. For example, in [C.1], [A] is connected with the [B] via **advcl**. As another example, in [C.2], there are four independent clauses. [A] and [B] are connected via **parataxis**, as are [A] and [D], and [C] and [D] are connected via **advcl**. None of these connections are syntactically marked. In fact, connections between clauses are rarely marked in Chinese, as the relationship between clauses is either clear via their relative placement or can be inferred in context [255].

ID: zhgsd/dev7
Sentence: [A]努克的經濟以漁業為主，[B]1960年代，鱈魚業相當發達，[C]現已沒落，[D]目前主要出產螃蟹和大比目魚。
Translation: [A]The economy of Nuuk is mainly fishing industry, [B]in the 1960s, the cod industry was highly developed, [C]it has since declined, [D]currently it mainly produces crabs and halibuts.



[C.2]

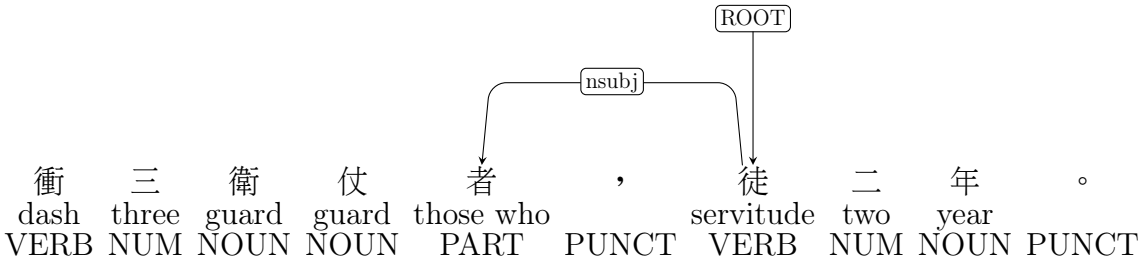
In this study, we had a different practice: we did not connect independent clauses at all. Instead, we treated each of them as having their own root. This is because in this study, we aimed to analyze texts from a strictly syntactic perspective: we calculated the average dependency length of different genres of texts, as they were presented. We were not trying to presume any syntactic connections that were not explicitly marked in the texts. In our view, any presumed connection between independent clauses is a connection at the discourse level, instead of at the syntactic level. For example, in [C.1], it's more appropriate to classify the connection between [A] and [B] as discourse instead of syntactic, as [B] serves as an elaboration [274] of [A].

In addition, as mentioned before, punctuations were not introduced into Chinese until approximately 100 years ago, and as a result, punctuations for all the classical Chinese texts were added only recently. However, there are no clear guidelines on when to use a period instead of a comma, and the practice varies across native speakers [275]. As a result, for the same classical Chinese text, different annotators, adopting their own practices, might segment the text into different number of sentences. If we assume each sentence only has one root, the resulting average dependency length will vary, dependent on the annotator.

Hence, in our analysis, within each sentence, we first identified the number of clauses. Then, we checked whether if each clause is an independent clause, and if not, how to connect the clause to an independent clause.

If a dependent clause serves as the subject of an independent clause, the two clauses are connected together. For example, in [C.3], the comma separates the first clause serving as the subject, and the second clause serving as the predicate.

ID: 7:74:1:1
Sentence: 衝三衛仗者，徒二年。
Translation [276]: Those who interfere with the three ranks of the imperial guardsmen are punished by two years of penal servitude.



[C.3]

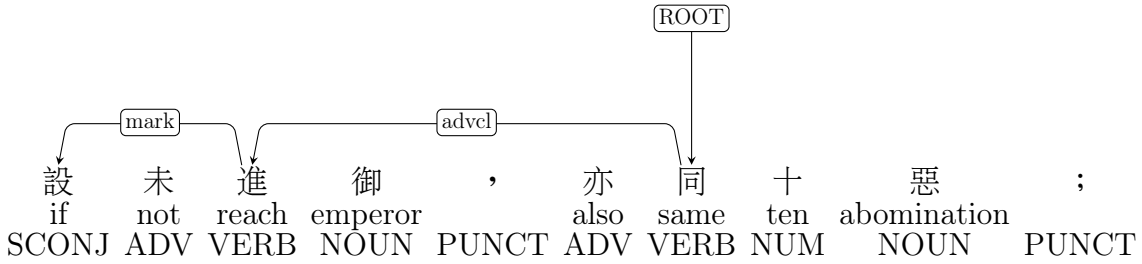
If a dependent clause serves as oblique of an independent clause, the two clauses are connected together. For example, in [C.4], the second clause is a complete sentence with a subject ("罪 / crime"), a verb ("如 / resemble"), and "之 / it". The first clause provides additional information about the crime, specifically on which occasions the crime is the same, and therefore, it connects to the second clause via the relation **obl**.

ID: 10:125:1:2
Sentence: 若依式應須遣使詣闕而不遣者，罪亦如之。
Translation [276]: In cases where, according to the statuses, documents should be sent to the palace by courier and they are not, the punishment is the same.

[C.4]

If a dependent clause is marked by a coordinate or subordinate conjunction, it connects to the independent clause as **advcl** or **conj**. A word is a conjunction if it is labeled as such (連詞) in the Classical Chinese Dictionary [277]. In [C.5], the connection (**advcl**) is licensed by a subordinate conjunction marker "設".

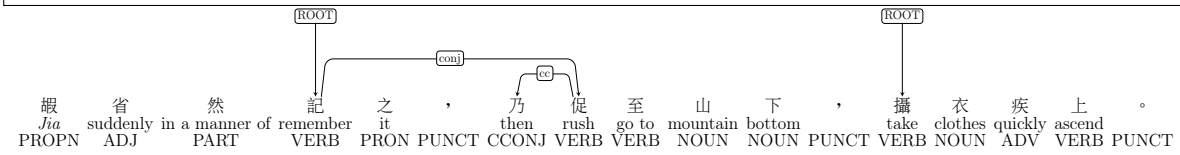
ID: 1:6:7:1c5:s3
Sentence: 設未進御，亦同十惡；
Translation [244]: Even if these items have not yet been presented to the emperor, the crime still comes under the ten abominations.



[C.5]

As another example, in [C.6], (B) is connected to (A) because the connection is licensed by a conjunction "乃". On the other hand, (C) is not connected to other clauses because there are no conjunctions licensing such connection.

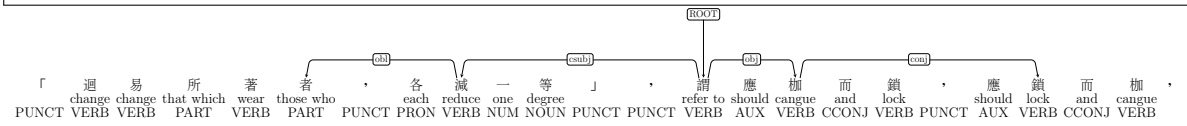
ID: liu-yi-zhuan:11:6
Sentence: (A)嘏省然記之，(B)乃促至山下，(C)攝衣疾上。
Translation: (A) *Jia* suddenly realized and realized it, (B) and then he rushed to the bottom of the mountain. (C) He grabbed his clothes and quickly ascended.



[C.6]

However, there are limited cases where seemingly independent clauses can be connected. Two clauses can be connected if they are both part of the subject, both part of the object, or both clausal modifiers of the same word, even if the connection between the two clauses is not explicitly enabled. In [C.7], the clause "迴易所著者" is connected to "各減一等" because the first clause serves as the subject, and the second clause serves as the predicate. Then these two clauses are connected to the rest because they serve as the clausal subject (**csubj**), and the rest serves as the predicate. In the predicate, the clauses "應枷而鎖" and "應鎖而枷" are connected by the relation **conj**, because both are the object of the verb "謂". Therefore, the entire set of 4 clauses are considered to share the same root.

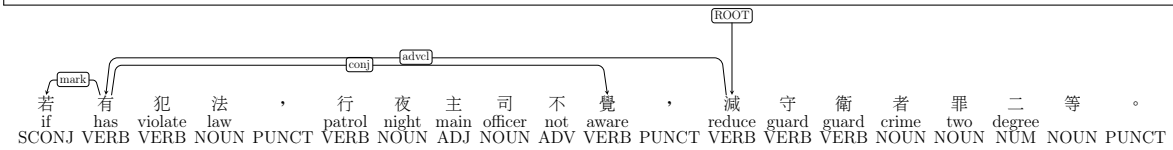
ID: 29:469:1:s5
Sentence: 「迴易所著者，各減一等」，謂應枷而鎖，應鎖而枷，
Translation [276]: "If the means of restraint are changed, the punishment is reduced one degree in each case" means that those who should be in the cangue are put into fetters or those who should be in fetters are put into the cangue.



[C.7]

As another example, in [C.8], there are three clauses: (A)"若有犯法 (if there is a violation of law)", (B)"行夜主司不覺 (the officials in charge of the night patrols are not aware)", and (C)"減守衛者罪二等 (the punishment of the night guard reduced by two degrees)". (A) contains a subordinate conjunction, which indicates that it is an adverbial clause. Semantically, both (A) and (B) states the condition of an offense, and (C) lays out punishment of the offense, so we interpret (A) and (B) as part of the adverbial clause, and (C) as the main clause. Therefore, (A) is connected to (C) with **advcl**, and (B) is connected to (A) with **conj**, even though (A) and (B) are not explicitly connected.

ID: 8:78:1:1
Sentence: 若有犯法，行夜主司不覺，減守衛者罪二等。
Translation [276]: if there is an offense, and the officials in charge of the night patrols are not aware of it, his punishment is reduced two degrees below that of the guardsmen.



[C.8]

C.2.4 List of universal dependency relations used in our analysis

We use the following dependency relations from Universal Dependencies (UD), [246] in our analysis.

nsubj: the nominal subject of a clause

In our analysis, if the subject is semantically the object of the clause, we still use **nsubj** as the dependency relation. In addition, the subject of sentennces with *bei* constructions is connected to the main verb via **nsubj**, which is simplified from **nsubj:pass** in UD.

Example:

ID: 4:34:4:1
Sentence: 庸、賃雖多，各不得過其本價。
Translation [244]: Punishment may not be fore more than the basic value of a particular thing, even though its use or rent produced more than its basic value.

[C.9]

obj: the object of a clause

Example:

ID: 4:34:4:1
Sentence: 庸、賃雖多，各不得過其本價。
Translation [244]: Punishment may not be fore more than the basic value of a particular thing, even though its use or rent produced more than its basic value.

[C.10]

In our analysis, if the object is a clause, the head of the clause is also connected with the main verb via **obj**. We did not use relations such as **ccomp** and **xcomp** in our analysis (Also See C.2.7).

Example:

ID: 9:101:1:1
Sentence: 諸廟享，知有緦麻以上喪，遣充執事者，笞五十；
Translation [244]: All cases of putting a person in charge of a court celebration while knowing that he is in mourning of a relative within the fifth degree of mourning are punished by fifty blows with the light stick.

[C.11]

csubj: clausal subject

Example:

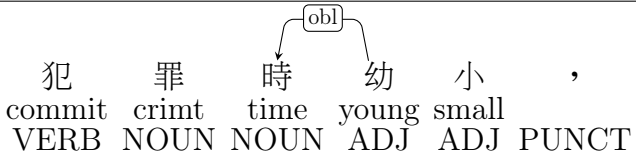
ID: lu-shan-cao-tang:2:3
Sentence: 砌階用石，冪窗用紙。
Translation: Building stairs uses stones. Covering windows uses papers.

[C.12]

obl: oblique

Example:

ID: 4:31:3:1
Sentence: 犯罪時幼小，
Translation [244]: An offense has been committed before a person is aged or infirm.



[C.13]

advcl: adverbial clause

Example:

ID: 4:34:4:1
Sentence: 庸、賃雖多，各不得過其本價。
Translation [244]: Punishment may not be fore more than the basic value of a particular thing, even though its use or rent produced more than its basic value.

[C.14]

acl: clausal modifier of a noun

Example:

ID: 15:199:3:2
Sentence: 監當主司知而聽者，併計所知，同私馱載法。
Translation [276]: If the officer in charge at the time has knowledge of this and yet allows it, everything about which he has knowledge is combined and calculated, and the punishment is the same as for carrying private possessions.

[C.15]

advmod: adverbial modifier

Example:

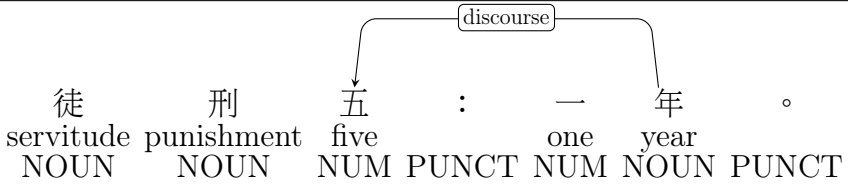
ID: 4:34:4:1
Sentence: 庸、賃雖多，各不得過其本價。
Translation [244]: Punishment may not be fore more than the basic value of a particular thing, even though its use or rent produced more than its basic value.

[C.16]

discourse: discourse element

In our study, we use **discourse** to connect the head of the clause before a colon to the head of the clause after the colon.
Example:

ID: 1:3:1:1
Sentence: 徒刑五：一年。
Translation [244]: The five punishments of penal servitude have a duration of one...

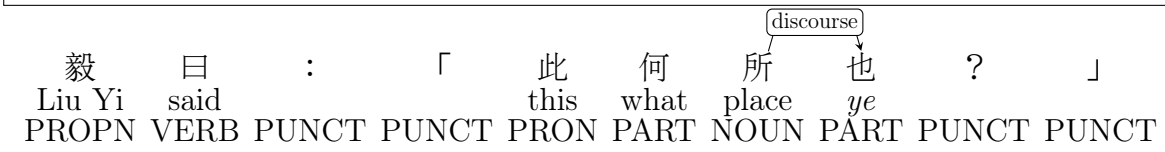


[C.17]

We also use **discourse** to connect a sentence final particle to its head (usually a verb). In UD this dependency relation is labeled as **discourse:sp**, but in our study, we simplified it to **discourse**.

Example:

ID: liu-yi-zhuan:2:12
Sentence: 毅曰：「此何所也？」
Translation: Liu Yi said, "what place is this?"



[C.18]

amod: adjective modifier

Example:

ID: 15:199:3:2
Sentence: 監當主司知而聽者，
Translation [276]: If the officer in charge at the time has knowledge of this and yet allows it....

[C.19]

aux: auxilary

Example:

ID: 4:34:4:1
Sentence: 庸、賃雖多，各不得過其本價。
Translation [244]: Punishment may not be fore more than the basic value of a particular thing, even though its use or rent produced more than its basic value.

[C.20]

The auxiliary word marking passive voice "被" is parsed as **aux:pass** in our analysis.

Example:

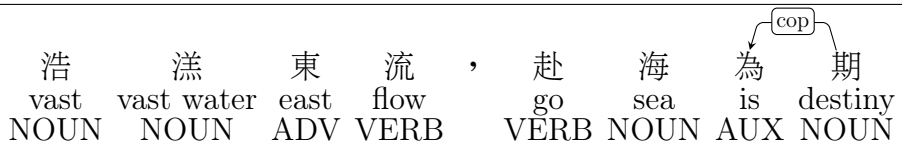
ID: 25:368:1:1c1
Sentence: 對制，謂親見被問。
Translation [276]: Replying to an imperial decree means to have a personal audience with the emperor an be questioned by him.

[C.21]

cop: copula

Example:

ID: ji-ji-ji:3:3
Sentence: 浩滔東流，赴海為期。
Translation: Vast water flows eastward. Going to the sea is its expected destiny.



[C.22]

mark: marker

mark connects a subordinate conjunction with the head of the subordinate clause.
Example:

ID: 4:34:4:1
Sentence: 庸、賃雖多，各不得過其本價。
Translation [244]: Punishment may not be fore more than the basic value of a particular thing, even though its use or rent produced more than its basic value.

[C.23]

det: determiner

Example:

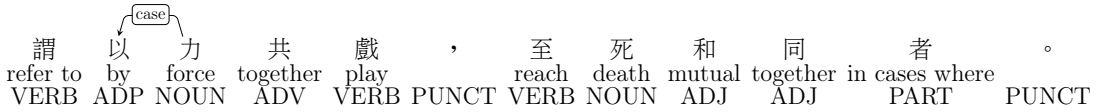
ID: 4:34:4:1
Sentence: 庸、賃雖多，各不得過其本價。
Translation [244]: Punishment may not be fore more than the basic value of a particular thing, even though its use or rent produced more than its basic value.

[C.24]

case: case marking

Example:

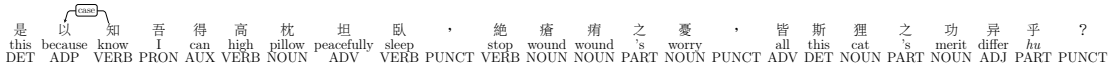
ID: 23:338:1:1c1
Sentence: 謂以力共戲，至死和同者。
Translation [276]: This refers to persons agreeing to a game where strength is used in a friendly way but death results.



[C.25]

Example:

ID: yang-li-shu:4:3
Sentence: 是以知吾得高枕坦臥, 絶瘡痍之憂，皆斯狸之功异乎？
Translation: Because of this I know I can sleep peacefully with a high pillow. I can be free from the worry of wounds. It's all the merit of this cat, isn't it?

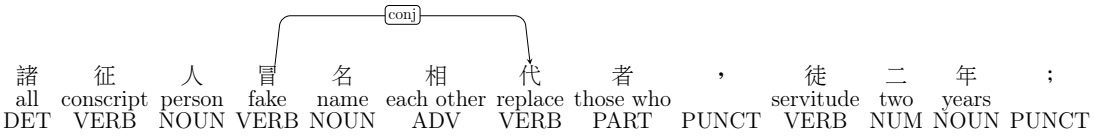


[C.26]

conj: conjunction

Example:

ID: 16:228:1:1
Sentence: 諸征人冒名相代者，徒二年；
Translation [276]: All cases of conscripts replacing each other under false names are punished by two years of penal servitude.

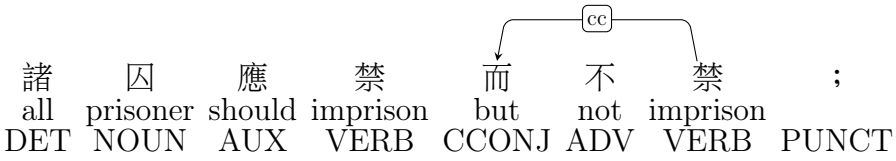


[C.27]

cc: coordinate conjunction

cc marks the coordinate conjunction with the head of the coordinating element.
Example:

ID: 29:469:1:1
Sentence: 諸囚應禁而不禁；
Translation [276]: All cases of not imprisoning persons who should be...

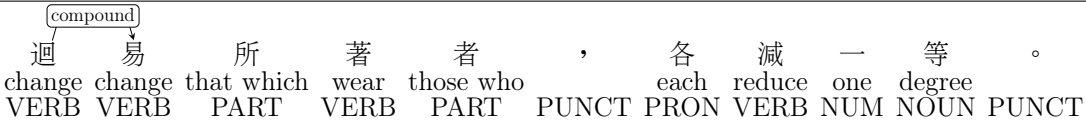


[C.28]

compound: compound words

Example:

ID: 29:469:1:1
Sentence: 迴易所著者，各減一等。
Translation [276]: If the means of restraint are changed, the punishment is reduced one degree in each case.

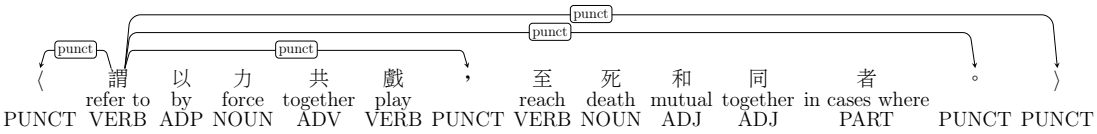


[C.29]

punct: punctuation

Example:

ID: 23:338:1:1c1
Sentence: 謂以力共戲，至死和同者。
Translation [276]: This refers to persons agreeing to a game where strength is used in a friendly way but death results.

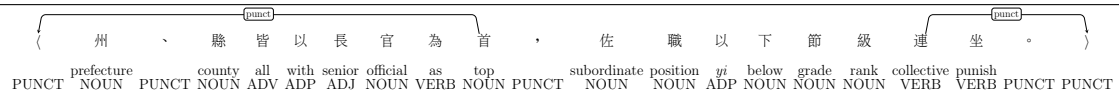


[C.30]

If there are multiple sentences within a quotation or a parenthesis, the opening quotation or parenthesis connects to the head of the first sentence, and the closing quotation or parenthesis connects to the head of the last sentence.

Example:

ID: 13:174:1:1c1
Sentence: 〈州、縣皆以長官為首，佐職以下節級連坐。〉
Translation [276]: At the prefecture and county levels, the senior official is the principal. He, his subordinates, and those who have lower grades of rank are collectively responsible.

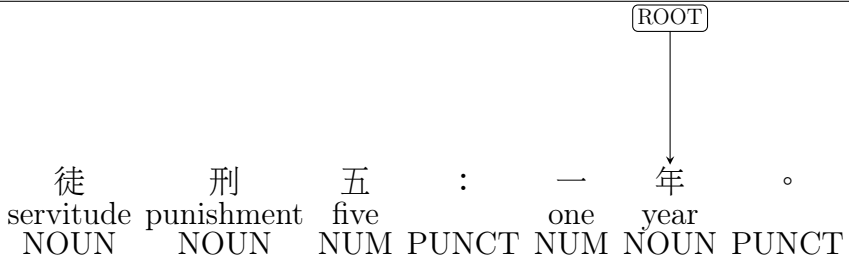


[C.31]

root: the root of a sentence

Example:

ID: 1:3:1:1
Sentence: 徒刑五：一年。
Translation [244]: The five punishments of penal servitude have a duration of one...

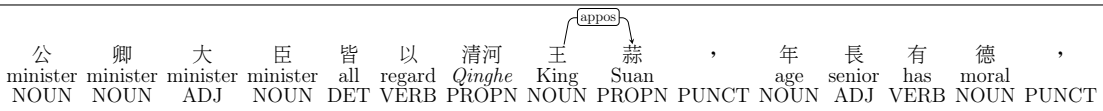


[C.32]

appos: appositional modifier

Example:

ID: liang-han-bian-wang:3:4
Sentence: 公卿大臣皆以清河王蒜，年長有德...
Translation: Every minister all regarded that the King of Qinghe, Suan, was senior and had good morals...

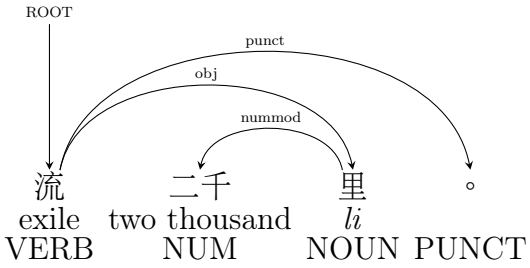


[C.33]

C.2.5 Word segmentation

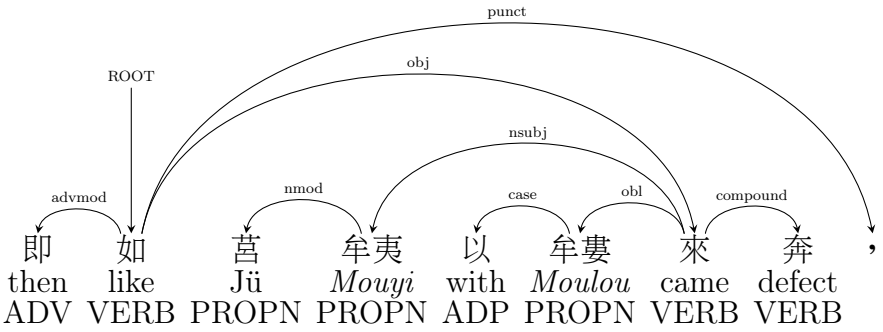
Each character is treated as one word, with the following two exceptions. First, numerals, no matter how many characters they have, are treated as one word. For example, in [C.34] the numeral 二千 (two thousand) is treated as one word, even though it consists of two characters. Second, proper nouns, such as personal names or place names, are also treated as one word. See [C.35].

ID: 7:71:1:1
Sentence: 流二千里。
Translation [276]: exile at a distance of 2000 *li*.



[C.34]

ID: 1:6:4:s1
Sentence: 即如莒牟夷以牟婁來奔，
Translation [244]: Like Mou-yi of Jü came as a fugitive, giving [the city of] Mou-Lou.



[C.35]

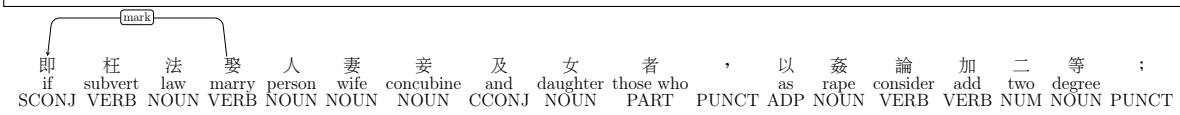
C.2.6 Conventions on common constructions

Constructions involving "者"

One of the most common constructions in the Tang Code is "< A >者, < B >"., where A describes the situation of the offense, and B states the punishment of the offense. In this case, the head of the highest node of A is "者" via the dependency relation **acl**, and "者" connects to the ROOT of the sentence via **nsubj** or **obl**, following conventions in C.2.7. See [C.58, C.59] for examples.

It is also common for these constructions to start with subordinate conjunctions such as "若", "即", or "其" (all translated as "if"). The head of these conjunctions are all the main verb of the clause, and the dependency relation is **mark**. See [C.36, C.37, C.38] for examples.

ID: 14:186:2:1
Sentence: 即枉法娶人妻妾及女者，以姦論加二等；
Translation [276]: If a supervisory official subverts the law by taking another man's wife, concubine, or daughter as his concubine, the punishment is for illicit sexual intercourse increased two degrees.



[C.36]

ID: 10:125:1:2
Sentence: 若依式應須遣使詣闕而不遣者，罪亦如之。
Translation [276]: In cases where, according to the statuses, documents should be sent to the palace by courier and they are not, the punishment is the same.

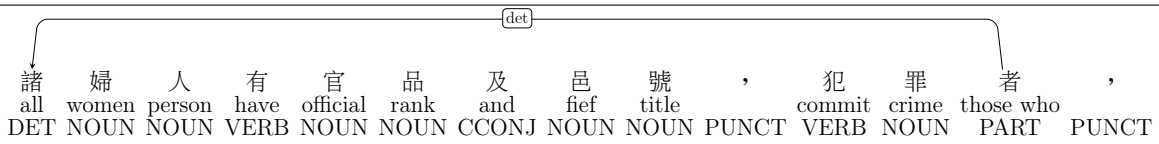
[C.37]

ID: 14:186:1:2
Sentence: 其在官非監臨者，減一等。
Translation [276]: For officials who are not supervisory officials, the punishment is reduced one degree.

[C.38]

Determiners like "諸" can also start a sentence, and in this case, the head of "諸" is also "者", and the dependency relation in this case is **det**. This translates to "All those who...", or "in all cases where...". An example is given in [C.24].

ID: 2:12:1:1
Sentence: 諸婦人有官品及邑號，犯罪者，
Translation [244]: All cases in which women who have official rank or fief-titles commit offenses...



[C.39]

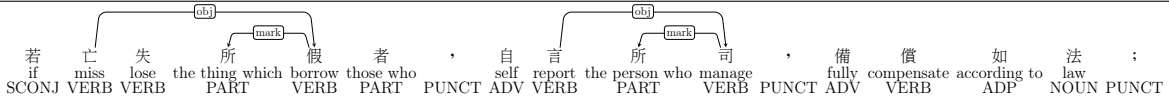
Constructions involving "所"

One function of "所" is to combine with verbs to form noun phrases representing the location, the personnel, or the entity carrying out the action [277]. Following [278], the head of "所" is the verb via the dependency relation **mark**. The dependency relation between the verb and its head depends on the function of the resulting 所-phrase.

[C.40] and [C.41] are two examples. There are three verbs: "假 (borrow)", "司 (manage)", and "著 (wear)". They combine with "所" to form noun phrases "所假", "所司", and "所著",

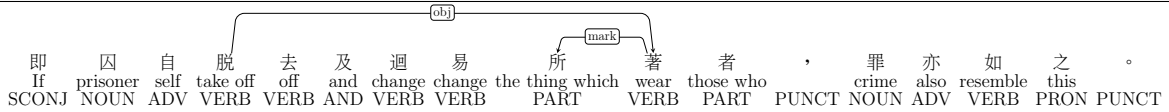
which are translated to “the object which they borrowed”, “the person who manages”, and “the object which they wear”, respectively. These noun phrases all serve as the object of the corresponding clause.

ID: 15:211:2:1
Sentence: 若亡失所假者，自言所司，備償如法；
Translation [276]: If what has been borrowed is lost and the person reports it to the government office himself, the value must be repaid in accordance with the law.



[C.40]

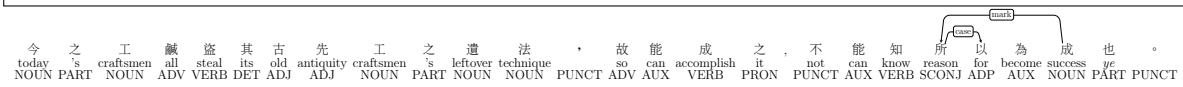
ID: 29:469:2:1
Sentence: 即囚自脱去及迴易所著者，罪亦如之。
Translation [276]: If the prisoners themselves remove the means of restraint or change them, the punishment is the same as above.



[C.41]

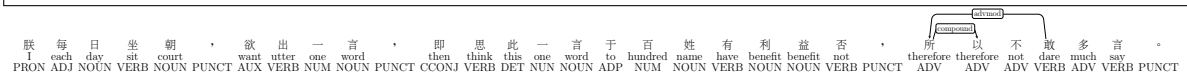
“所” usually combines with “以” to form the construction “所以”. It can either mean “therefore” or “the reason of”. When it means “therefore”, we treat it as a compound adverb. When it means “the reason of”, the head of “以” is “所”, with the dependency relation **case**, while “所” connects to its head as **mark**; see [C.42, C.43].

ID: ji-ji-ji:3:7
Sentence: 今之工鹹盜其古先工之遺法，故能成之,不能知所以為成也。
Translation: The craftsmen today all steal old techniques left by the master craftsmen of antiquity, so they are able to accomplish the work, but they do not know what makes it successful.



[C.42]

ID: shen-yan-yu-di:1:1
Sentence: 朕每日坐朝，欲出一言，即思此一言于百姓有利益否，所以不敢多言。
Translation: Every day when I hold court, whenever I am about to say a word, I ask myself whether it will benefit my people, so I do not speak excessively.

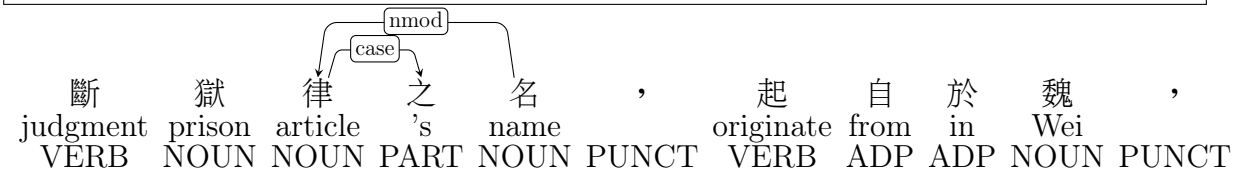


[C.43]

Constructions involving "之"

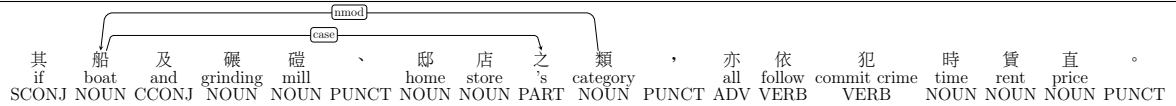
"之" as a particle usually appears in the form of "< A >之< B >". If A is a noun phrase, then "之" marks the genitive case. In this case, it roughly translates to "< A >'s< B >". "之" connects to the highest node of A, with the dependency relation **case**. Meanwhile, A connects to B via **nmod**. See [C.44, C.45] for examples.

ID: 29:s1
Sentence: 斷獄律之名，起自於魏，
Translation [276]: The name of the Articles on Judgment and Prison began during the Wei dynasty.



[C.44]

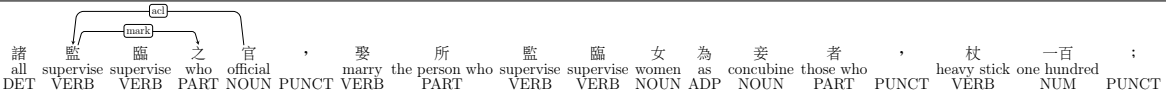
ID: 4:34:3:1
Sentence: 其船及碾磑、邸店之類，亦依犯時賃直。
Translation [244]: Assessment of the value of such things as boats, grinding mills, warehouses, and wholesale stores is also according to the value of the rent at the time of the offense.



[C.45]

On the other hand, if A is an adjective or a relative clause, "之" serves as a relativizer and connects to the highest node of A with **mark:rel**, which is merged with **mark** in our analysis. See [C.46] for an example.

ID: 14:186:1:1
Sentence: 諸監臨之官，娶所監臨女為妾者，杖一百；
Translation [276]: All cases of supervisory officials who take women within their area of jurisdiction as concubines are punished by one hundred blows with the heavy stick.



[C.46]

Constructions involving "為"

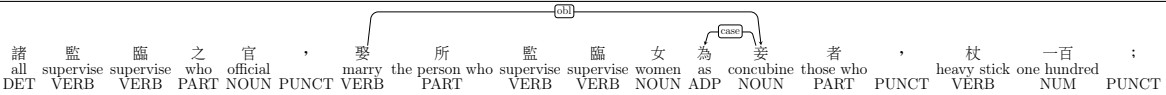
One construction where "為" appears is "為$< A >< B >$", where A is a noun and B is a verb. In this case, "為" specifies the beneficiary of the verb, which translates to the preposition "for". "為" connects to $< A >$ with **case**, and $< A >$ connects to $< B >$ with **obl**. An example is provided in [C.47].

ID: 14:186:1:1
Sentence: 若為親屬娶者，亦如之。
Translation [276]: If a supervisory official takes a woman as concubine for one of his relatives, the punishment is the same.

[C.47]

"為" can also appear in constructions such as "$< A >< B >$為$< C >$", where A is a verb, and B, C are nouns. In this case, "為" specifies the role of B as a result of action A, and it roughly translates to the preposition "as". "為" connects to C with **case**, and C connects to A with **obl**. An example is shown in [C.48]

ID: 14:186:1:1
Sentence: 諸監臨之官，娶所監臨女為妾者，杖一百；
Translation [276]: All cases of supervisory officials who take women within their area of jurisdiction as concubines are punished by one hundred blows with the heavy stick.

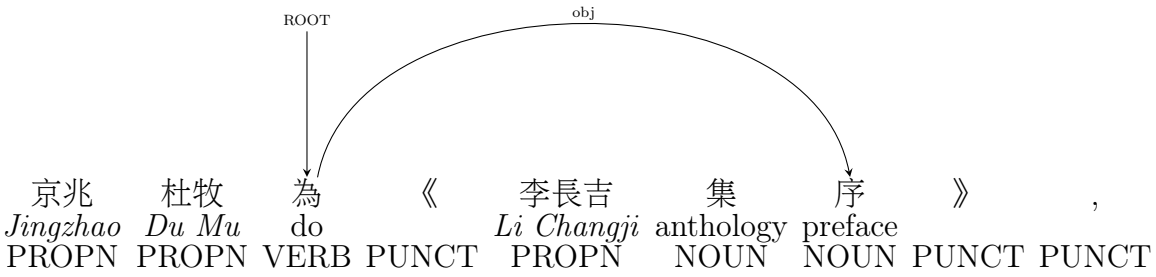


[C.48]

"為 $< A >$所$< B >$" is another common construction. It signifies the passive voice: A is the agent of the action of verb B. In this case, "為" connects to A with **case**, and following conventions on "所", "所" connects to B with **mark**. Finally, A connects to B with **obl**.

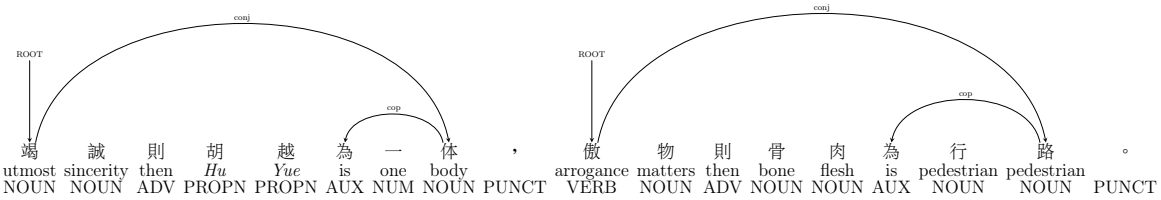
Apart from as a preposition, "為" can also be a verb meaning "to do" or a copula. For example, in [C.49], "為" is broadly translated as "to do" ("to compose" to be more specific), so it is annotated as verb and parsed as the root of the sentence. In contrast, in [C.50], "為" is broadly translate as a copula, so it is annotated as an auxiliary verb and is connected to the predicate as **cop**.

ID: li-he-xiao-zhuan:1:1
Sentence: 京兆杜牧為《李長吉集序》,
Translation: Du Mu from Jingzhao composed *The Preface for the Li Changji Anthology.*



[C.49]

ID: jian-tai-zong-shi:2:5
Sentence: 竭誠則胡越為一体，傲物則骨肉為行路。
Translation: If one has utmost sincerity, even peoples as remote as Hu and Yue become one body. If one is arrogant towards others, even one's kins become strangers.

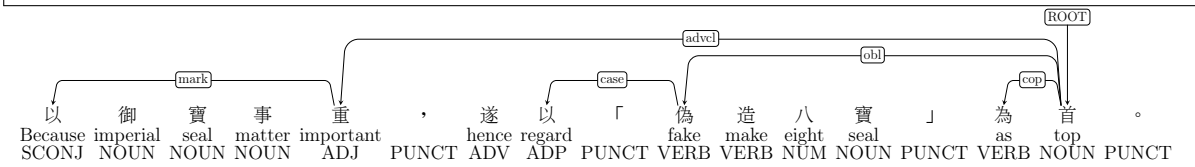


[C.50]

Constructions involving "以"

"以" could be a subordinate conjunction licensing an adverbial clause to express reason or purpose. In this case, "以" connects to the top node of the adverbial clause with **mark**. Another common construction involving "以" is "以< A >為< B >", where A is a noun. This translates to "to regard A as B". We analyze "以" as an adposition and "為" as an auxiliary verb, and therefore the head of "以" is A with the dependency relation **case**. Accordingly, A connects to B with **obl**. See [C.51] for an example. The the first "以" is an subordinate conjunction meaning "because", whereas the second "以" is an adposition meaning "regard as".

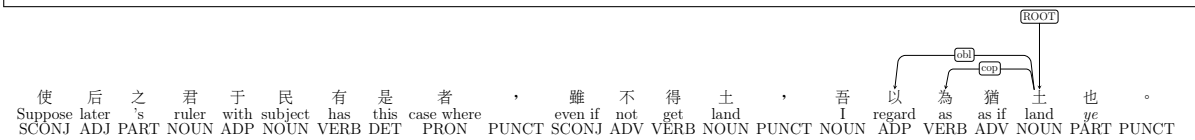
ID: 25:s2
Sentence: 以御寶事重，遂以「偽造八寶」為首。
Translation [276]: Because offenses involving the imperial seals are the most serious, therefore counterfeiting the eight seals is put at the head of this section.



[C.51]

A is often omitted in the construction "以< A >為< B >", which roughly translates to "to think". In this case, we still analyzed "以" as an adposition, but its dependency relation would be **obl**. See [C.52].

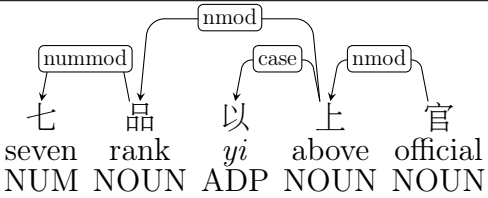
ID: du-si-ma-fa:1:10
Sentence: 使后之君于民有是者，雖不得土，吾以為猶土也。
Translation: If the rulers in later generations can treat their subjects like this, even if they do not conquer the world, I think it is as if they conquered the world.



[C.52]

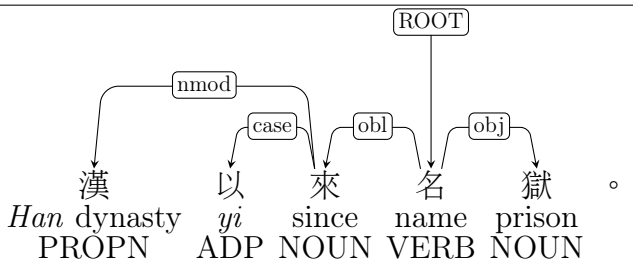
Other constructions involving "以" include "< A >以上 (and above)", "< A >以來 (since)", and "< A >以下 (below)", where A can be a noun or a verb. In these cases, we regard "上", "來", and "下" as the top node of the corresponding phrase, with "以" connecting to them with **case** and A connecting to them with **nmod** or **acl**. Two examples are provided in [C.53] and [C.54].

ID: 2:10:1:1
Sentence: 七品以上官
Translation [244]: officials of the seventh rank and above...



[C.53]

ID: 29:s4
Sentence: 漢以來名獄。
Translation: It is named *yu* since the *Han* dynasty.

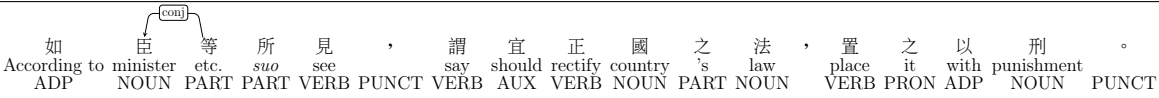


[C.54]

Constructions involving "等"

Apart from a verb meaning "to wait", or a noun meaning "degree", "等" could also be a pronoun marking the ending of an enumeration. In this case, "等" connects to the first element of the enumeration with **conj**, just like the rest of the elements. An example is shown in [C.55].

ID: fu-chou-yi-zhuang:1:20
Sentence: 如臣等所見，謂宜正國之法，置之以刑。
Translation: According to what we, the ministers, see, it would be appropriate to rectify the laws of the country and to deal with it by punishment.



[C.55]

C.2.7 Misc

xcomp and ccomp

In our analysis, open clausal complement (**xcomp**) and clausal complement (**ccomp**) are not used. Instead, they are broadly labelled as **conj** or **obj**.

For example, in [C.56], "有 (there is)" is the top node of a clause that serves as the object of "知 (to know)". In Universal Dependencies [246], this relation would be labelled as **ccomp**, but in our analysis, for simplicity, it is labelled as **obj**.

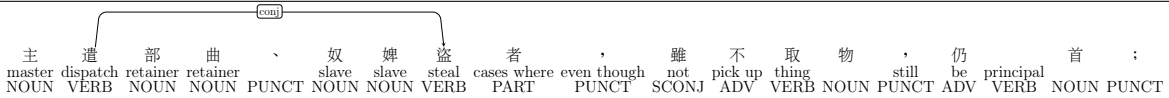
Example:

ID: 9:101:1:1
Sentence: 諸廟享，知有緦麻以上喪，遣充執事者，笞五十；
Translation [244]: All cases of putting a person in charge of a court celebration while knowing that he is in mourning of a relative within the fifth degree of mourning are punished by fifty blows with the light stick.

[C.56]

As another example, in [C.57], there are two verbs in the subject clause: "遣 (dispatch)" and "盜 (rob)". The object of the first verb "部曲、奴婢 (retainers and slaves)" are implicitly the subject of the second verb. In Universal Dependencies [246], the second verb connects to the first verb by **xcomp**, but in our analysis, for simplicity, the relation merged with **conj**.

ID: 20:297:4:1
Sentence: 主遣部曲、奴婢盜者，雖不取物，仍首；
Translation [276]: If a master sends a personal retainer or a slave to commit a robbery, even if he gets none of the goods, he is still the principal.



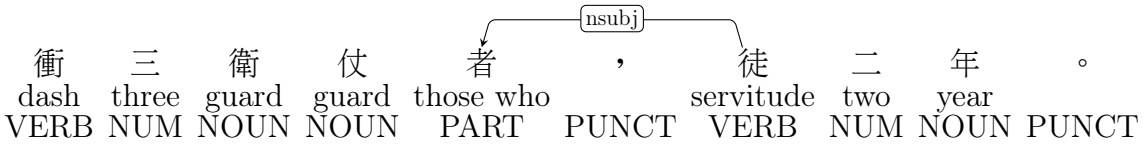
[C.57]

nsubj, obj, or obl

It's common in the Tang Code that a noun phrase precedes a verb phrase, and in many cases, the noun phrase serves as the subject, and the verb phrase serves as the predicate. However, this is not always the case, as the noun phrase can serve as the patient of the verb phrase, or it can serve as a topicalizer (translated as "in cases where"). For the former case, the dependency relation is **nsubj:pass** in Universal Dependencies, but in our analysis, it is merged with **nsubj**. For the latter case, the dependency relation is **obl**.

For example, in [C.58], the noun phrase ("those who interfere with the three ranks of the imperial guardsmen") is the patient of the verb ("punish"), because the violators are being punished, rather than doing the punishment. Therefore, the two clauses are linked together via the dependency relation **nsubj:pass**, which is merged into **nsubj** in our analysis.

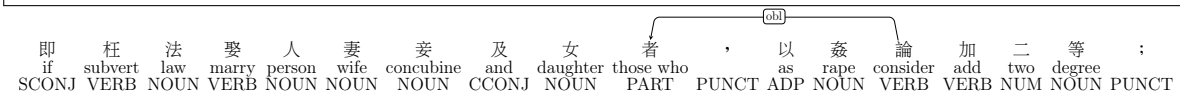
ID: 7:74:1:1
Sentence: 衝三衛仗者，徒二年。
Translation [276]: Those who interfere with the three ranks of the imperial guardsmen are punished by two years of penal servitude.



[C.58]

On the other hand, in [C.59], although the sentence structure is similar, the noun phrase here serves as a topicalizer and is roughly translated as "in cases where". Instead of being the agent or the patient of the predicate, it provides additional information, and therefore the dependency relation here is **obl**. The real subject of the sentence is "the punishment", which is omitted and should be inferred from context.

ID: 14:186:2:1
Sentence: 即枉法娶人妻妾及女者，以姦論加二等；
Translation [276]: If a supervisory official subverts the law by taking another man's wife, concubine, or daughter as his concubine, the punishment is for illicit sexual intercourse increased two degrees.

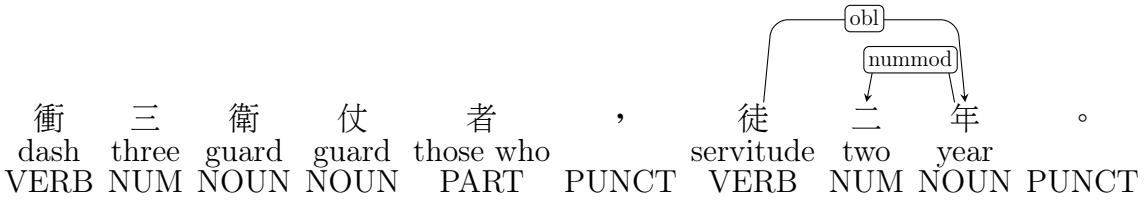


[C.59]

In our analysis, we followed the principle that if a sentence segment can be reasonably interpreted as the subject or the patient of the following predicate, the dependency relation will be **nsubj**. On the other hand, the dependency relation will be **obl.**

In the *Tang Code*, the punishments are generally described in two parts: the type of the punishment and the degree of the punishment. For example, in [C.60], the verb "徒" describes the type of the punishment (penal servitude), and the phrase "二年" describes the degree, or in this case, the duration of the punishment (two years). In this paper, we parsed the degree of punishment as **obl**. See [C.60] for an example.

ID: 7:74:1:1
Sentence: 衝三衛仗者，徒二年。
Translation [276]: Those who interfere with the three ranks of the imperial guardsmen are punished by two years of penal servitude.



[C.60]

Another common way to describe punishments is to state the added/reduced degree compare to the punishment of another offense. In [C.61], the punishment of killing or wounding a person in play is defined as two degrees less than that of killing or wounding a person in affray. In our convention, the punishment being compared to is analyzed as **obl**, and the degree reduced is analyzed as **obj**.

ID: 23:338:1:1
Sentence: 諸戲殺傷人者，減鬥殺傷二等；
Translation [276]: All cases of killing or wounding a person in play are punished two degrees less than for killing or wounding a person in affray.

[C.61]

compound and amod/nmod

amod/nmod are used in multi-character expressions (usually 2 characters) where the first character modifies the second character. If the first character is a noun, the relation is **nmod**, and if the first character is an adjective, then the relation is **amod**. Examples include "獄官 (prison's official; **nmod**)", "長官 (senior official; **amod**)", and "案狀 (case document; **nmod**)". In all these cases, the first character modifies the second character. The head of these expressions is the second character.

On the other hand, **compound** is used in multi-character expressions where both character are of equal status, or they don't modify each other. Examples include "車馬 (carriages + horses)", "殺傷 (kill + injure)", "街巷 (streets + alleys)", and "參驗 (refer + investigate)". In these cases, both characters within each expression are equal to each other. The head of these expressions is the first character.

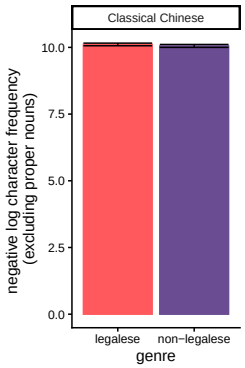


Figure C.1: **In classical Chinese, legalese texts exhibit similar average word frequency as non-legalese ones**. Mean negative log word frequency per sentence ($y-$axis) for classical Chinese legalese and non-legalese texts ($x-$axis).

C.3 Analysis on word frequency

As a proxy for comprehension difficulty, besides average dependency length, we also looked at a second variable in this study: average word frequency per sentence. Similar to dependency length, it is also shown to contribute to comprehension difficulty in legalese texts [45], although to a lesser extent. For classical Chinese, we drew the frequency data from a classical Chinese character frequency list, published by the State Language Commission (國家語言文字工作委員會)[1]. Since in our analysis, one character corresponds to one word, word frequency is hence approximated by character frequency. However, because numbers and proper nouns are treated as single words regardless of their length, they were excluded from the average frequency calculation. For modern Chinese, we drew word frequency data from the Python library wordfreq (Version 3.1.1), [279], which aggregates Chinese word frequency data from Wikipedia, subtitles, news, books, web texts, Twitter/X, and Reddit.

Table C.4 and Figure C.1 show the average negative log frequency for each genre in classical Chinese. We ran a mixed-effects linear regression on the data, with the same random effect as the previous analysis. We found no difference in negative log word frequency between the two genres ($\beta = -0.04$, SE $= 0.068$, $p = 0.497$). Results in modern Chinese are shown in Table C.4 and Figure C.2A. We adopted the subsampling analysis approach outlined in the main text to control for the unbalanced sample size. We also included results from [222] to compare the effect size. We set the reference levels for language and genre to be Chinese and legalese, respectively. Among the 500 subsampling analysis, the main effect of genre were all above zero (Figure C.2B): non-legalese texts on average contain lower frequency words than legalese ones. This result contrasts with what was reported in [45], where American legalese texts contain lower frequency words than non-legalese ones. It also leads to the significant positive interaction between language and genre (Figure C.2C).

We also calculated the negative log word frequency for additional legalese texts, including laws in Hong Kong and Taiwan, and sub-statutory regulations in mainland China. The results are shown in Table C.5 and Figure C.3A. Subsampling analyses show that all these

[1] https://github.com/bedlate/cn-corpus/tree/master/

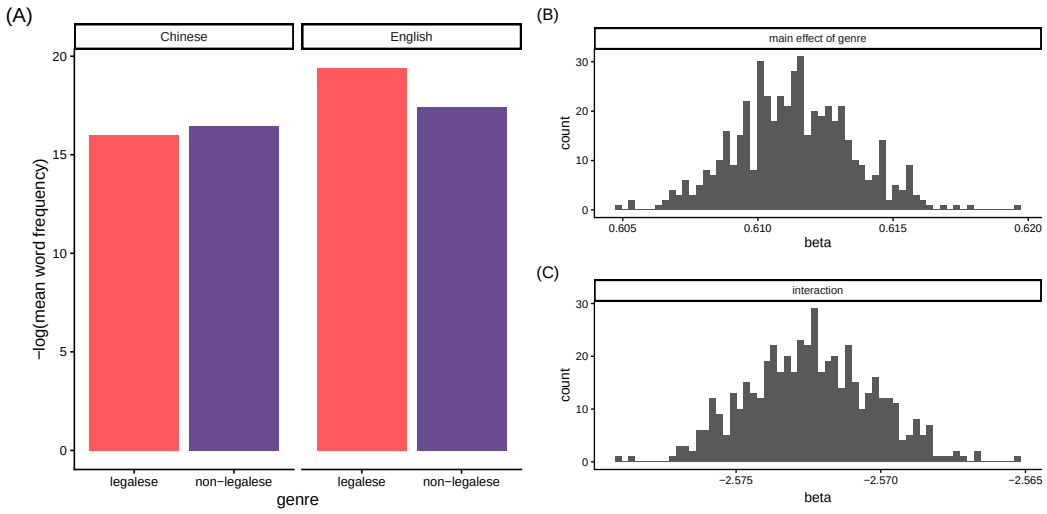


Figure C.2: **Legalese texts in mainland China exhibit higher average word frequency than non-legalese texts**. (A) Mean negative log word frequency per sentence for modern Chinese texts. The English data is drawn from [222] as a comparison. (B-C) Results from the sub-sampling analysis: (B) Distribution of genre effect estimates (beta) obtained from 500 comparisons between the legalese data and the size-matched subsamples of the non-legalese data. (C) Distribution of interaction between genre and language, obtained from 500 comparisons between the legalese data and the size-match subsamples of the non-legalese data, along with legalese and non-legalese data from a previous study on American English [222].

legalese texts on average contain significantly more frequent words than non-legalese texts (See Figure C.3B-C), which is also in the contrary of the findings in [45] on American English. These results imply that legalese written in both classical Chinese and modern Chinese do not contain low-frequency jargons that impede understanding like in American legalese.

C.4 Reliability of Stanza

In our modern Chinese analysis, we used Stanza [252] to automatically annotate and parse the legalese and non-legalese texts. In this analysis, we checked how much Stanza is reliable in annotation and parsing, and more importantly, how much Stanza is reliable in identifying long dependency sentences. To do so, we randomly sampled sentences from each genre, hand parsed them, and compared the results with the ones by Stanza. We repeated this process 3 times and hence obtained 3 batches of texts. Each batch contained approximately 2000 dependencies.

In each batch, to assess whether Stanza is accurate in identifying each word's head and its dependency relation with the head, we calculated the proportion of words where Stanza identifies the same head as us ("head accuracy") and the proportion of words where Stanza assigns the same dependency relation as us ("deprel accuracy"). The results are shown in Table C.6. On average, the parser agreed with our parsing only 61% of the time in identifying the head, and 66% in identifying the dependency relation.

Then, to assess whether Stanza could at least identify sentences with longer average dependency length reasonably accurately, we calculated two average dependency lengths for each sentence: one from Stanza, and the other from our hand parsing. We removed the dependencies labeled as **root**, **parataxis**, **flat:foreign**, and unlicensed **advcl** from the calculation. We then plotted both dependency lengths against one another. Ideally, if Stanza parsing and our hand parsing agree with each other, we would expect all the data points to be close to the diagonal line. The results were shown in Figure C.4. Across batches, the average dependency lengths calculated by Stanza's parsing and our manual parsing are reasonably similar. The results suggest that although our manual parsing and Stanza often disagree on the heads and dependency relations of individual words, both identify the same sentences as having long dependencies.

Table C.2: (Continued) the genre, article name, and author of the 40 non-legalese texts in this study.

Genre	Article	Author
Diary	右溪記 (Record on Youxi)	元結 (Yuan Jie, 719-772 CE)
Fiction	長恨傳 (Biography of Everlasting Sorrow)	陳鴻 (Chen Hong, 9th century)
Fiction	霍小玉傳 (Biography of Huo Xiaoyu)	蔣防 (Jiang Fang, 792-835 CE)
Biography	府君墓誌銘 (Epitaph for Zhang Fujun)	張説 (Zhang Shuo, 663-730 CE)
Commentary	吳季子箚論 (On Wu's Nobleman Zha)	獨孤及 (Dugu Ji, 726-777 CE)
Short story	説天雞 (On the Outstanding Rooster)	羅隱 (Luo Ying, 833-910 CE)
Diary	醉鄉記 (Record on the Land of Drunkenness)	王績 (Wang Ji, 585-644 CE)
Letter	賀行軍陸大夫書 (Letter to Grand Master Lu on His Military Campaign)	李翺 (Li Ao, 774-836 CE)
Letter	上常州盧使君書 (Letter to Governor Lu of Changzhou)	孟郊 (Meng Jiao, 751-814 CE)
Diary	太原王公同州修堰記 (Record of Duke Wang of Taiyuan Repairing the Weir in Tongzhou)	司空圖 (Sikong Tu, 837-908 CE)
Letter	與馮陶書 (Letter to Feng Tao)	沈亞之 (Shen Yazhi, 781-832 CE)
Commentary	定祀九宮儀注議 (Commentary on the Ritual Regulations for the Nine Palace Device)	王起 (Wang Qi, 760-847 CE)
Commentary	請東北將吏刊石紀功德狀 (Document Requesting the Northeastern Commanders to Carve a Stone Inscription Commemorating Their Merits)	張九齡 (Zhang Jiulin, 678-740 CE)
Commentary	三國論 (On the Three Kingdoms)	王勃 (Wang Bo, 650-676 CE)
Diary	撫州寶應寺翻經臺記 (Record on the Sutra-Translation Platform at Baoying Temple in Fuzhou)	顔真卿 (Yan Zhenqing, 709-785 CE)
Short story	唐城客夢 (A Traveler's Dream in Tang City)	黃滔 (Huang Tao, 840-911 CE)
Short story	甘露述 (Account of the Sweet Dew)	歐陽詹 (Ouyang Zhan, 755-800 CE)
Biography	周書·李遷哲列傳 (Book of Zhou - Biography of Li Qianzhe)	令狐德棻 (Linghu Defen, 582-666 CE)
Short story	貞觀政要·慎言語第二十二 (The Political Program in Zhenguan Times - On caution in speech)	吳兢 (Wu Jing, 670-749 CE)
Biography	晉書·曹志列傳	房玄齡

Table C.3: Speeches included in our analysis

Speaker	Title / Occasion
Bai Chunli (白春禮)	The science, technology, and innovation in China
Fernando Chui Sai On (崔世安)	The inaugural speech of the 4th Chief Executive Officer of the Macau Special Administrative Region
Hu Shih (胡適)	There is no effort in vein
Huai Jinpeng (懷進鵬)	Keynote speech at the World Digital Education Conference
Liu Xiaoming (劉曉明)	Speech at the Eton College
Mao Zedong (毛澤東)	Speech at the Founding Ceremony of the People's Republic of China
Wen Jiabao (温家寶)	Speech at Harvard University
Wen Yiduo (聞一多)	The last speech
Xi Jinping (習近平)	Speech at the opening ceremony of the Belt and Road Forum
Zhou Xiaochuan (周小川)	Keynote speech at the 2009 Lujiazui Forum

Language	Genre	# sentences	Avg. $-\log p_{\text{word}}$
Classical Chinese	Legalese	614	10.2
	Non-legalese	485	10.2
Modern Chinese	Laws, mainland China	1556022	15.9
	Non-legalese	13876111	16.5
English	Legalese	1308424	19.4
	Non-legalese	6074672	17.4

Table C.4: Number of sentences and the mean negative log word frequency in each type of text. Higher values suggest higher comprehension difficulty. Note: the average negative log word frequency have different magnitudes in different languages because the frequencies are calculated from different sources.

Language	Genre	# sentences	Avg. $-\log p_{\text{word}}$
Modern Chinese	Statute laws, mainland China	35145	20.3
	Laws, Hong Kong	177363	20.7
	Laws, Taiwan	652191	21.9
	Sub-statutory regulations, mainland China	1520877	21.0
English	Legalese	1308424	19.4
	Non-legalese	6074672	17.4

Table C.5: Number of sentences and the mean negative log word frequency in each type of text. Higher values suggest higher comprehension difficulty. The English data is drawn from [222] as a comparison. Note: the average negative log word frequency have different magnitudes in different languages because the frequencies are calculated from different sources.

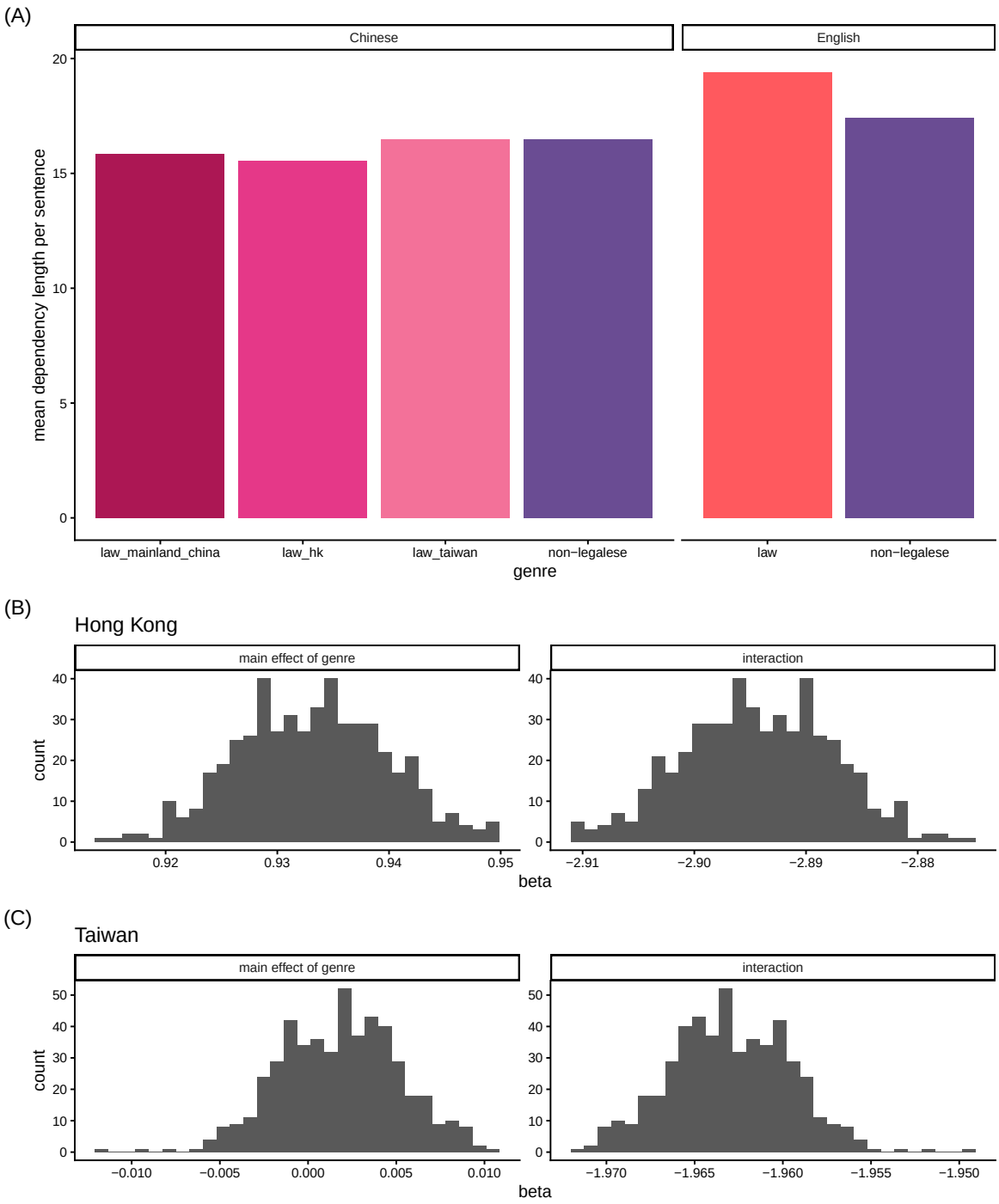


Figure C.3: **Legalese texts from Hong Kong and Taiwan exhibit higher average word frequency than non-legalese texts**. (A) Mean negative word frequency per sentence for legalese texts from Hong Kong and Taiwan, along with non-legalese texts in modern Chinese. Results from American English are pulled from [222], as a comparison of the effect size. Acronyms: HK - Hong Kong. (B) Distribution of (left) main effect of genre and (right) interaction between genre and language, obtained from 500 comparisons between the Hong Kong legalese data and the size-match subsamples of the non-legalese data, along with legalese and non-legalese data from a previous study on American English [222]. (C) Same as (B) but the legalese data is from Taiwan.

Table C.6: Parsing accuracy by genre and batch. The genre and batch with the highest accuracy is bolded, whereas the genre and batch with the lowest accuracy is underlined. Acronyms: CN - mainland China, HK - Hong Kong, TW - Taiwan.

Genre	Batch	n dependencies	Head Acc.	Deprel Acc.
law (CN)	batch1	127	0.677	0.795
law (CN)	batch2	232	0.659	**0.802**
law (CN)	batch3	194	0.598	0.680
regulation	batch1	129	0.690	0.651
regulation	batch2	235	0.740	0.753
regulation	batch3	216	0.685	0.736
interpretation	batch1	165	0.636	0.636
interpretation	batch2	207	0.715	0.725
interpretation	batch3	229	0.729	0.738
resolution	batch1	141	0.582	0.688
resolution	batch2	255	0.639	0.714
resolution	batch3	197	0.614	0.756
law (HK)	batch1	130	0.600	0.677
law (HK)	batch2	255	0.533	0.612
law (HK)	batch3	222	0.523	0.572
law (TW)	batch1	136	0.485	0.559
law (TW)	batch2	139	0.612	0.655
law (TW)	batch3	169	0.562	0.580
news	batch1	148	0.696	0.723
news	batch2	241	0.564	0.627
news	batch3	193	0.580	0.606
wikipedia	batch1	195	0.579	0.646
wikipedia	batch2	220	0.586	0.659
wikipedia	batch3	185	0.627	0.692
wechat	batch1	175	0.560	0.691
wechat	batch2	221	0.615	0.652
wechat	batch3	211	0.720	0.754
novels	batch1	163	0.436	0.589
novels	batch2	204	<u>0.392</u>	<u>0.480</u>
novels	batch3	122	0.615	0.664
hotel_review	batch1	163	0.491	0.564
hotel_review	batch2	236	0.636	0.661
hotel_review	batch3	238	0.479	0.597
speech	batch1	135	0.741	0.741
speech	batch2	229	0.646	0.699
speech	batch3	193	**0.793**	0.694

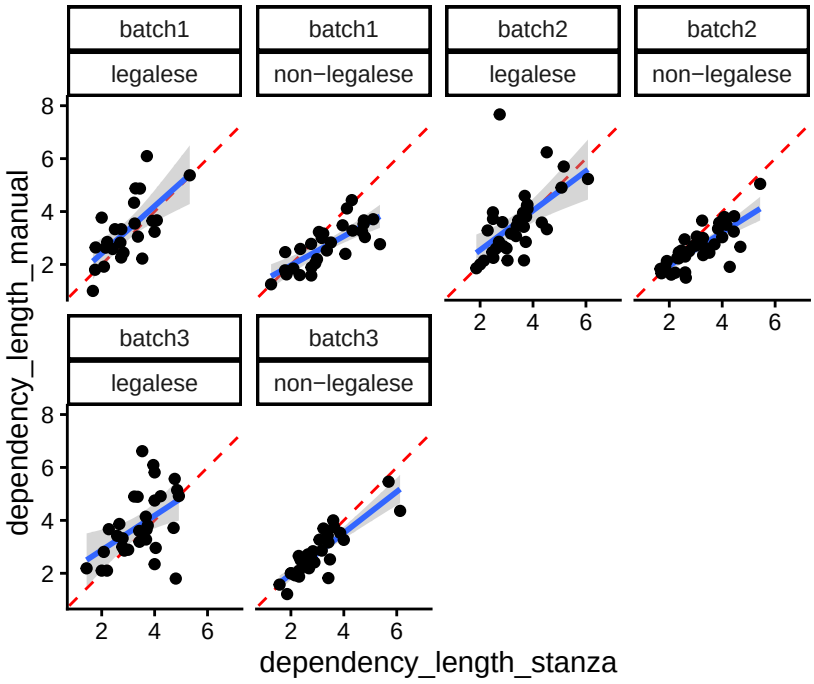


Figure C.4: The average dependency length per sentence by Stanza on the $x-$axis [252] and by our manual parsing ($y-$axis), organized by batches (rows) and genres (legalese vs. non-legalese, columns).

References

[1] E. Fedorenko, S. T. Piantadosi, and E. A. F. Gibson. "Language Is Primarily a Tool for Communication Rather than Thought." *Nature*, **630**(8017), 2024, pp. 575–586. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07522-w.

[2] M. Tomasello, A. C. Kruger, and H. H. Ratner. "Cultural learning." *Behavioral and brain sciences*, **16**(3), 1993, pp. 495–511.

[3] C. Green and S. Moran. "Buwal sound inventory (GM)." In: *PHOIBLE 2.0*. Ed. by S. Moran and D. McCloy. Jena: Max Planck Institute for the Science of Human History, 2019. URL: https://phoible.org/inventories/view/1402.

[4] S. Moran. "Danish sound inventory (UZ)." In: *PHOIBLE 2.0*. Ed. by S. Moran and D. McCloy. Jena: Max Planck Institute for the Science of Human History, 2019. URL: https://phoible.org/inventories/view/2167.

[5] G. G. Corbett. "Number of Genders (v2020.4)." In: *The World Atlas of Language Structures Online*. Ed. by M. S. Dryer and M. Haspelmath. Zenodo, 2013. DOI: 10.5281/zenodo.13950591. URL: https://doi.org/10.5281/zenodo.13950591.

[6] M. S. Dryer. "Order of Subject, Object and Verb (v2020.4)." In: *The World Atlas of Language Structures Online*. Ed. by M. S. Dryer and M. Haspelmath. Zenodo, 2013. DOI: 10.5281/zenodo.13950591. URL: https://doi.org/10.5281/zenodo.13950591.

[7] C. Everett, D. E. Blasi, and S. G. Roberts. "Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots." *Proceedings of the National Academy of Sciences*, **112**(5), 2015, pp. 1322–1327.

[8] C. Everett. "Languages in drier climates use fewer vowels." *Frontiers in psychology*, **8**, 2017, p. 1285.

[9] I. Maddieson and C. Coupé. "Human Language Diversity and the Acoustic Adaptation Hypthesis." *Proceedings of Meetings on Acoustics*, **25**(1), 2016, p. 060005. ISSN: 1939-800X. DOI: 10.1121/2.0000198.

[10] D. E. Blasi, S. Moran, S. R. Moisik, P. Widmer, D. Dediu, and B. Bickel. "Human sound systems are shaped by post-Neolithic changes in bite configuration." *Science*, **363**(6432), Mar. 2019. ISSN: 1095-9203. DOI: 10.1126/science.aav3218. URL: http://dx.doi.org/10.1126/science.aav3218.

[11] C. Everett and S. Chen. "Speech adapts to differences in dentition within and across populations." *Scientific Reports*, **11**(1), Jan. 2021. ISSN: 2045-2322. DOI: 10.1038/s41598-020-80190-8. URL: http://dx.doi.org/10.1038/s41598-020-80190-8.

[12] A. Wray and G. W. Grace. "The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form." *Lingua*, **117**(3), Mar. 2007, pp. 543–578. ISSN: 0024-3841. DOI: 10.1016/j.lingua.2005.05.005. URL: http://dx.doi.org/10.1016/j.lingua.2005.05.005.

[13] G. Lupyan and R. Dale. "Language structure is partly determined by social structure." *PloS one*, **5**(1), 2010, e8559.

[14] K. Smith. "Simplifications made early in learning can reshape language complexity: An experimental test of the Linguistic Niche Hypothesis." In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 46. 2024.

[15] O. Shcherbakova, S. M. Michaelis, H. J. Haynie, S. Passmore, V. Gast, R. D. Gray, S. J. Greenhill, D. E. Blasi, and H. Skirgård. "Societies of strangers do not speak less complex languages." *Science Advances*, **9**(33), 2023, eadf7704.

[16] E. Gibson, R. Futrell, J. Jara-Ettinger, K. Mahowald, L. Bergen, S. Ratnasingam, M. Gibson, S. T. Piantadosi, and B. R. Conway. "Color naming across languages reflects color use." *Proceedings of the National Academy of Sciences*, **114**(40), 2017, pp. 10785–10790.

[17] M. C. Frank, D. L. Everett, E. Fedorenko, and E. Gibson. "Number as a cognitive technology: Evidence from Pirahã language and cognition." *Cognition*, **108**(3), 2008, pp. 819–824.

[18] E. Gibson, R. Futrell, S. P. Piantadosi, I. Dautriche, K. Mahowald, L. Bergen, and R. Levy. "How efficiency shapes human language." *Trends in cognitive sciences*, **23**(5), 2019, pp. 389–407.

[19] T. Regier, P. Kay, and N. Khetarpal. "Color naming reflects optimal partitions of color space." *Proceedings of the National Academy of Sciences*, **104**(4), 2007, pp. 1436–1441.

[20] T. Regier, C. Kemp, and P. Kay. "11 Word Meanings across Languages Support Efficient Communication." *The Handbook of Language Emergence*, **87**, 2015, p. 237.

[21] N. Zaslavsky, C. Kemp, T. Regier, and N. Tishby. "Efficient compression in color naming and its evolution." *Proceedings of the National Academy of Sciences*, **115**(31), 2018, pp. 7937–7942.

[22] N. Zaslavsky, C. Kemp, N. Tishby, and T. Regier. "Communicative need in colour naming." *Cognitive Neuropsychology*, 2019, pp. 1–13.

[23] C. Kemp and T. Regier. "Kinship categories across languages reflect general communicative principles." *Science*, **336**(6084), 2012, pp. 1049–1054.

[24] N. Zaslavsky, M. Maldonado, and J. Culbertson. "Let's talk (efficiently) about us: Person systems achieve near-optimal compression." In: *Proceedings of the annual meeting of the cognitive science society*. Vol. 43. 43. 2021.

[25] F. Mollica, G. Bacon, N. Zaslavsky, Y. Xu, T. Regier, and C. Kemp. "The forms and meanings of grammatical markers support efficient communication." *Proceedings of the National Academy of Sciences*, **118**(49), 2021, e2025993118.

[26] N. Imel and S. Steinert-Threlkeld. "Modal semantic universals optimize the simplicity/informativeness trade-off." In: *Semantics and Linguistic Theory*. 2022, pp. 227–248.

[27] Y. Xu, E. Liu, and T. Regier. "Numeral systems across languages support efficient communication: From approximate numerosity to recursion." *Open Mind*, **4**, 2020, pp. 57–70.

[28] E. Gibson. "Linguistic complexity: Locality of syntactic dependencies." *Cognition*, **68**(1), 1998, pp. 1–76.

[29] E. Gibson. *Syntax: A cognitive approach*. MIT Press, 2026.

[30] E. Gibson and J. Thomas. "Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical." *Language and cognitive processes*, **14**(3), 1999, pp. 225–248.

[31] F. Hsiao and E. Gibson. "Processing relative clauses in Chinese." *Cognition*, **90**(1), 2003, pp. 3–27.

[32] D. Grodner and E. Gibson. "Consequences of the serial nature of linguistic input for sentenial complexity." *Cognitive science*, **29**(2), 2005, pp. 261–290.

[33] R. Futrell, K. Mahowald, and E. Gibson. "Large-scale evidence of dependency length minimization in 37 languages." *Proceedings of the National Academy of Sciences*, **112**(33), 2015, pp. 10336–10341.

[34] R. Ferrer i Cancho, R. V. Solé, and R. Köhler. "Patterns in syntactic dependency networks." *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, **69**(5), 2004, p. 051915.

[35] R. Keller. *On language change: The invisible hand in language*. Psychology Press, 1994.

[36] N. Levshina. *Communicative efficiency*. Cambridge University Press, 2022.

[37] K. Mahowald, E. Fedorenko, S. T. Piantadosi, and E. Gibson. "Info/information theory: Speakers choose shorter words in predictive contexts." *Cognition*, **126**(2), 2013, pp. 313–318.

[38] K. Smith, S. Kirby, and H. Brighton. "Iterated learning: A framework for the emergence of language." *Artificial Life*, **9**(4), 2003, pp. 371–386.

[39] S. Kirby, H. Cornish, and K. Smith. "Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language." *Proceedings of the National Academy of Sciences*, **105**(31), 2008, pp. 10681–10686.

[40] G. Roberts and M. Fedzechkina. "Social biases modulate the loss of redundant forms in the cultural evolution of language." *Cognition*, **171**, 2018, pp. 194–201.

[41] S. L. Walter and K. R. Ringenberg. "Language Policy, Literacy, and Minority Languages." *Review of Policy Research*, **13**(3-4), 1994, pp. 341–366.

[42] W. Van Langendonck. *Theory and typology of proper names*. De Gruyter Mouton, 2008.

[43] J. C. Scott. *Seeing like a state*. Yale University Press, 2008.

[44] Great Britain. Committee on the Preparation of Legislation. *The Preparation of Legislation*. Tech. rep. Chairman: Sir David Renton. London: HMSO, 1975.

[45] E. Martínez, F. Mollica, and E. Gibson. "Poor writing, not specialized concepts, drives processing difficulty in legal language." *Cognition*, **224**, 2022, p. 105070.

[46] P. M. Tiersma. "Some myths about legal language." *Law, Culture and the Humanities*, **2**(1), 2006, pp. 29–50.

[47] P. Tiersma. "The plain English movement." *Electronic version:< http://< grammar. ucsd. edu/courses/lign105/student-court-cases/plain% 20english. pdf*, 2007.

[48] P. Tiersma. "The nature of legal language." In: *Dimensions of forensic linguistics*. John Benjamins Publishing Company, 2008, pp. 7–25.

[49] J. Nintemann, M. Robbers, and N. Hober. *Here–Hither–Hence and related categories: A cross-linguistic study*. Vol. 26. Walter de Gruyter GmbH & Co KG, 2020.

[50] N. Tishby, F. C. Pereira, and W. Bialek. "The information bottleneck method." *arXiv preprint physics/0004057*, 2000.

[51] S. C. Levinson. "Language and space." *Annual review of Anthropology*, **25**(1), 1996, pp. 353–382.

[52] T. Stolz, N. Levkovych, A. Urdze, J. Nintemann, and M. Robbers. *Spatial interrogatives in Europe and beyond*. De Gruyter Mouton, 2017.

[53] M. Maldonado and J. Culbertson. "Here, there and everywhere: An experimental investigation of the semantic features of indexicals." *https://ling.auf.net/lingbuzz/005628*, 2020.

[54] S. C. Levinson, S. Kita, D. B. Haun, and B. H. Rasch. "Returning the tables: Language affects spatial reasoning." *Cognition*, **84**(2), 2002, pp. 155–188.

[55] S. P. Gennari, S. A. Sloman, B. C. Malt, and W. T. Fitch. "Motion events in language and cognition." *Cognition*, **83**(1), 2002, pp. 49–79.

[56] E. Pederson, E. Danziger, D. Wilkins, S. Levinson, S. Kita, and G. Senft. "Semantic typology and spatial conceptualization." *Language*, **74**(3), 1998, pp. 557–589.

[57] S. Jarvis and A. Pavlenko. *Crosslinguistic influence in language and cognition*. Routledge, 2008.

[58] T. Cadierno. "Expressing motion events in a second language: A cognitive typological perspective." *Cognitive linguistics, second language acquisition, and foreign language teaching*, 2004, pp. 13–49.

[59] E. Danziger. "Deixis, gesture, and cognition in spatial frame of reference typology." *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, **34**(1), 2010, pp. 167–185.

[60] S. C. Levinson and D. P. Wilkins. *Grammars of space: Explorations in cognitive diversity*. Cambridge University Press, 2006.

[61] J. Hawkins. *A performance theory of order and constituency*. Vol. 73. Cambridge, UK: Cambridge University Press, 1994.

[62] C. Kemp and T. Regier. "Kinship categories across languages reflect general communicative principles." *Science*, **336**(6084), 2012, pp. 1049–1054.

[63] F. Mollica, G. Bacon, Y. Xu, T. Regier, and C. Kemp. "Grammatical marking and the tradeoff between code length and informativeness." In: *CogSci*. 2020.

[64] N. Zaslavsky, T. Regier, N. Tishby, and C. Kemp. "Semantic categories of artifacts and animals reflect efficient coding." In: *41st Annual Meeting of the Cognitive Science Society*. 2019.

[65] C. Kemp, A. Gaby, and T. Regier. "Season naming and the local environment." In: *CogSci*. 2019, pp. 539–545.

[66] N. Tishby and N. Zaslavsky. "Deep learning and the information bottleneck principle." In: *2015 IEEE Information Theory Workshop (ITW)*. IEEE. 2015, pp. 1–5.

[67] D. Strouse and D. J. Schwab. "The deterministic information bottleneck." *Neural computation*, **29**(6), 2017, pp. 1611–1630.

[68] R. M. Dixon. "Demonstratives: A cross-linguistic typology." *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, **27**(1), 2003, pp. 61–112.

[69] K. R. Coventry, D. Griffiths, and C. J. Hamilton. "Spatial demonstratives and perceptual space: Describing and remembering object location." *Cognitive Psychology*, **69**, 2014. Publisher: Elsevier, pp. 46–70.

[70] H. Diessel. "13 Deixis and demonstratives." *Semantics-Interfaces*, 2019. Publisher: Walter de Gruyter GmbH & Co KG, p. 463.

[71] H. Diessel and K. R. Coventry. "Demonstratives in spatial language and social interaction: An interdisciplinary review." *Frontiers in Psychology*, **11**, 2020. Publisher: Frontiers Media SA, p. 555265.

[72] H. Diessel. "Demonstratives, joint attention, and the emergence of grammar." *Cognitive Linguistics*, 4, 2006.

[73] K. R. Coventry, B. Valdés, A. Castillo, and P. Guijarro-Fuentes. "Language within your reach: Near–far perceptual space and spatial demonstratives." *Cognition*, **108**(3), 2008, pp. 889–895.

[74] S. C. Levinson. "Introduction: demonstratives: patterns in diversity." In: *Demonstratives in cross-linguistic perspective*. Cambridge University Press, 2018, pp. 1–42.

[75] S. C. Levinson and S. C. Levinson. *Space in language and cognition: Explorations in cognitive diversity*. 5. Cambridge University Press, 2003.

[76] H. Diessel. "Bühler's two-field theory of pointing and naming and the deictic origins of grammatical morphemes." *Grammaticalization and language change: New reflections*, 2012. Publisher: John Benjamins Amsterdam, pp. 37–50.

[77] W. F. Hanks. "11. Deixis and indexicality." *Foundations of pragmatics*, **1**, 2011. Publisher: Walter de Gruyter, p. 315.

[78] W. F. Hanks. *Referential practice: Language and lived space among the Maya*. University of Chicago Press, 1990.

[79] C. J. Fillmore. *Lectures on deixis*. CSLI publications, 1997.

[80] R. D. Perkins. "Deixis, grammar, and culture." *Deixis, Grammar, and Culture*, 1992. Publisher: John Benjamins Publishing Company, pp. 1–255.

[81] K. Bühler. *Sprachtheorie*. Fischer, 1934.

[82] N. J. Enfield. "The definition of what-d'you-call-it: semantics and pragmatics of recognitional deixis." *Journal of pragmatics*, **35**(1), 2003. Publisher: Elsevier, pp. 101–117.

[83] D. Peeters and A. Özyürek. "This and That Revisited: A Social and Multimodal Approach to Spatial Demonstratives." *Frontiers in Psychology*, **7**, 2016. ISSN: 1664-1078. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00222 (visited on 10/26/2022).

[84] K. R. Coventry, D. Lynott, A. Cangelosi, L. Monrouxe, D. Joyce, and D. C. Richardson. "Spatial language, visual attention, and perceptual simulation." *Brain and language*, **112**(3), 2010. Publisher: Elsevier, pp. 202–213.

[85] J. O. P. García, K. R. Ehlers, and K. Tylén. "Bodily constraints contributing to multimodal referentiality in humans: the contribution of a de-pigmented sclera to proto-declaratives." *Language & Communication*, **54**, 2017, pp. 73–81.

[86] A. Bangerter. "Using pointing and describing to achieve joint focus of attention in dialogue." *Psychological science*, **15**(6), 2004, pp. 415–419.

[87] K. Cooperrider. "The co-organization of demonstratives and pointing gestures." *Discourse Processes*, **53**(8), 2016, pp. 632–656.

[88] D. Kemmerer. "Near and far in language and perception." *Cognition*, **73**(1), 1999. Publisher: Elsevier, pp. 35–63.

[89] R. Rocca, M. Wallentin, C. Vesper, and K. Tylén. "This is for you: social modulations of proximal vs. distal space in collaborative interaction." *Scientific reports*, **9**(1), 2019, pp. 1–14.

[90] S. R. Anderson and E. L. Keenan. "Deixis." *Language typology and syntactic description*, **3**(4), 1985, pp. 259–308.

[91] R. Jackendoff. *Semantics and cognition*. Vol. 8. MIT press, 1983.

[92] L. Lakusta and B. Landau. "Language and memory for motion events: Origins of the asymmetry between source and goal paths." *Cognitive science*, **36**(3), 2012, pp. 517–544.

[93] T. Nikitina. "Subcategorization pattern and lexical meaning of motion verbs: a study of the source/goal ambiguity." 2009.

[94] M. Haspelmath. "The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison." In: *The new psychology of language*. Psychology Press, 2003, pp. 217–248.

[95] T. Stolz, S. Lestrade, and C. Stolz. "The crosslinguistics of zero-marking of spatial relations." In: *The Crosslinguistics of Zero-Marking of Spatial Relations*. De Gruyter (A), 2014.

[96] T. Georgakopoulos and P. Karatsareas. "A diachronic take on the Source–Goal asymmetry." *Space in Diachrony*, **188**, 2017.

[97] R. Jackendoff. "The architecture of the linguistic-spatial interface." *Language and space*, **1**, 1996, p. 30.

[98] R. Jackendoff and B. Landau. "Spatial language and spatial cognition." In: *Bridges between psychology and linguistics*. Psychology Press, 2013, pp. 157–182.

[99] R. W. Langacker. "Reference-point constructions." In: *Mouton Classics*. De Gruyter Mouton, 2013, pp. 413–450.

[100] D. B. Haun, C. J. Rapold, G. Janzen, and S. C. Levinson. "Plasticity of human spatial cognition: Spatial language and cognition covary across cultures." *Cognition*, **119**(1), 2011, pp. 70–80.

[101] E. Ünal, C. Richards, J. C. Trueswell, and A. Papafragou. "Representing agents, patients, goals and instruments in causative events: A cross-linguistic investigation of early language and cognition." *Developmental Science*, 2021.

[102] E. Ünal, Y. Ji, and A. Papafragou. "From event representation to linguistic meaning." *Topics in cognitive science*, **13**(1), 2021, pp. 224–242.

[103] A. Papafragou. "Source-goal asymmetries in motion representation: Implications for language production and comprehension." *Cognitive science*, **34**(6), 2010, pp. 1064–1092.

[104] T. Regier and M. Zheng. "Attention to endpoints: A cross-linguistic constraint on spatial meaning." *Cognitive Science*, **31**(4), 2007, pp. 705–719.

[105] L. Lakusta and B. Landau. "Starting at the end: The importance of goals in spatial language." *Cognition*, **96**(1), 2005, pp. 1–33.

[106] T. Regier. *The human semantic potential: Spatial language and constrained connectionism*. MIT Press, 1996.

[107] M. Srinivasan and D. Barner. "The Amelia Bedelia effect: World knowledge and the goal bias in language acquisition." *Cognition*, **128**(3), 2013, pp. 431–450.

[108] M. Johanson, S. Selimis, and A. Papafragou. "The Source-Goal asymmetry in spatial language: language-general vs. language-specific aspects." *Language, Cognition and Neuroscience*, **34**(7), 2019, pp. 826–840.

[109] C. Pléh, Z. Vinkler, and L. Kálmán. "EARLY MORPHOLOGY OF SPATIAL EXPRESSIONS IN HUNGARIAN CHILDREN: A CHILDES STUDY." *Acta Linguistica Hungarica*, **44**(1/2), 1997, pp. 249–260. ISSN: 12168076, 15882624. URL: http://www.jstor.org/stable/44306789 (visited on 11/10/2022).

[110] E. Dromi. "More on the acquisition of locative prepositions: An analysis of Hebrew data." *Journal of Child Language*, **6**(3), 1979, pp. 547–562.

[111] M. L. Do, A. Papafragou, and J. Trueswell. "Cognitive and pragmatic factors in language production: Evidence from source-goal motion events." *Cognition*, **205**, 2020, p. 104447.

[112] Y. Chen, J. Trueswell, and A. Papafragou. "Source-Goal asymmetry in motion events: Sources are robustly encoded in memory but overlooked at test." 35th Annual Conference on Human Sentence Processing. 2022.

[113] M. Denić, S. Steinert-Threlkeld, and J. Szymanik. "Complexity/informativeness trade-off in the domain of indefinite pronouns." In: *Semantics and linguistic theory*. Vol. 30. 2021, pp. 166–184.

[114] S. Steinert-Threlkeld. "Quantifiers in natural language optimize the simplicity/informativeness trade-off." In: *Proceedings of the 22nd Amsterdam colloquium*. 2020, pp. 513–522.

[115] G. Zipf. *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley, 1949.

[116] M. Haspelmath. "Explaining grammatical coding asymmetries: Form–frequency correspondences and predictability." *Journal of Linguistics*, **57**(3), 2021, pp. 605–633. DOI: 10.1017/S0022226720000535.

[117] C. E. Shannon. "Coding theorems for a discrete source with a fidelity criterion." In: *IRE National Convention Record, Pt. 4*. 1959, pp. 142–163.

[118] T. Berger. "Rate-distortion theory." *Wiley Encyclopedia of Telecommunications*, 2003.

[119] P. Harremoës and N. Tishby. "The information bottleneck revisited or how to choose a good distortion measure." In: *Information Theory, 2007. ISIT 2007. IEEE International Symposium on*. IEEE. 2007, pp. 566–570.

[120] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, Apr. 2005. ISBN: 9780471748823. DOI: 10.1002/047174882x. URL: http://dx.doi.org/10.1002/047174882X.

[121] C. R. Twomey, G. Roberts, D. H. Brainard, and J. B. Plotkin. "What we talk about when we talk about colors." *Proceedings of the National Academy of Sciences*, **118**(39), 2021, e2109237118.

[122] R. Bhui, L. Lai, and S. J. Gershman. "Resource-rational decision making." *Current Opinion in Behavioral Sciences*, **41**, 2021, pp. 15–21.

[123] A. Zénon, O. Solopchuk, and G. Pezzulo. "An information-theoretic perspective on the costs of cognition." *Neuropsychologia*, **123**, 2019, pp. 5–18.

[124] R. Futrell. "An information-theoretic account of semantic inference in word production." *Frontiers in Psychology*, **12**, 2021, p. 672408.

[125] P. Kay, B. Berlin, L. Maffi, W. R. Merrifield, and R. Cook. *The world color survey*. CSLI Publications Stanford, CA, 2009.

[126] N. J. Sloane et al. "The on-line encyclopedia of integer sequences." *Published electronically at https://oeis. org*, 2018.

[127] C. Saldana, B. Herce, and B. Bickel. "More or less unnatural: Semantic similarity shapes the learnability and cross-linguistic distribution of unnatural syncretism in morphological paradigms." *Open Mind*, **6**, 2022, pp. 183–210.

[128] R. R. Noyer. "Features, positions and affixes in autonomous morphological structure." PhD thesis. Massachusetts Institute of Technology, 1992.

[129] M. Baerman. "Directionality and (un) natural classes in syncretism." *Language*, 2004, pp. 807–827.

[130] G. G. Corbett. "Morphosyntactic complexity: A typology of lexical splits." *Language*, 2015, pp. 145–193.

[131] M. Cysouw. *The paradigmatic structure of person marking*. OUP Oxford, 2009.

[132] K. Pertsova. *Learning form-meaning mappings in presence of homonymy: A linguistically motivated model of learning inflection*. University of California, Los Angeles, 2007.

[133] T. Johnson, K. Gao, K. Smith, H. Rabagliati, and J. Culbertson. "Investigating the effects of i-complexity and e-complexity on the learnability of morphological systems." *Journal of Language Modelling*, **9**(1), 2021, pp. 97–150.

[134] M. Maldonado and J. Culbertson. "Person of interest: Experimental investigations into the learnability of person systems." *Linguistic Inquiry*, 2020, pp. 1–42.

[135] K. Pertsova. "Grounding systematic syncretism in learning." *Linguistic Inquiry*, **42**(2), 2011, pp. 225–266.

[136] K. Pertsova. "Logical Complexity in Morphological Learning: effects of structure and null/overt affixation on learning paradigms." In: *Annual Meeting of the Berkeley Linguistics Society*. Vol. 38. 2012, pp. 401–413.

[137] A. Nevins. "Productivity and Portuguese morphology." *Romance Languages and Linguistic Theory 2013: Selected papers from'Going Romance'Amsterdam 2013*, **8**, 2015, p. 175.

[138] A. Nevins, C. Rodrigues, and K. Tang. "The rise and fall of the L-shaped morphome: diachronic and experimental studies." *Probus*, **27**(1), 2015, pp. 101–155.

[139] M. Fedzechkina, T. F. Jaeger, and E. L. Newport. "Language learners restructure their input to facilitate efficient communication." *Proceedings of the National Academy of Sciences*, **109**(44), 2012, pp. 17897–17902.

[140] J. M. Hupp, V. M. Sloutsky, and P. W. Culicover. "Evidence for a domain-general mechanism underlying the suffixation preference in language." *Language and Cognitive Processes*, **24**(6), 2009, pp. 876–909.

[141] M. Maldonado, C. Saldana, and J. Culbertson. "Learning biases in person-number linearization." *PsyArXiv Preprints*, (o. Nr.), 2020, pp. 163–176.

[142] A. Martin and J. Culbertson. "Revisiting the suffixing preference: Native-language affixation patterns influence perception of sequences." *Psychological Science*, **31**(9), 2020, pp. 1107–1116.

[143] M. Haspelmath. "On system pressure competing with economic motivation." In: *Competing Motivations in Grammar and Usage*. Ed. by B. MacWhinney, A. Malchukov, and E. Moravcsik. Oxford University Press, 2014, pp. 197–208.

[144] F. Ackerman and R. Malouf. "Morphological organization: The low conditional entropy conjecture." *Language*, 2013, pp. 429–464.

[145] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2026. URL: http://www.R-project.org/.

[146] H. Hammarström and R. Forkel. "Glottocodes: Identifiers Linking Families, Languages and Dialects to Comprehensive Reference Information." *Semantic Web Journal*, **13**(6), 2022, pp. 917–924. URL: https://content.iospress.com/articles/semantic-web/sw212843.

[147] A. Papafragou. "Spatial representations in language and thought." In: *ITRW on Experimental Linguistics*. 2006.

[148] P.-C. Bürkner. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software*, **80**(1), 2017, pp. 1–28. DOI: 10.18637/jss.v080.i01.

[149] M. Li and P. Vitanyi. "An introduction to kolmogorov complexity and its applications: Preface to the first edition." 1997.

[150] K. Ehret. "Kolmogorov complexity of morphs and constructions in English." *Linguistic Issues in Language Technology*, **11**, 2014, pp. 43–71.

[151] B. W. Fortson. *Indo-European language and culture: An introduction*. John Wiley & Sons, 2011.

[152] N. Zaslavsky, K. Garvin, C. Kemp, N. Tishby, and T. Regier. "The evolution of color naming reflects pressure for efficiency: Evidence from the recent past." *Journal of Language Evolution*, Apr. 2022. ISSN: 2058-458X. DOI: 10.1093/jole/lzac001. eprint: https://academic.oup.com/jole/advance-article-pdf/doi/10.1093/jole/lzac001/43361425/lzac001.pdf. URL: https://doi.org/10.1093/jole/lzac001.

[153] J. McWhorter. *Language interrupted: Signs of non-native acquisition in standard language grammars*. Oxford University Press on Demand, 2007.

[154] M. Ramscar, A. H. Smith, M. Dye, R. Futrell, P. Hendrix, H. Baayen, and R. Starr. "The 'universal'structure of name grammars and the impact of social engineering on the evolution of natural information systems." In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 35. 35. 2013.

[155] J. d. Pina-Cabral. "Names and Naming." In: *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, 2015, pp. 183–187. ISBN: 9780080970875. DOI: 10.1016/b978-0-08-097086-8.12213-5. URL: http://dx.doi.org/10.1016/B978-0-08-097086-8.12213-5.

[156] R. Alford. "Naming and identity: A cross-cultural study of personal naming practices." 1987.

[157] B. Schlücker and T. Ackermann. "The morphosyntax of proper names: An overview." *Folia linguistica*, **51**(2), 2017, pp. 309–339.

[158] E. Mensah and S. Mekamgoum. "The communicative significance of Ngêmbà personal names." *African Identities*, **15**(4), 2017, pp. 398–413.

[159] D. Moore. "The indexing of Welsh personal names." *The Indexer*, **17**(1), 1990, pp. 12–19.

[160] C. E. Shannon. "A Mathematical Theory of Communication." *Bell System Technical Journal*, **27**(3), 1948, pp. 379–423. DOI: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[161] M. Dye, P. Milin, R. Futrell, and M. Ramscar. "Alternative solutions to a language design problem: The role of adjectives and gender marking in efficient communication." *Topics in cognitive science*, **10**(1), 2018, pp. 209–224.

[162] R. Futrell and M. Hahn. "Information theory as a bridge between language function and language form." *Frontiers in Communication*, **7**, 2022, p. 657725.

[163] J. D. Jescheniak and W. J. Levelt. "Word frequency effects in speech production: Retrieval of syntactic information and of phonological form." *Journal of experimental psychology: learning, Memory, and cognition*, **20**(4), 1994, p. 824.

[164] S. Smith-Bannister. *Names and naming patterns in England, 1538-1700*. Oxford University Press, 1997.

[165] R. Hagan. "Traditional Australian Aboriginal naming processes." *Proof of Birth*, 2015, pp. 87–101.

[166] W. W. Snell. "Kinship relations in Machiguenga." PhD thesis. Hartford Seminary Foundation, 1964.

[167] D. F. Littlefield and L. E. Underhill. "Renaming the American Indian: 1890-1913." *American Studies*, **12**(2), 1971, pp. 33–45.

[168] G. Palsson. "Personal names: Embodiment, differentiation, exclusion, and belonging." *Science, Technology, & Human Values*, **39**(4), 2014, pp. 618–630.

[169] R. Du, Y. Yuan, J. Hwang, J. Mountain, and L. L. Cavalli-Sforza. "Chinese surnames and the genetic differences between north and south China." *Journal of Chinese Linguistics Monograph Series*, (5), 1992, pp. 1–93.

[170] J. Chen, L. Chen, Y. Liu, X. Li, Y. Yuan, and Y. Wang. "An index of Chinese surname distribution and its implications for population dynamics." *American journal of physical anthropology*, **169**(4), 2019, pp. 608–618.

[171] E. T. Jaynes. "Information theory and statistical mechanics." *Physical review*, **106**(4), 1957, p. 620.

[172] S. K. Baek, H. A. T. Kiet, and B. J. Kim. "Family name distributions: Master equation approach." *Physical Review E*, **76**(4), 2007, p. 046113.

[173] H. A. T. Kiet, S. K. Baek, H. Jeong, and B. J. Kim. "Korean family name distribution in the past." *Journal of the Korean Physical Society*, **51**(5), 2008, pp. 1812–1816.

[174] S. Lieberson and F. B. Lynn. "Popularity as taste." *Onoma*, **38**(0), 2003, pp. 235–276.

[175] A. Crook. *Personal naming patterns in Scotland, 1700-–1800: a comparative study of the parishes of Beith, Dingwall, Earlston, and Govan (MPhil Thesis)*. Scotland: Glasgow University, 2012.

[176] P. Ebrey. "The Chinese Family and the Spread of Confucian Values." In: *The East Asian Region: Confucian Heritage and Its Modern Adaptation*. Ed. by G. Rozman. Princeton: Princeton University Press, 1991, pp. 45–83.

[177] D. A. Galbi. *Names from Northumberland and Durham, 1530-1830*. https://www.galbithink.org/names/ncumb.txt, accessed on May 25, 2025. 2025.

[178] Korean Statistical Information Service. *Population by surname and clan (5 or more people) - Nationwide*. https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1IN15SD&conn_path=I2. 2015.

[179] Taiwanese Ministry of Interior. *Quan Guo Xing Ming Tong Ji Fen Xi (National statistics and analyses of personal names)*. https://www.ris.gov.tw/documents/data/5/2/107namestat.pdf. 2018.

[180] U.S. Census Bureau. *Frequently Occurring Surnames from the 2010 Census*. https://www.census.gov/topics/population/genealogy/data/2010_surnames.html. 2010.

[181] Z. Handel. "What is Sino-Tibetan? Snapshot of a Field and a Language Family in Flux." *Language and Linguistics Compass*, **2**(3), 2008, pp. 422–441.

[182] J. J. Song. *The Korean language: Structure, use and context*. Routledge, 2006.

[183] P. Sidwell and R. Blench. "14 the austroasiatic urheimat: the southeastern riverine hypothesis." *Dynamics of human diversity*, 2011, p. 315.

[184] E. Malmi, A. Gionis, and A. Solin. "Computationally inferred genealogical networks uncover long-term trends in assortative mating." In: *Proceedings of the 2018 World Wide Web Conference*. 2018, pp. 883–892.

[185] S. Kotilainen. "Who Sets the Norms of Naming?: Traditional Naming Practices in Agrarian Finland and Onomastic Legislation." *Onoma: Journal of the International Council of Onomastic Sciences*, (47), 2012, pp. 137–161.

[186] Social Security Administration. *Beyond the Top 1000 Names*. https://www.ssa.gov/oact/babynames/limits.html. 2023.

[187] P. Ohm. "Broken promises of privacy: Responding to the surprising failure of anonymization." *UCLA Law Review*, **57**, 2009, p. 1701.

[188] D. A. Galbi. "Long-term trends in personal given name frequencies in the UK." *Available at SSRN 366240*, 2002.

[189] G. Clark. *A farewell to alms: A brief economic history of the world*. Princeton University Press, 2007.

[190] J. Jefferies. "The UK population: past, present and future." In: *Focus on people and migration*. Ed. by R. Chappell, D. Pearce, F. Carlos-Bovagnet, and D. Till. Springer, 2005, pp. 1–17.

[191] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily. "Random effects structure for confirmatory hypothesis testing: Keep it maximal." *Journal of Memory and Language*, **68**(3), 2013, pp. 255–278.

[192] M. de Carvalho and F. J. Marques. "Jackknife euclidean likelihood-based inference for Spearman's rho." *North American Actuarial Journal*, **16**(4), 2012, pp. 487–492.

[193] M. Ramscar, P. Hendrix, C. Shaoul, P. Milin, and H. Baayen. "The myth of cognitive decline: Non-linear dynamics of lifelong learning." *Topics in cognitive science*, **6**(1), 2014, pp. 5–42.

[194] Z. Bin and C. Millward. "Personal names in Chinese." *Names*, **35**(1), 1987, pp. 8–21.

[195] Wikipedia. *List of members of the Chinese Academy of Sciences*. https://en.wikipedia.org/wiki/List_of_members_of_the_Chinese_Academy_of_Sciences, accessed on May 25, 2025. 2025.

[196] Suomalainen Tiedeakatemia. *Varsinaiset jäsenet*. https://acadsci.fi/jasenet/kotimaiset-jasenet/, accessed on May 25, 2025. 2025.

[197] Wikipedia. *Category:Members of the French Academy of Sciences*. https://en.wikipedia.org/wiki/Category:Members_of_the_French_Academy_of_Sciences, accessed on May 25, 2025. 2025.

[198] Daehan Mingug Hagsulwon. *Current members*. https://www.nas.go.kr/page/567dff03-b5f6-43e2-8131-bf67bdc9d530, accessed on May 25, 2025. 2025.

[199] Wikipedia. *List of members of the National Academy of Sciences*. https://en.wikipedia.org/wiki/List_of_members_of_the_National_Academy_of_Sciences, accessed on May 25, 2025. 2025.

[200] Rossiyskaya Akademiya Nauk. *Akademiki Rossiyskoy akademii nauk*. https://www.ras.ru/members/personalstaff/fullmembers.aspx, accessed on May 25, 2025. 2025.

[201] D. Dahan, J. S. Magnuson, M. K. Tanenhaus, and E. M. Hogan. "Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition." *Language and Cognitive Processes*, **16**(5-6), 2001, pp. 507–534.

[202] L. Zhaoping. "For your name's sake." *Current Biology*, **20**(8), 2010, R341.

[203] J. Qiu. "Scientific publishing: identity crisis." *Nature News*, **451**(7180), 2008, pp. 766–767.

[204] E. Rotenberg and A. Kushmerick. "The author challenge: Identification of self in the scholarly literature." *Cataloging & Classification Quarterly*, **49**(6), 2011, pp. 503–520.

[205] A. Strotmann and D. Zhao. "Author name disambiguation: What difference does it make in author-based citation analysis?" *Journal of the American Society for Information Science and Technology*, **63**(9), 2012, pp. 1820–1833.

[206] J. Goyes Vallejos. "What's in a name?" *Science*, **372**(6543), 2021, pp. 754–754. ISSN: 0036-8075. DOI: 10.1126/science.372.6543.754. eprint: https://science.sciencemag.org/content/372/6543/754.full.pdf. URL: https://science.sciencemag.org/content/372/6543/754.

[207] D. McKie. *What's in a Surname?: A Journey from Abercrombie to Zwicker*. Random House, 2014.

[208] G. P. Kelly. *From Vietnam to America: A chronicle of the Vietnamese immigration to the United States*. Routledge, 2019.

[209] D. S. Lauderdale and B. Kestenbaum. "Asian American ethnic identification by surname." *Population Research and Policy Review*, **19**(3), 2000, pp. 283–300.

[210] V. M. Taylor, T. T. Nguyen, H. Hoai Do, L. Li, and Y. Yasui. "Lessons learned from the application of a Vietnamese surname list for survey research." *Journal of immigrant and minority health*, **13**(2), 2011, pp. 345–351.

[211] U.S. Census Bureau. *Population*. https://www2.census.gov/programs-surveys/decennial/2020/data/apportionment/population-change-data-table.xlsx. 2020.

[212] U.S. Census Bureau. *U.S. Census Bureau Announces 2010 Census Population Counts Apportionment Counts Delivered to President*. https://www.census.gov/newsroom/releases/archives/2010_census/cb10-cn93.html. 2010.

[213] U.S. Census Bureau. *California*. https://www.census.gov/geographies/reference-files/2010/geo/state-local-geo-guides-2010/california.html. 2010.

[214] U.S. Census Bureau. *Delaware*. https://www.census.gov/geographies/reference-files/2010/geo/state-local-geo-guides-2010/delaware.html. 2010.

[215] U.S. Census Bureau. *The Asian population: 2010*. https://www.census.gov/content/dam/Census/library/publications/2012/dec/c2010br-11.pdf. 2010.

[216] Department of Household Registration. *Census files*. See "戶籍人口統計速報(9701)" https://www.ris.gov.tw/app/portal/346. 2018.

[217] Statistics Korea. *Complete Enumeration Results of the 2015 Population and Housing Census)*. https://kostat.go.kr/board.es?mid=a20107020000&bid=11739&tag=&act=view&list_no=420179&ref_bid=&keyField=&keyWord=&nPage=1. 2016.

[218] M. Ramscar, S. Chen, R. Futrell, and K. Mahowald. *Zenodo Repository of Cross-Cultural Structures of Personal Name Systems Reflect General Communicative Principles*. https://doi.org/10.5281/zenodo.13937788. 2025.

[219] Clause 40 Foundation. *Brief of Amicus Curiae in Support of Petitioner in Jackson v. United States*. U.S. Supreme Court, No. 22-6640. July 2023. URL: https://www.supremecourt.gov/DocketPDF/22/22-6389/272385/20230719143933658_22-6640%20tsac%20Clause%2040%20Foundation.pdf.

[220] Legal Information Institute. *Vagueness Doctrine*. Cornell Law School, 2024. URL: https://www.law.cornell.edu/wex/vagueness_doctrine (visited on 03/09/2026).

[221] C. Felsenfeld. "The plain English movement." *Can. Bus. LJ*, **6**, 1981, p. 408.

[222] E. Martínez, F. Mollica, and E. Gibson. "So much for plain language: An analysis of the accessibility of US federal laws over time." *Journal of Experimental Psychology: General*, **153**(5), 2024, p. 1153.

[223] C. Shain, I. A. Blank, E. Fedorenko, E. Gibson, and W. Schuler. "Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex." *Journal of Neuroscience*, **42**(39), 2022, pp. 7412–7430.

[224] E. Martínez, F. Mollica, and E. Gibson. "Even lawyers do not like legalese." *Proceedings of the national academy of sciences*, **120**(23), 2023, e2302672120.

[225] S. Pinker. *The sense of style: The thinking person's guide to writing in the 21st century*. Penguin Books, 2015.

[226] D. Mellinkoff. *The language of the law*. Wipf and Stock Publishers, 2004.

[227] E. Martínez, F. Mollica, and E. Gibson. "Even laypeople use legalese." *Proceedings of the National Academy of Sciences*, **121**(35), 2024, e2405564121.

[228] F. Mollica, A. Deloucas, E. Martínez, and E. Gibson. "The Ancient Origins of Modern Legalese: Content of Law." in prep.

[229] J. Huehnergard. *Key to a Grammar of Akkadian*. Brill, 2013.

[230] K. A. Adams. "A manual of style for contract drafting." In: American Bar Association. 2004.

[231] H. J. Berman. *Law and Revolution, I: The Formation of the Western Legal Tradition*. Harvard University Press, 1985.

[232] E. J. Bickerman and M. Smith. *The Ancient History of Western Civilization*. Harper & Row, 1976.

[233] D. Bodde. "Basic concepts of Chinese law: The genesis and evolution of legal thought in traditional China." *Proceedings of the American Philosophical Society*, **107**(5), 1963, pp. 375–398.

[234] J. W. Jones. *The Law and Legal Theory of the Greeks: An Introduction*. Oxford: Clarendon Press, 1956.

[235] S. Yang. *The Book of Lord Shang: A Classic of the Chinese School of Law*. Trans. by J. J. L. Duyvendak. London: Arthur Probsthain, 1928.

[236] A. Liu. *The Essential Huainanzi*. Ed. by J. S. Major, S. A. Queen, A. S. Meyer, and H. D. Roth. Translated, edited, and selected by John S. Major, Sarah A. Queen, Andrew Seth Meyer, and Harold D. Roth. New York: Columbia University Press, 2012.

[237] X. Yu. "Legal pragmatism in the People's Republic of China." *J. Chinese. L.*, **3**, 1989, p. 29.

[238] C. Wang and X. Zhang. *Introduction to Chinese Law*. Chapter: "Introduction: An Emerging Legal System" by Wang Chenguang. Hong Kong: Sweet & Maxwell, 1997.

[239] V. Behr. "Development of a new legal system in the People's Republic of China." *Louisiana Law Review*, **67**, 2006, p. 1161.

[240] T.-T. Hsia. "The language revolution in communist China." *Far Eastern Survey*, **25**(10), 1956, pp. 145–154.

[241] P. Wesley-Smith. *The Sources of Hong Kong Law: China's Nestorian Monument and Its Reception in the West, 1625-1916*. Vol. 1. Hong Kong University Press, 1994.

[242] T.-s. Wang. “The legal development of Taiwan in the 20th century: Toward a liberal and democratic country.” *Pac. Rim L. & Pol'y J.*, **11**, 2002, p. 531.

[243] E. Wilkinson. *Chinese History: a new manual*. Harvard University Press, 2015, p. 119.

[244] W. Johnson. *The T'ang Code, Volume I: General Principles*. Princeton University Press, Mar. 1979. ISBN: 9780691606323. DOI: 10.2307/j.ctvdtph1z. URL: http://dx.doi.org/10.2307/j.ctvdtph1z.

[245] 長孫無忌 et al. 故唐律疏議. Digitized copy from the Sibu Congkan (四部叢刊), hosted by Wikisource. 潘氏滂喜齋藏宋刊本 653. URL: https://zh.wikisource.org/wiki/File:Sibu_Congkan_Sanbian195-%E9%95%B7%E5%AD%AB%E7%84%A1%E5%BF%8C-%E6%95%85%E5%94%90%E5%BE%8B%E7%96%8F%E8%AD%B0-12-10.djvu (visited on 08/14/2026).

[246] M.-C. de Marneffe, C. D. Manning, J. Nivre, and D. Zeman. “Universal Dependencies.” *Computational Linguistics*, May 2021, pp. 1–54. ISSN: 1530-9312. DOI: 10.1162/coli_a_00402. URL: http://dx.doi.org/10.1162/coli_a_00402.

[247] D. Bates, M. Mächler, B. Bolker, and S. Walker. “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software*, **67**(1), 2015, pp. 1–48. DOI: 10.18637/jss.v067.i01.

[248] Y. K. Tse. 現代漢語歐化語法概論. 光明圖書公司, 1990.

[249] M. Cutts. “Unspeakable acts.” *English Today*, **9**(3), 1993, pp. 34–38.

[250] C. Xu. *History of the Constitution of the People's Republic of China*. Fujian People's Publishing House, 2003. ISBN: 978-7-211-04326-2.

[251] T. Zhu. “民法典編纂中的立法語言規範化.” 中國法學, **1**, 2017, pp. 230–248.

[252] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages.” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2020. URL: https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

[253] M. Davies. “Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English.” *Corpora*, **7**(2), 2012, pp. 121–157.

[254] C. H.-y. Chan. *Legal translation and bilingual law drafting in Hong Kong: Challenges and interactions in Chinese regions*. Routledge, 2020.

[255] D. C. Li and Z. P.-s. Luk. *Chinese-English contrastive grammar: An introduction*. Hong Kong University Press, 2017.

[256] D. Qian. “《唐律疏議》結構及書名辨析.” 史研究, (4), 2000, pp. 110–118.

[257] P. Keller. “Legislation in the People's Republic of China.” *U. Brit. Colum. L. Rev.*, **23**, 1988, p. 653.

[258] B. Zhang. “語體差異和語法規律.” 修辭學習, **2**, 2007, pp. 1–9.

[259] P. Brown and S. C. Levinson. “" Uphill" and" downhill" in Tzeltal.” *Journal of Linguistic Anthropology*, **3**(1), 1993, pp. 46–74.

[260] S. Kirby, M. Tamariz, H. Cornish, and K. Smith. "Compression and communication in the cultural evolution of linguistic structure." *Cognition*, **141**, 2015, pp. 87–102.

[261] C. Silvey, S. Kirby, and K. Smith. "Communication increases category structure and alignment only when combined with cultural transmission." *Journal of Memory and Language*, **109**, 2019, p. 104051.

[262] J. Aggarwal, J. Park, B. Beekhuizen, and S. Stevenson. "The Integrated Information Bottleneck: A theoretically-grounded model unifying category-and system-level structure in communicative efficiency." In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 48. 2026.

[263] S. Trott and B. Bergen. "Languages are efficient, but for whom?" *Cognition*, **225**, 2022, p. 105094.

[264] M. Zhan and R. P. Levy. "Comparing theories of speaker choice using a model of classifier production in Mandarin Chinese." In: Association for Computational Linguistics. 2018.

[265] V. S. Ferreira. "Ambiguity, accessibility, and a division of labor for communicative success." *Psychology of Learning and motivation*, **49**, 2008, pp. 209–246.

[266] M. Gimenes and B. New. "Worldlex: Twitter and blog word frequencies for 66 languages." *Behavior research methods*, **48**(3), 2016, pp. 963–972.

[267] P. Lison and J. Tiedemann. "Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles." *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 2016.

[268] N. Kirielle, P. Christen, and T. Ranbaduge. "Outlier detection based accurate geocoding of historical addresses." In: *Data Mining: 17th Australasian Conference, AusDM 2019, Adelaide, SA, Australia, December 2–5, 2019, Proceedings 17*. Springer. 2019, pp. 41–53.

[269] C. Xiao. "A Literature Review on Chinese Run-Ons." *Cross-Cultural Communication*, **17**(2), 2021, pp. 118–129.

[270] Wikisource. 唐律疏議. [Online; accessed 24-August-2025]. 2007. URL: https://zh.wikisource.org/wiki/%E5%94%90%E5%BE%8B%E7%96%8F%E8%AD%B0.

[271] H.-H. Chen. "The contextual analysis of Chinese sentences with punctuation marks." *Literary and linguistic computing*, **9**(4), 1994, pp. 281–289.

[272] M. Jin, M.-Y. Kim, D. Kim, and J.-H. Lee. "Segmentation of Chinese long sentences using commas." In: *Proceedings of the Third SIGHAN Workshop on Chinese Language Processing*. 2004, pp. 1–8.

[273] W. Wang and Z. Zhao. "On the Syntactic Categorization of Chinese Run-on Sentences." *Chinese Teaching in the World*, **31**(2), 2017.

[274] F. Wolf and E. Gibson. "Representing discourse coherence: A corpus-based study." *Computational linguistics*, **31**(2), 2005, pp. 249–287.

[275] R. Song, G. Shili, Y. Shang, and D. Lu. "面向文本信息處理的漢語句子和小句." 中文信息學報, **31**(2), Mar. 2017, pp. 18–24.

[276] W. Johnson. *The T'ang Code, Volume II: Specific Articles*. Princeton University Press, 1997. URL: http://www.jstor.org/stable/j.ctt7ztrgh (visited on 08/25/2025).

[277] Editorial Office for the Hanyu Da Cidian. *Gu Dai Han Yu Ci Dian (Classical Chinese Dictionary)*. Sichuan Ci Shu Chu Ban She, 2019.

[278] K. Yasuoka. "Universal dependencies treebank of the four books in Classical Chinese." In: *DADH2019: 10th International Conference of Digital Archives and Digital Humanities*. Digital Archives and Digital Humanities. 2019, pp. 20–28.

[279] R. Speer. *rspeer/wordfreq: v3.0*. Version v3.0.2. Sept. 2022. DOI: 10.5281/zenodo.7199437. URL: https://doi.org/10.5281/zenodo.7199437.