On the diverse language experiences of humans and machines

Naomi Tachikawa Shapiro

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2023

Reading Committee:
Shane Steinert-Threlkeld, Chair
Qi Cheng
Naja Ferjan Ramírez
Fei Xia

Program Authorized to Offer Degree:
Linguistics

University of Washington

Abstract

On the diverse language experiences of humans and machines

Naomi Tachikawa Shapiro

Chair of the Supervisory Committee:
Shane Steinert-Threlkeld
Department of Linguistics

Human experience is characterized by remarkable linguistic and sociocultural diversity. At the same time, much of this diversity is neglected in the language-related fields of linguistics, cognitive science, and natural language processing. This dissertation thus explores how diverse language experiences can lead to overlooked variation in humans and machines. Beginning with English-speaking families in the Pacific Northwest, Chapter 1 observes how children's language environments—in particular, differences in maternal and paternal input—can lead to variation in infant volubility, illustrating the delicate relationship between experience and behavior. Chapter 2 then devises an iconic approach to artificial language learning to investigate variation in cognitive biases across diverse language communities; importantly, seeking this variation directly can illuminate aspects of cognition that may be rooted in language experience rather than innate constraints. Finally, turning to machines, Chapter 3 shows how multilingual language models can encode many morphosyntactic properties crosslinguistically, but only occasionally uncover when a property is *shared* by multiple languages—beckoning the questions: To what extent does this variable behavior arise from crosslinguistic variation present in multilingual training data? And to what extent is it indeed humanlike? Together, these studies underscore the importance of minding diverse language experiences for understanding language processing in humans and machines.

To Kevin, my love:

These chapters are finally behind us

But every man is more than just himself; he also represents the unique, the very special and always significant and remarkable point at which the world's phenomena intersect, only once in this way and never again.

– Hermann Hesse, *Demian*

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would first like to thank my dissertation committee: Shane Steinert-Threlkeld has been a kind, patient, and thoughtful advisor and an equally brilliant collaborator. I have learned so much from him. I am profoundly grateful to Shane for the rigor he fosters in my work, for his encouragement to embrace my generalist ways, and for training me in the art of work-life balance and discerning which lemons are worth the squeeze. (I promise to keep practicing!) I am further grateful to Naja Ferjan Ramírez and Qi Cheng for their enthusiastic support and for immersing me in the discourses of language development and psycholinguistics. My research has also benefited enormously from the longtime support of Fei Xia. Last but not least, I would like to thank Noah A. Smith for being an engaging and thought-provoking Graduate School Representative.

I would like to thank my earliest mentor, Laura McGarrity, for introducing me to perspectives on language, for teaching me how to teach, and for her unwavering guidance and support these past 15 years. I would not have graduated college without her, nor believed that I held any promise in research. Laura is proof that great teaching and mentorship can change lives. I also wish to thank my longtime mentor and collaborator, Arto Anttila. Since my first quarter at Stanford, he has been in my corner, nourishing my passion for theory and phonology. The creativity Arto brings to research is inspiring and his joy infectious. I am also grateful and excited for my newly formed collaborations with Barbara "Basia" Citko and Andrew Hedding. Finally, I am grateful to Akira Omaki for taking a chance on me. His excitement for us to adventure into computational psycholinguistics meant the world to me. It was an honor to know him and to be his mentee, however briefly.

UW Linguistics has been my home away from home. I am extremely grateful to my colleagues behind the scenes who have kept the department running in tip-top shape: Joyce Parvi, Mike Furr, Misha Burgess, Monica Cohn, and Kyung Lim. In addition, I am grateful to Richard Wright for giving me the opportunity to revitalize the department's *proseminar*. I would also like to thank my wonderful mentees, Junyin Chen and Amelia Stockdill.

My research has further been supported by the Foreign Language & Area Studies Fellowship, the UW Presidential Dissertation Fellowship, the UW Linguistics Excellence in Research Graduate Award, and the UW Computing Research Club.

Academia would not be worth it without the friends we make along the way. I am forever grateful to Robert Xu and Ciyang Qing, who encouraged—*brainwashed?!*—me to pursue a PhD. I miss our late night chats and bike rides home together. I am also grateful to the extraordinary "friends slash colleagues" (Woo, 2019) I made at UW: Brent Woo, Ian Rigby, Molly FitzMorris, Amandalynne Paullada, Angie McMillan-Major, Rob Squizzero, Tsudoi Wada, Sara Ng, C.M. Downey, Trent Ukasick, Cassie Maz, Gita Dhungana, Haotian Zhu, Katie Lindekugel, Kaveri Sheth, Saiya Karamali, Adeline Braverman, Liz Conrad, Ray Gagné, Siyu Liang, Vipasha Bansal, Matt Kelley, Yadi Peng, and Preston Jiang. I am especially thankful to Amandalynne for inspiring me and for being a refuge throughout the PhD process.

I am deeply grateful to my friends of yore for keeping me grounded. Thank you to my dear friends Lucy Zhao, Nhi Nguyen, and Seoyoung Kil. Thank you to Jewel "Amanda" Bourne, Cat Fang, and Daniel Oh for their unconditional love and for always coaxing me out from under my rock. Thank you to the Emu and Elephant, to Chris Sundita, and to Kenny Mead.

I have had the good fortune of being in therapy for most of life. I am forever grateful to Cecile and Jessica. I especially wish to thank my current therapist, Jonathan, for challenging me to be happy, healthy, and well-rested; for broadening my perspectives on academia; and for helping me triumph over my *Pooh health Decision* (PhD).

My gratitude for my family could fill a second dissertation. But, I also love my family too much to write another dissertation, so these paragraphs will have to do:

I am grateful to my stepdad, Aki, for his quiet humor and for sharing his Raisin Wiches and love of travel with me. He has supported my every adventure and taken such good care of my mom (and the four-legged Tachikawas). I am grateful to my dad, Carl, for setting my mind alight with the wonderful world of words. My dad taught me how to write, edit, and analyze. He is my proofreader, sounding board, and "life fixer". I am proud to be his slow and deliberate daughter. I am grateful to mom, Julia, who shattered every glass ceiling ~~before~~ *for* me. She is my constant inspiration. My mom's otherworldly strength, her exuberance and wit, and her profound love meant that I was never afraid to chart my own path. I am grateful to my stepmom, Yolanda, for being a force of wisdom and stability; for loving me so much, she faints when I'm in pain; and for being an eternal source of pep talks and giggled swear words. I would not be a functional human without her.

It has been my privilege to be the youngest of four children. (Henceforth, siblings: *Please call me Dr. Shapiro.*) I am grateful to my sisters, Lena and Lauren. So much of who I am is thanks to them. Thank you, Lena, for tying balloons around my waist and for always guiding me back to my Golden Egg. Thank you, Lauren, for loving me fiercely and protecting me just as fiercely. I also

wish to thank my brother Ben and his wife Shaina for raising three monsters, whom I cherish. I almost forgive Abby, Izzy, and Jack for being taller than me #$@%!

I am grateful to my Grandma Sara for her love, for reminding me I'm not a delicate little prairie flower, and for taking too much pride in me. She's already telling anyone who'll listen that I've gotten a PhD. I am also grateful to and dearly miss my *akon*, Tomoki. He would have loved this too.

I am grateful to the Chens and extended clan for loving me and embracing me in their family. They have filled much needed reprieves from school with laughter, crafts, board games, yummy cooking, and fun floats down the Blackfoot River.

I am grateful to Chibi for being mine when I needed her most.

Lastly, to my husband, Kevin:

Thank you for liking my shoes.

Thank you for studying with me at Suzzallo.

Thank you for our neverending playdate since then.

I could not have done this dissertation without you.

Thank you for being my co-author in life.

I love you so much.

1 Introduction: Diversity in language experience

1.1 Juxtaposing humans and machines

Linguistic theory has long posited that human languages are far more similar than they are dissimilar, from possessing lexical categories (e.g., nouns and verbs); reflecting notions of plurality, tense, and subjects versus objects; to exhibiting hierarchical structure; and more. In turn, much psycholinguistics research has asked what drives these similarities: Are certain aspects of language innate and therefore universal (e.g., M. C. Baker, 2001; Chomsky, 1965, 1980; Fitch, Hauser, & Chomsky, 2005; Hauser, Chomsky, & Fitch, 2002)? Do crosslinguistic patterns stem from general cognitive pressures that constrain how we learn and use language (e.g., E. Bates & MacWhinney, 1982; Christiansen & Chater, 2008; Jackendoff & Pinker, 2005; Pinker & Jackendoff, 2005)?

In the time since I embarked on this PhD in 2017, advances in deep learning (e.g., Vaswani et al., 2017) and computing power have led to large language models (LLMs) that exhibit impressive language-like functionality, bringing new energy to the *nature vs. nurture* debate. Generative models like ChatGPT (OpenAI, 2022) are capable of producing seemingly fluent, novel utterances. Crucially, this is accomplished without the language-specialized genetic endowment that humans have long been hypothesized to possess. Their facility for language instead arises primarily from the statistics that the models have gleaned from raw text. LLMs have thus been argued to evidence a nurture-based pathway to language (e.g., Contreras Kallens, Kristensen-McLachlan, & Christiansen, 2023; Piantadosi, 2023), challenging Chomsky's "poverty of the stimulus" argument that languages cannot be learned without language-specific hardware (Chomsky, 1980; see Pearl, 2022, for review). These discussions, however, are in their early days, with recent literature commenting on the ecologically-invalid amounts of training data LLMs receive, among other important differences (e.g.,

Mahowald et al., 2023; Warstadt & Bowman, 2022; Yedetore, Linzen, Frank, & McCoy, 2023).

Harking back to Hockett's "design features" of language, whatever genetic endowment we may have for language, it is widely understood that we cannot acquire languages without *cultural transmission* (Hockett, 1959). In other words, **language experience**—for both humans and machines—is paramount. Identifying what aspects of language are biological or architectural in origin requires us also to seek out *experiential* factors that shape language learning and structure. This dissertations aids this endeavor by investigating how diverse language experiences can lead to variation in linguistic behaviors in both humans and machines. In turn, I ask: What does this variation say about the nature of human and machine language processing?

For humans, language experience is deeply characterized by linguistic and sociocultural diversity. In addition to the typological variation that exists across languages, individuals within the same language community will often vary in their social contexts and their exposure to other languages, correlating with crosslinguistic, language-internal, and intra-speaker variation (cf. Boland, Kaan, Valdés Kroff, & Wulff, 2016). Despite this diversity, a pervasive problem across language-related fields is that they have focused predominantly on English and the experiences of a small subset of English speakers (Blasi, Henrich, Adamou, Kemmerer, & Majid, 2022; Bylund, Khafif, & Berghoff, 2023; Kidd & Garcia, 2022; Majid, 2023; *inter alia*), amounting to an *Anglocentric bias* (Blasi, Henrich, et al., 2022). Research on child language acquisition, for instance, has mainly sampled children who belong to speakers of English and other "well-studied" Indo-European languages, particularly within wealthy countries (Kidd & Garcia, 2022). Within this narrow demographic, studies on child language environments have largely focused on maternal contributions, overlooking the distinct and complex, context-dependent roles that caregivers can play in shaping children's linguistic development (cf. Pancsofar, 2020; Tamis-LeMonda, Baumwell, & Cabrera, 2012).

Research in psycholinguistics has further assumed that "the paradigmatic language user [is] an unrealistically invariant monolingual" (Boland et al., 2016). Yet, it is widely believed that the majority of the world's population is familiar with two or more languages (e.g., Grosjean, 1982, 2008, 2010; Marian & Shook, 2012; Romaine, 2013; Trask, 1999), exhibiting diverse learning trajectories, patterns of language use, and proficiency profiles (Byers-Heinlein et al., 2019) that blur the lines between monolingualism, bilingualism, and multilingualism (cf. Cheng et al., 2021; Gullifer et al., 2021; Luk & Bialystok, 2013; Ortega, 2020). Further complicating this picture, language users will

often vary in the processing of their second (L2), third (L3), and L*n* languages as a function of their prior language experience (cf. Lago, Mosca, & Stutter Garcia, 2021). Recent work has even begun to problematize the concept of "native speakers", which can overemphasize monolingual experiences while perpetuating oversimplified, normative assumptions about the experiences, behaviors, and identities of language users (Cheng et al., 2021; see also Weissler et al., 2023, for discussion). In sum, uncountably diverse human experiences lead to a multidimensional continuum of sociocultural and linguistic profiles that the cognitive sciences oft neglect (Blasi, Henrich, et al., 2022; Levinson, 2012; Majid, 2023).

Similar operating assumptions have run amok in natural language processing (NLP) research. Numerous studies have called attention to the field's overblown focus on English and other "high-resource" languages, detailing subsequent harms and inequities in the technology arena (Bender, Gebru, McMillan-Major, & Shmitchell, 2021; Blasi, Anastasopoulos, & Neubig, 2022; Joshi, Santy, Budhiraja, Bali, & Choudhury, 2020; Ranathunga & de Silva, 2022; Søgaard, 2022; *inter alia*). Compounding this issue is the fact that languages are often treated as monolithic entities in NLP. If psycholinguistics glosses over language-internal variation, NLP research altogether erases it—such that it will collapse varietal, idiolectal, and sociolinguistic variation into a single dataset and label it "English" (when it doesn't otherwise leave this variation out)—raising the question, who's language experiences do LLMs model? And in what social contexts? To recast this comparison, a *single* artificial agent is tasked with modeling *countless* human agents, and will inevitably fail for a subset of these individuals due to imbalanced representation in the training data.

Furthermore, while many over the years have pushed back against the alleged "language agnosticism" of machine learning approaches (Bender, 2011; Joshi et al., 2020), recent human-machine comparisons easily regress: Notably, the cutting-edge LLMs like ChatGPT that have been invoked to challenge psycholinguistic theory are primarily English-based models "gorging on hundreds of terabytes of data" (Chomsky, Roberts, & Watumull, 2023) that are available to few languages. Crucially, these comparisons often ignore non-English and multilingual experiences as they attempt to weigh in on human cognition.

Taken together, a consequence of this myopia is that we risk overgeneralizing observations from a socioculturally narrow sliver of "English speakers' behaviors, brains, and cognition to our entire species" (Blasi, Henrich, et al., 2022; see also Levinson, 2012, and Majid, 2023)—in other words,

mistaking nurture for nature. Likewise, we risk citing English-based LLMs as evidence for or against hypotheses of a human-universal language endowment, when we have yet to establish that (i) the parallels hold for other languages and (ii) that the models handle variation and multilingualism in humanlike ways. Accordingly, this dissertation takes a step back from juxtaposing humans and machines, tracing the diverse influences of language experience on their individual behaviors. My hope is that, in doing so, it will help recalibrate these respective discourses and human-machine comparisons to be more mindful of diverse language experiences.

1.2 Overview: Language experience across learners and contexts

The remainder of this introduction provides an overview of the chapters that follow. Each chapter portrays an agent (human or machine) learning language (real or artificial), while exploring the impacts of language experience on the learner's behavior. Beginning with humans then turning to machines, I explore how overlooked variation in experience can drive variation in behavior. Each chapter is self-contained and can be read on its own, stapled together as vignettes of language experience. Since the studies are the fruit of interdisciplinary collaborations, I will gradually shift to using *we*—also to acknowledge that this dissertation is entering the public domain, where together we can discuss, debate, and advance our collective understanding of language.

1.2.1 Human agents learning human languages

Chapter 2 illustrates the delicate relationship between human language experience and linguistic behavior by zooming in on familial factors as sources of variability in child language learning. Focusing on English-speaking families in the U.S. Pacific Northwest, we explore how differences in maternal and paternal language can lead to variation in infant volubility ("chattiness"). Notably, fathers have long been underrepresented and their distinct roles under-considered in the child development literature. This chapter devotes special attention to paternal usage of *parentese*, a near-universal style of infant-directed speech that is distinguished by its higher pitch, slower tempo, and exaggerated intonation (Ferjan Ramírez, Lytle, Fish, & Kuhl, 2018; Ferjan Ramírez, Lytle, & Kuhl, 2020; Golinkoff, Can, Soderstrom, & Hirsh-Pasek, 2015; Kuhl et al., 1997; Kuhl, Tsao, & Liu, 2003; H. M. Liu, Kuhl, & Tsao, 2003; Ramírez-Esparza, García-Sierra, & Kuhl, 2014, 2017a,

2017b; Singh, Nestor, Parikh, & Yull, 2009; Song, Demuth, & Morgan, 2010; Thiessen, Hill, & Saffran, 2005).

We show that the infants in our sample of mother-father families were exposed to substantially fewer words and less parentese from their fathers than their mothers. Even so, an interesting asymmetry emerged where maternal word counts and paternal parentese predicted child vocalizations, but paternal word counts and maternal parentese did not. Our findings verify that, while infants may hear less input from their fathers in some families, paternal parentese still plays a unique role in shaping their linguistic development. The chapter concludes by calling for future work to study child language environments within more linguistically and socioculturally diverse families. Such work should also seek to explain parental gender differences in terms of family dynamics and beliefs surrounding family roles in child development. This study was first published in the *Journal of Speech, Language, and Hearing Research* in collaboration with Naja Ferjan Ramírez and Daniel Hippe (Shapiro, Hippe, & Ferjan Ramírez, 2021). With the journal's permission, it is reproduced unaltered in Chapter 2, along with its abstract.

1.2.2 Human agents learning artificial languages

Chapter 3 begins by introducing the *artificial language learning* (ALL) methodology, wherein participants are taught miniature languages in controlled experiments. In particular, we review the strengths and weaknesses of using ALL to test for universal cognitive biases. We discuss how *crosslinguistic influence*—transfer effects from participants' language experiences—can lead to overlooked explanations for biases observed in ALL experiments, especially given the minimal language representation in this body of literature. We suggest this conundrum may be due in part to limitations of artificial languages as they're traditionally conceived.

Accordingly, we propose a novel iconicity-based paradigm that replaces phonologically-grounded lexemes with pictographic stimuli in artificial languages. Crucially, iconic artificial languages may be more accessible to understudied populations, in addition to making it easier to re-use the same stimuli with different communities in controlled crosslinguistic comparisons. They may therefore enable a more holistic understanding of human cognition by exposing which behaviors originate from innate biases versus crosslinguistic influence.

We validate our approach by using pictographic lexemes to reproduce previously attested word and

morpheme ordering effects. In Experiments 1 and 2, we verify that English speakers prefer syntactic and morphological structures that are *scope-isomorphic*, mirroring the semantic scope relations of their individual parts (Culbertson & Adger, 2014; Martin, Holtz, Abels, Adger, & Culbertson, 2020; Martin, Ratitamkul, Abels, Adger, & Culbertson, 2019; Saldana, Oseki, & Culbertson, 2021). Next, the chapter foreshadows our ongoing work with Polish speakers (Experiment 3), where we attempt to tease apart biological and experiential explanations for scope-isomorphic biases by teaching the same iconic artificial language to speakers of different languages. These experiments illustrate how iconic ALL can be used to study crosslinguistic influence and test the universality of hypothesized cognitive constraints. Shane Steinert-Threlkeld and I presented Experiment 1 at *CogSci 2023* (Shapiro & Steinert-Threlkeld, 2023), while Experiments 2 and 3 reflect our joint work with Qi Cheng and Barbara Citko.

1.2.3 Artificial agents learning human languages

Returning to machines, Chapter 4 describes deep learning and LLMs in greater detail. It then spotlights *multilingual* LLMs, setting the stage for the rest of the chapter, which explores variable *crosslingual* functionality given the challenges posed by linguistic diversity in multilingual training data. In particular, we explore how a multilingual model can vary in its encodings of crosslinguistic morphosyntactic properties.

Following prior work to "probe" LLMs for humanlike linguistic representations (e.g., Bacon & Regier, 2019; Chi, Hewitt, & Manning, 2020; Conneau, Kruszewski, Lample, Barrault, & Baroni, 2018; Futrell & Levy, 2019; Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018; Hewitt & Manning, 2019; Hupkes, Veldhoen, & Zuidema, 2018; Jawahar, Sagot, & Seddah, 2019; N. F. Liu, Gardner, Belinkov, Peters, & Smith, 2019; Marvin & Linzen, 2018; Tenney, Das, & Pavlick, 2019; Zhang & Bowman, 2018), we put forward a multilabel probing task to analyze multilingual LLMs, which we demonstrate with the model Multilingual BERT (*mBERT*; Devlin, Chang, Lee, & Toutanova Kristina, 2019). We train probes for seven typologically diverse languages of varying morphological complexity (Afrikaans, Croatian, Finnish, Hebrew, Korean, Spanish, and Turkish), then evaluate the probes on six held-out languages (Arabic, Chinese, Marathi, Slovenian, Tagalog, and Yorùbá). This style of probing has the benefit of revealing which crosslinguistic properties a language model recognizes as being shared by multiple languages.

Overall, we find that mBERT recognizes many morphosyntactic features in different languages, but only occasionally uncovers when a property is shared by those languages (e.g., connecting pronouns in the related languages Hebrew and Arabic, but not nominative case in Turkish and Spanish). Importantly, this inconsistent crosslingualism likely emerges from variation in how the morphosyntactic features surface crosslinguistically, as reflected in the model's training experience. Chapter 4 concludes by considering the implications of these findings for multilingual LLMs and, briefly, for *human* language processing. The chapter adapts joint work with Amandalynne Paullada and Shane Steinert-Threlkeld, published in *Findings of the ACL: EMNLP 2021* (Shapiro, Paullada, & Steinert-Threlkeld, 2021).

1.3 Summary

This dissertation takes a step back from human-machine comparisons, posing the simple question: How do behaviors reflect prior language experience? Across learners and contexts, I relate variation in language experience to variation in behavior. I begin with a microcosm of this intricate relationship, shown with infants among their families. As the dissertation progresses, it delves further into the relationship between experience and behavior with adult and machine language learners, increasingly complicating the *nature vs. nurture* debate in often overlooked ways. By the end, I hope to reinforce the importance of accounting for diverse language experiences in the cognitive and NLP sciences and, moreover, for the juxtaposition of humans and machines.

2 How chatty are daddies? An exploratory study of infants' language environments

Abstract

Purpose: Fathers play a critical but underresearched role in their children's cognitive and linguistic development. Focusing on two-parent families with a mother and a father, the present longitudinal study explores the amount of paternal input infants hear during the first 2 years of life, how this input changes over time, and how it relates to child volubility. We devote special attention to parentese, a near-universal style of infant-directed speech, distinguished by its higher pitch, slower tempo, and exaggerated intonation. **Method:** We examined the daylong recordings of the same 23 infants at ages 6, 10, 14, 18, and 24 months, given English-speaking families. The infants were recorded in the presence of their parents (mother-father dyads), who were predominantly White and ranged from mid to high socioeconomic status (SES). We analyzed the effects of parent gender and child age on adult word counts and parentese, as well as the effects of maternal and paternal word counts and parentese on child vocalizations. **Results:** On average, the infants were exposed to 46.8% fewer words and 51.9% less parentese from fathers than from mothers, even though paternal parentese grew at a 2.8-times faster rate as the infants aged. An asymmetry emerged where maternal word counts and paternal parentese predicted child vocalizations, but paternal word counts and maternal parentese did not. **Conclusions:** While infants may hear less input from their fathers than their mothers in predominantly White, mid-to-high SES, English-speaking households, paternal parentese still plays a unique role in their linguistic development. Future research on sources of variability in child language outcomes should thus control for parental differences since parents' language can differ substantially and differentially predict child language.

2.1 Introduction

Sociocultural frameworks have long emphasized child development as a socially mediated process, in which caregivers scaffold their children's cognitive and linguistic development through social interactions (e.g., Bruner, 1981; Kuhl, 2007, 2011; Snow, 1977, 1999; Vygotsky, 1978). While research on parental language within these frameworks has largely focused on maternal contributions, emerging studies have highlighted the invaluable roles that fathers play in their children's linguistic development (for reviews, see Pancsofar, 2020; Tamis-LeMonda, Baumwell, & Cabrera, 2012). This work is set against an evolving backdrop, as family structures diversify, more women pursue careers, and fathers become more directly involved in family life and childcare (Cabrera, Tamis-LeMonda, Bradley, Hofferth, & Lamb, 2000; Cabrera, Volling, & Barr, 2018; Jones & Mosher, 2013). Controlling for maternal input and demographic factors, research has begun to chart fathers' language input during early childhood, revealing its unique associations with children's concurrent and subsequent language skills (C. E. Baker & Vernon-Feagans, 2015; Conica, Nixon, & Quigley, 2020; Majorano, Rainieri, & Corsano, 2013; Malin, Cabrera, & Rowe, 2014; Pancsofar & Vernon-Feagans, 2006; Pancsofar, Vernon-Feagans, & The Family Life Project Investigators, 2010; Quigley & Nixon, 2020; Reynolds, Vernon-Feagans, Bratsch-Hines, Baker, & the Family Life Project Key Investigators, 2019; Tamis-LeMonda, Baumwell, & Cristofaro, 2012). In the present longitudinal study, we continue this endeavor, tracing exposure to paternal parentese during the first 2 years of life (i.e., infancy) to better understand sources of variability in children's language outcomes.

Several studies have connected paternal input *quality* during infancy to child language skills in later years. For example, fathers'use of metalingual talk and repetitions of children's utterances at 24 months have both been tied to children's vocabulary skills at 48 months and beyond (Conica et al., 2020; Malin et al., 2014). Likewise, paternal usage of *wh*-questions at 24 months is positively associated with concurrent child vocabulary and verbal reasoning skills at 36 months (Rowe, Leech, & Cabrera, 2017). At the same time, fathers are widely reputed to differ *quantitatively* from mothers, with many studies suggesting that fathers talk less overall (Golinkoff & Ames, 1979; Hladik & Edwards, 1984; Leaper, Anderson, & Sanders, 1998; Majorano et al., 2013; Pancsofar & Vernon-Feagans, 2006). For instance, in a study of Italian families during 20-min triadic free-play sessions, Majorano et al. (2013) found that fathers' but not mothers' noun frequency at 15 months

predicted child language production and comprehension at 30 months, even though mothers had produced more words, greater vocabulary diversity, and longer utterances. In similar settings, Pancsofar and Vernon-Feagans (2006) found that fathers' but not mothers' vocabulary diversity at 24 months predicted children's expressive language skills at 36 months, even though fathers had produced fewer utterances, word types, and *wh*-questions, and took shorter conversational turns. These studies demonstrate the complex and unique associations between paternal and child language, even when fathers provide less input than mothers.

2.1.1 Daylong audio recordings

Research using Language ENvironment Analysis (LENA) technology has added new dimensions to quantitative comparisons between mothers' and fathers' language input. LENA's pocket-sized recording devices are wearable by infants and facilitate daylong snapshots of their natural environments. These recordings offer a more ecologically valid glimpse of parent-child interactions and at a scale that exceeds traditional observations in a laboratory or from brief visits to infants' homes (Christakis et al., 2009; Oller et al., 2010; Xu, Richards, & Gilkerson, 2014; Zimmerman et al., 2009). In addition, LENA's proprietary software segments and classifies speech, tabulating volubility measures such as adult word counts. Recent work using LENA has further pointed to significant disparities between mothers and fathers in the amount of language input they provide. Gilkerson and Richards (2009) reported that mothers accounted for 75% of the total adult words spoken in their semilongitudinal study of children between 2 and 48 months of age. Pairing LENA's automatic measures with manual coding of child-directed speech (CDS), Bergelson et al. (2018) similarly found that infants hear 2–3 times more CDS from women than from men.

However, neither Gilkerson and Richards nor Bergelson et al. address how family dynamics may have contributed to the disparities they observed between maternal and paternal speech. In both studies, it is unclear whether the recordings came from single-parent or two-parent households and, in the latter families, whether the parents were of the same or different gender, or whether both parents were present during the recordings (e.g., a parent could have been away at work). Moreover, neither study connected parental differences to child language outcomes. Notably, Gilkerson and Richards found that parents' word counts *overall* predicted child vocalizations ("talkative parents have talkative children", p. 21; see also Hart & Risley, 1995), but did not consider how mothers and

fathers may individually contribute to this effect.

2.1.2 Parentese

Bergelson et al.'s (2018) finding that men produce less CDS motivates an interesting avenue of study when we consider the wealth of research that has shown different *registers* of CDS to vary in their impact on child language learning. Specifically, infants favor parentese, an acoustically exaggerated style of CDS that benefits infants' concurrent and subsequent language skills (Ferjan Ramírez et al., 2018, 2020; Golinkoff et al., 2015; Kuhl et al., 1997, 2003; H. M. Liu et al., 2003; Ramírez-Esparza et al., 2014, 2017a, 2017b; Singh et al., 2009; Song et al., 2010; Thiessen et al., 2005). Parentese is distinct from adult-directed speech and "standard/adult" registers of infant-directed speech (Farran, Lee, Yoo, & Oller, 2016; Ramírez-Esparza et al., 2014) in terms of its simplified lexicon and syntax, slower tempo, and melodic intonation contours (Fernald, 1985; Fernald & Kuhl, 1987; Genovese et al., 2020). The contributions of parentese to child language development are rooted in these characteristics, which evoke social responses from infants and enhance parent-child interactions (Golinkoff et al., 2015; Tartter, 1980). Accordingly, multiple studies have tied parentese to infant vocal activity, such as babbling (Ferjan Ramírez et al., 2018; Ramírez-Esparza et al., 2014), word production (Ferjan Ramírez et al., 2020; Ramírez-Esparza et al., 2014), and conversational turns (Ferjan Ramírez et al., 2020). In general, adult vocalizations that are higher in pitch and amplitude are more likely to be followed by infant vocalizations (Ritwika et al., 2020).

Despite parentese formerly being called *motherese*, fathers produce parentese cross-linguistically (Broesch & Bryant, 2018; Quigley, Nixon, & Lawson, 2019; see also Saint-Georges et al., 2013), though they exhibit some prosodic differences from mothers (Fernald et al., 1989; Gergely, Faragó, Galambos, & Topál, 2017; Warren-Leubecker & Bohannon III, 1984) Nevertheless, the majority of the work on associations between parentese and child language has either not distinguished maternal and paternal parentese in their analyses (Ferjan Ramírez et al., 2018, 2020; Ramírez-Esparza et al., 2014, 2017a, 2017b) or has focused exclusively on mothers (Kuhl et al., 1997; H. M. Liu et al., 2003). It thus remains unknown how mothers and fathers differ in the amount of parentese they produce and how their parentese might differentially relate to infant vocalizations and child language learning during the first 2 years of life and beyond.

2.1.3 The present study

In the present exploratory study, we seek to contrast paternal and maternal input, posing the following questions: How much paternal input, especially parentese, do infants hear during the first 2 years of life and, relatedly, how might this input change throughout infancy? Moreover, how might the input of mothers and fathers differ in their associations to infant vocalizations?

We analyzed previously collected longitudinal data from the same group of 23 infants at ages 6, 10, 14, 18, and 24 months. All of the infants came from predominantly White, English-speaking families and were raised by mother-father parents. Using LENA technology, naturalistic daylong audio recordings were obtained from the 23 families at each age, during times when both parents were asked to be home with their child. This allowed us to control for parental disparities that could arise from a parent being absent during a recording (e.g., if one parent was away at work). In our analysis, we focused on three response variables: the total number of words heard by infants (Adult Word Count [AWC]), the amount of parentese they heard, and the number of linguistic vocalizations they produced (Child Vocalization Count [CVC]). Both AWC and CVC are measures of volubility ("chattiness"). In addition to looking at the effects of child age and parent gender (i.e., mother vs. father) on these variables, we also controlled for socioeconomic status (SES), which has been shown time and again to predict child language learning (for review, see Rowe, 2018). The participating families ranged from mid to high SES.

While past research has explored the benefits of parentese and the prosodic differences displayed by mothers and fathers, our study is, to our knowledge, the first to compare the amount of maternal and paternal parentese infants hear, and to do so longitudinally. Likewise, our study is the first to relate adult volubility and parentese to child volubility while simultaneously examining parental differences. As this analysis was exploratory, our only hypothesis was that the input from mothers and fathers would vary from one another, both synchronically and across infancy. Our primary goal was to study sources of variability in children's language environments and to see how this variation might relate to child volubility. More broadly, a thorough understanding of infants' language environments can inform theories of language acquisition, shape family-centered policy, and identify circumstances that might benefit from intervention.

2.2 Method

2.2.1 Participants and data collection

We analyzed daylong recordings collected from the same 23 infants at ages 6, 10, 14, 18, and 24 months. The participating families were part of the control group of a larger longitudinal study on parent-infant verbal interactions (see Ferjan Ramírez et al., 2018, 2020). The original study recruited 79 English-speaking families, of which 55 families participated in a parent coaching intervention and 24 families served as the "no treatment" control group. Out of the 24 families, we excluded one single-parent household from our analysis, leaving 23 families to constitute our present dataset on parental differences. The original study recruited the families in the greater Seattle area via the University of Washington Subjects Pool. All of the parents provided informed written consent. The study and its experimental procedures were approved by the Institute Review Board of the University of Washington and conformed to the U.S. Federal Policy for the Protection of Human Subjects.

The families were recruited when the participating infants were 5 months of age; each infant was born full-term (± 14 days of due date), of normal birth weight (6–10 lbs.), and without birth or postnatal complications. The parents of the 23 infants were all mother-father dyads. According to demographic data collected prior to the audio recordings, 12 of the infants were girls and 11 were boys. The families ranged from mid to high SES, as measured by the widely used Hollingshead Index (Hollingshead, 1975, 2011), a composite SES score (range: 8–66) based on parent education, occupational prestige, family income, and related factors. On the Hollingshead scale, the families fell between 30 (e.g., both parents had high school diplomas and worked in sales or construction) and 66 (e.g., both parents had advanced degrees and worked as engineers or attorneys; M = 49.5, SD = 10.9). Twenty-one of the infants were White, one was of unknown race, and one was of mixed race. All of the parents spoke English varieties standard to the U.S. Pacific Northwest.

The daylong recordings were collected between October 1, 2016, and August 5, 2018. The collection timepoints were set as close as possible to each infant's 6-, 10-, 14-, 18-, and 24-month birthdays (on average, within 3 days of the date). These timepoints were initially selected to parallel milestones in child language development (i.e., babbling, transition to first words, individual words, transition to word combinations, and combinatorial speech). At each timepoint, the infants were recorded over two consecutive weekend days, when both parents were home and not working.

Variable	Type	6 months	10 months	14 months	18 months	24 months
AWC	LENA	16,621.0 ± 7,605.6	15,380.3 ± 7,782.6	15,467.0 ± 7,416.3	16,164.3 ± 6,297.0	16,674.1 ± 6,425.7
FAN	LENA	10,956.7 ± 5,517.6	10,473.2 ± 5,681.7	9,589.8 ± 5,129.0	9,966.3 ± 4,691.0	10,217.1 ± 4,190.3
MAN	LENA	5,664.3 ± 3,230.7	4,907.1 ± 3,427.9	5,877.2 ± 4,016.6	6,198.0 ± 3,352.1	6,457.0 ± 4,198.2
% Parentese	manual	44.6 ± 18.7	46.1 ± 20.7	52.4 ± 20.7	58.5 ± 26.0	66.9 ± 21.6
% *M.* parentese	manual	33.3 ± 15.5	35.7 ± 18.2	38.4 ± 18.8	40.7 ± 22.2	45.7 ± 21.4
% *P.* parentese	manual	14.9 ± 12.7	14.3 ± 11.9	18.7 ± 12.5	23.2 ± 17.1	30.3 ± 16.5
Proportion of *M.* input containing parentese	manual	0.50 ± 0.18	0.50 ± 0.21	0.57 ± 0.22	0.60 ± 0.22	0.67 ± 0.20
Proportion of *P.* input containing parentese	manual	0.30 ± 0.21	0.34 ± 0.21	0.40 ± 0.19	0.45 ± 0.27	0.57 ± 0.22
CVC	LENA	1,177.9 ± 393.9	1,270.0 ± 472.0	1,146.5 ± 441.3	1,639.6 ± 585.0	2,604.2 ± 1,165.5

Table 2.1: Speech variables and their distributions (M ± SD) when infants were 6, 10, 14, 18, and 24 months old. *Note:* AWC = adult word count; FAN = female adult nearby words; MAN = male adult nearby words; CVC = child vocalization count; *M.* = maternal; *P.* = paternal; LENA = Language ENvironment Analysis estimate; manual = manually coded.

Parents were instructed to start each recording in the morning when their child awoke, to go about their day as usual, then to turn off the recorder at night when the child went to sleep. Throughout the day, the infants wore the lightweight LENA device inside the front pocket of a specially designed vest. The average duration of the daylong recordings was 12.8 hr (range: 8.7–16); recording lengths did not differ significantly between the five data collection timepoints (p = .312).

2.2.2 Key variables

The key variables in our analysis and their distributions are summarized in Table 2.1. Parent and child speech were quantified through a combination of automatic annotation by LENA software and manual (human) annotation. LENA's acoustic modeling software supplies various estimates of child speech and exposure to adult speech (cf. Gilkerson & Richards, 2020). Regarding the accuracy of these estimates, recent efforts have sought to assess and validate LENA's classification performance (Bulgarelli & Bergelson, 2020; Cristia, Bulgarelli, & Bergelson, 2020; Cristia et al., 2021; Lehet, Arjmandi, Houston, & Dilley, 2020; Wang, Williams, Dilley, & Houston, 2020). According to one meta-analysis, LENA achieves a mean recall and precision of 0.59 and 0.68, respectively, for

recognizing adult words and a mean recall of 0.77 for recognizing child vocalizations (Cristia et al., 2020). Such validation studies demonstrate that LENA is a useful tool for studying infants' language environments, but one that should be supplemented by manually quantified measures—as we do in the present study with parentese.

We drew on several automatic metrics from LENA: Adult speech was measured in AWC, the estimated number of adult words spoken near the infant, whether child-directed or adult-directed. The LENA Advanced Data EXtractor tool subdivides AWC into words spoken by women and those spoken by men, what they term "female adult nearby" (FAN) words and "male adult nearby" (MAN) words. We used FAN and MAN to approximate maternal and paternal word counts, respectively. Infant vocal activity was measured in CVC, the estimated number of segments that contain meaningful child speech (excluding nonspeech signals like cries and vegetative sounds). Child vocalization segments can be of any length, as long as they are surrounded by 300+ ms of nonspeech. These variables—AWC, FAN, MAN, and CVC—are all considered measures of volubility and, as such, do not index the quality of utterances. Each variable was measured over the length of each daylong recording, then averaged across each timepoint, producing a single estimate per child at each age.

Following Ramírez-Esparza et al. (2014, 2017a, 2017b), we supplemented LENA's estimates with manual annotations of parentese. We segmented the daylong recordings into 30-s intervals, then selected the 50 intervals with the highest AWC from each recording day. This yielded 100 30-s segments per family at each age. Past studies have shown that 30-s clips of ambient sounds provide sufficient information for characterizing observed behaviors (Mehl, Gosling, & Pennebaker, 2006; Orena, Byers-Heinlein, & Polka, 2019; Ramírez-Esparza, Mehl, Álvarez-Bermúdez, & Pennebaker, 2009). To collect a broad range of environments, we further required that the selected intervals be spaced at least 3 min apart. Ten research assistants then manually annotated the selected segments for three binary variables: (i) the presence/absence of *any* parentese, (ii) the presence/absence of *maternal* parentese, and (iii) the presence/absence of *paternal* parentese. Note that any interval could contain both maternal and paternal parentese. The annotators identified parentese by its higher pitch and wider pitch range, showing high intercoder agreement (0.99 intraclass correlation). Finally, for each infant at each timepoint, "% parentese" was quantified as the percentage of intervals that contained parentese, intended to reflect the proportion of parental input that is parentese. Percent

maternal parentese and % paternal parentese were likewise quantified. For readability, we will refer to "% parentese" as parentese, "% maternal parentese" as maternal parentese, and "% paternal parentese" as paternal parentese.

2.2.3 Statistical analysis

We evaluated associations of child age, SES, and parent gender with the outcomes AWC and parentese, using multivariable linear mixed-effects regression. Both AWC and parentese were log-transformed to reduce right-skewness. Child age and SES were included as continuous covariates, while parent gender was treated as a binary variable. We also included random intercepts per subject and parent nested within subject to account for the repeated measures at each age. To assess how adult linguistic input might vary by parent with child age, we added an interaction term between child age and parent gender to the main effects models for both AWC and parentese.

We again used linear mixed-effects regression to analyze CVC and its associations with child age, SES, FAN, MAN, maternal parentese, and paternal parentese. We log-transformed CVC, FAN, and MAN to reduce right-skewness and included random intercepts per subject to account for the repeated measures. FAN, MAN, and maternal and paternal parentese were incorporated as continuous covariates. We subsequently added interaction terms between child age and each of the four adult speech variables to explore how their associations with CVC might vary with child age.

For our analyses, we used the statistical computing language R (version 3.6.2; R Core Team, 2013) and, in particular, the lme4 package (D. Bates, Mächler, Bolker, & Walker, 2015) to fit the linear regression models. With the help of the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017), we conducted two-sided tests to determine statistical significance, defining the threshold for significance as $\alpha = .05$.

2.3 Results

2.3.1 Adult words and parentese

On average, infants in the present sample heard 16,061.4 adult words per day: 10,240.6 words from women and 5,794.8 words from men. In relation to AWC, we found a significant main effect of parent gender ($p < .001$), such that infants heard on average 46.8% fewer words from men than from

	Outcome: AWC			Outcome: % Parentese		
Main effect	%Δ	95% CI	*p*	%Δ	95% CI	*p*
SES (per 1-point increase)	0.13	(-1.26, 1.54)	0.860	0.86	(-0.84, 2.60)	0.334
Child age (per 1-month increase)	0.42	(-0.30, 1.13)	0.255	3.22	(2.21, 4.24)	<0.001
Parent gender (Male)	-46.78	(-56.45, -34.97)	<0.001	-51.92	(-63.22, -37.14)	<0.001
Interaction						
Child age × Parent gender (Male)	1.16	(-0.26, 2.61)	0.112	3.10	(1.15, 5.09)	0.002

Table 2.2: Associations with adult word count (AWC) and % parentese. *Note:* AWC = adult word count; %Δ = mean change in outcome per unit increase in variable; CI = confidence interval; SES = socioeconomic status. The main effect terms are from the main effects models and the interaction terms are from the interaction models.

women. However, AWC was not significantly associated with SES ($p = .860$), child age ($p = .255$), or with an interaction between child age and parent gender ($p = .112$). Table 2.2 summarizes the results for both AWC and parentese.

Across all of the timepoints, 53.7% of the manually coded intervals on average contained parentese, 38.8% included maternal parentese, and 20.3% included paternal parentese. (Recall that an interval could contain parentese from both parents; hence, maternal and paternal parentese do not sum to total parentese, 53.7%.) We found that parentese was significantly associated with parent gender ($p < .001$). Based on the model that included only main effects, fathers produced on average 51.9% less parentese than mothers. While there was no main effect of SES ($p = .334$), we did find a significant main effect of child age ($p < .001$) and, furthermore, that child age interacted significantly with parent gender ($p = .002$), as depicted in Figure 2.1. In particular, maternal parentese increased by 1.7% each month, whereas paternal parentese increased by 4.8% each month. At 6 months of age, infants heard on average 62.8% less parentese from fathers than from mothers, but, by 24 months, they heard only 35.5% less parentese from fathers.

Since our results show that mothers produced significantly more words than fathers in our sample, a potential concern is that the present dataset favors maternal parentese. By only annotating % parentese in the 30-s segments with the highest AWC (see Ferjan Ramírez et al., 2018, 2020), the original study could have incidentally biased the selection towards segments that contain more female words, contriving or inflating the gap between maternal and paternal parentese. In a post hoc analysis, we thus looked at the relative proportions of each parent's input that contained

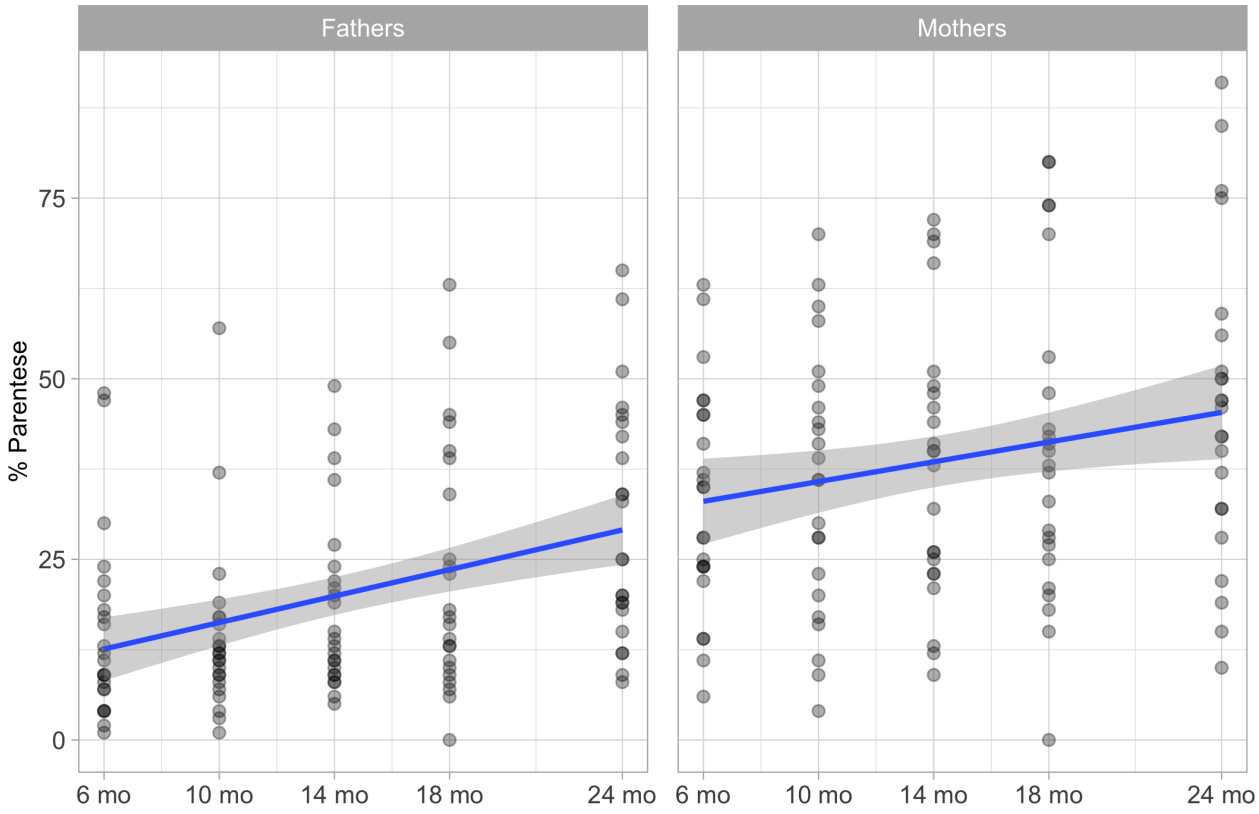


Figure 2.1: The percentage of coded input containing paternal and maternal parentese for the same 23 infants at ages 6, 10, 14, 18, and 24 months. The blue regression lines reflect linear predictions of % paternal and maternal parentese and the gray bands 95% confidence intervals. Note that, at 18 months, there were two separate households in which one parent did not produce any parentese in the coded segments (one father and one mother).

parentese—hereafter, their "relative parentese proportions." We did so by identifying which of the coded segments contained maternal speech and which contained paternal speech, then calculated their respective proportions of parentese. From this analysis, we excluded segments that additionally contained speech from a third (i.e., nonparent) adult to avoid how this might affect parents' usage of parentese. The distributions of these proportions are included in Table 2.1. Mixed-effects linear regression revealed that parentese constituted a significantly smaller proportion of fathers' input than mothers' input (41.0% vs. 56.6%, on average; $p < .001$), as shown in Table 2.3. We discuss the implications of these results in the Discussion.

Main effect	%Δ	95% CI	p
SES (per 1-point increase)	0.03	(-0.39, 0.46)	0.873
Child age (per 1-month increase)	0.84	(0.66, 1.02)	<0.001
Parent gender (Male)	-10.06	(-14.21, -5.70)	<0.001
Interaction			
Child age × Parent gender (Male)	0.42	(0.06, 0.78)	0.024

Table 2.3: Associations with relative parentese proportion. *Note:* %Δ = mean change in relative parentese proportion per unit increase in variable; CI = confidence interval; SES = socioeconomic status. The main effect terms are from the main effects model and the interaction term is from the interaction model.

2.3.2 Child vocalizations

The infants in our study produced on average 1,567.6 vocalizations per day. As relayed in Table 2.4, CVC was significantly associated with child age ($p < .001$) but not with SES ($p = .959$). While we found no main effect of MAN ($p = .275$), CVC was significantly associated with FAN ($p < .001$), with child vocalizations increasing by 4.8% on average per 10% increase in FAN. Conversely, CVC was significantly associated with paternal parentese ($p = .023$), but not with maternal parentese ($p = .313$), with child vocalizations increasing by 0.8% on average per 1.0% increase in fathers' usage of parentese in the coded intervals. In the interaction model, child age and FAN also interacted significantly ($p = .004$). At 6 months of age, infant vocalizations increased on average by 2.1% per 10% increase in FAN; however, by 24 months, this rate had grown to 8.9%. On the other hand, CVC was not significantly associated with interactions between child age and MAN ($p = .924$), paternal parentese ($p = .358$), or maternal parentese ($p = .356$).

2.4 Discussion

This study examines mother and father differences in parental language input, focusing on the number of adult words and parentese heard by infants in English-speaking households. We quantified these measures through a combination of automatic and manual annotation. To our knowledge, this is the first study to perform a longitudinal comparison of the amount of maternal and paternal parentese infants hear during the first 2 years of life, as well as the first to probe how parental input differences

Main effect	%Δ	95% CI	*p*
SES (per 1-point increase)	-0.03	(-0.99, 0.94)	0.959
Child age (per 1-month increase)	3.74	(2.69, 4.84)	<0.001
FAN (per 10% increase)	4.78	(2.83, 6.69)	< .001
MAN (per 10% increase)	-0.87	(-2.33, 0.72)	.275
% Maternal parentese (per 1% increase)	-0.26	(-0.74, 0.22)	.313
% Paternal parentese (per 1% increase)	0.76	(0.12, 1.38)	.023
Interaction			
Child age × FAN	0.39	(0.14, 0.65)	.004
Child age × MAN	-0.01	(-0.21, 0.19)	.924
Child age × Maternal parentese	-0.03	(-0.08, 0.03)	.356
Child age × Paternal parentese	0.04	(-0.04, 0.11)	.358

Table 2.4: Associations with child vocalization count (CVC). *Note:* %Δ = mean change in CVC per unit increase in variable; CI = confidence interval; SES = socioeconomic status; FAN = female adult nearby words; MAN = male adult nearby words. The main effect terms are from the main effects model and the interaction terms are from the interaction model.

relate to child volubility. Using LENA technology, we analyzed the daylong recordings of 23 infants with their mothers and fathers at ages 6, 10, 14, 18 and 24 months. The parents were predominantly White and ranged from mid to high SES.

After controlling for SES and asking families to record on weekends when both parents were present, we found that children heard significantly more words and parentese from mothers than from fathers. This gap persisted throughout infancy, even as fathers increased their usage of parentese over time at a faster rate than mothers. In a follow-up analysis, we analyzed relative parentese proportions to allay the concern that the reported parentese gap fell out of coding parentese in the intervals that had the highest AWCs. Mirroring our % parentese findings, parentese constituted a significantly smaller proportion of paternal input than maternal input, with fathers' relative parentese proportions increasing over time at a faster rate than mothers' relative parentese proportions. These trends suggest that the gap in infants' paternal parentese exposure does not merely stem from fathers producing fewer words in the coded intervals or overall, and instead reflect genuine differences in parents' usage of parentese.

We additionally found that maternal word counts and paternal parentese predicted child vocalizations, with the relation between child vocalizations and maternal word counts strengthening over

time. Interestingly, paternal word counts and maternal parentese did not predict child vocalizations. These asymmetries exemplify how the language of mothers and fathers can differentially relate to child language. Building on earlier work, we thus show that fathers play a unique role in children's linguistic development and deserve further study in order to better understand sources of variability in child language outcomes. We now turn to more detailed discussions of parental input and child volubility.

2.4.1 Parental input

We approximated parental word counts with LENA's AWC estimates. As expected, we found no apparent effects of child age on parent volubility (cf. Gilkerson et al., 2017). The infants in our sample heard on average 46.8% fewer words from fathers than from mothers, consistent with prior literature that has attested significant gaps between total maternal and paternal input (Gilkerson & Richards, 2009; Golinkoff & Ames, 1979; Hladik & Edwards, 1984; Leaper et al., 1998; Majorano et al., 2013; Pancsofar & Vernon-Feagans, 2006).

All of the fathers in the present study produced at least some parentese. In the analyzed coded segments, paternal parentese constituted on average 41.0% of all paternal input and 20.3% of children's language exposure. Nonetheless, we found that infants heard on average 51.9% less parentese from fathers than from mothers. This finding most resembles that of Bergelson et al. (2018), who observed that women produced 2-3 times more CDS than men (including both standard CDS and parentese). One possible explanation for these trends is that fathers took on comparatively fewer caregiver responsibilities, leading to fewer direct interactions with their infants (cf. Cabrera et al., 2000; Lamb & Tamis-LeMonda, 2004; Pleck, 2010). Concomitantly, these trends may be explained by differing beliefs regarding child language development. For instance, in a survey of 180 female and 120 male undergraduates in the Midwest, Kennison and Byrd-Craven (2015) found that men were significantly less likely to believe that infant-directed speech was beneficial to infants' development. Moreover, significantly more men reported that using "baby talk" had been discouraged in their families. Future studies that compare maternal and paternal input should also collect data that probes parents' beliefs and attitudes surrounding childcare responsibilities and child language development.

Our analysis further revealed that both parents increased their usage of parentese over time,

with fathers doing so at a faster rate. This may reflect a trend of parents, especially fathers, talking more to their infants as their children became more socially active with age. This analysis is partially supported by our finding that paternal parentese predicts CVC, as discussed in the following subsection. Nevertheless, while the fathers in our sample increased their parentese at a 2.8-times faster rate than mothers, mothers still produced substantially more parentese overall. As was visualized in Figure 2.1, the amount of parentese that fathers produced when the infants were 24 months old was, on average, the same amount that mothers had produced when the infants were 6 months old. Our finding that exposure to parentese increases over time, however, does diverge from Bergelson et al. (2018), who did not encounter any age-related effects for CDS with infants from 3 to 20 months of age. A possible explanation for this divergence is that their cross-sectional sample (i.e., recordings subsampled from infants of different ages) obscured relative increases in CDS over time that our longitudinal dataset captures.

It is also notable that gaps arose between maternal and paternal input even though both parents were asked to be present at home when the recordings in our study were collected. In addition to differing beliefs and familial responsibilities, contrasting workforce obligations may diminish or intensify these gaps during the work week—something worthy of explicit investigation in the future.

2.4.2 Child volubility

Consistent with prior work, child vocalizations in our sample increased with child age (cf. Gilkerson & Richards, 2009; Gilkerson et al., 2017). When looking at child vocalizations and its associations with parental input, asymmetries emerged between maternal and paternal speech. In particular, we found maternal total words to predict child vocalizations at each age, with this association strengthening over time. However, we found no such effects of paternal word counts. In contrast, and perhaps most surprisingly, we found that paternal parentese predicted child vocalizations, but maternal parentese did not.

These effects appear to reflect a quantity-quality distinction, where the sheer quantity of maternal input relates more to child volubility, whereas the quality of paternal input is more relevant than its quantity. What we are thus seeing is that mothers' volubility predicts child volubility. This dovetails with a threshold/saliency analysis of the parentese asymmetry, resembling the "threshold effect" postulated by Pancsofar and Vernon-Feagans (2006): It is possible that the mothers in the present

sample all provided ample verbal input, including parentese, to the point where maternal parentese was not predictive of child volubility; alternatively, their extensive usage of parentese could have prompted infants to monitor and model maternal volubility. In contrast, since paternal word counts and parentese were less likely to exceed such a threshold or to match maternal input, this allowed paternal parentese to become more salient and differentially relate to child vocalizations.

Paternal parentese and child volubility may have also been mutually reinforcing. When the fathers engaged with their children in parentese, it may have been especially salient, prompting more social and vocal responses from the infants. Likewise, when the infants were more socially and vocally interactive, this may have spurred paternal engagement, including the use of parentese. To infer causality, future studies should examine maternal and paternal parentese in the context of conversational turns and temporally contingent language (cf. Pretzer, Lopez, Walle, & Warlaumont, 2019). If paternal parentese does *lead* to child volubility as a result of its saliency, these transactions may be partially enabled by the contrast provided by high maternal volubility. Specifically, it is possible that the quality of paternal speech was salient not just because fathers provided input below some threshold, but because the mothers in our sample had also created rich communicative environments. Future research can delve into the asymmetries displayed by parents, focusing on the relationships between the quantity, quality, and saliency of parental input.

It is also worth noting that, even though maternal parentese did not predict child volubility in the present study, it has still been tied to many aspects of infant language development, such as phonetic learning (Kuhl et al., 1997; H. M. Liu et al., 2003) and vocabulary acquisition (Hartman, Ratner, & Newman, 2017; Kalashnikova & Burnham, 2018; Newman, Rowe, & Bernstein Ratner, 2016). At the same time, much of the work in this area has excluded fathers and has not controlled for parental differences. The asymmetries discussed here underscore the importance of distinguishing maternal and paternal input, as their effects on child language do not necessarily parallel one another. Paternal parentese is thus worthy of deeper investigation, while controlling for maternal language, and vice versa.

2.4.3 Additional future directions

First and foremost, efforts that seek to relate caregivers' language to child language development should control for parental differences, since the quantity and quality of parents' language can differ

substantially and differentially predict child language outcomes. As our analysis was exploratory, future work should continue to compare maternal and paternal parentese and how they relate to child volubility, as well as to more fine-grained child language measures (e.g., babbling, vocabulary diversity, utterance complexity). These relationships should also be explored during the first 6 months of life to further illuminate the trajectory and role of parentese during infancy. Related research can also assess the impact of family-centered interventions (cf. Bagner, 2013; Bagner & Eyberg, 2003) that target closing the gap between mothers' and fathers' usage of parentese.

With regards to LENA, future endeavors to compare parental input in daylong recordings should attempt to track interlocutors and activities, at least within a subset of segments. This is especially important, since many laboratory-based studies have shown mothers' and fathers' language to vary across contexts (e.g., dyadic and triadic interactions, as in Bingham, Kwon, & Jeon, 2013) and activities (e.g., shared book reading, as in Malin et al., 2014). This would enable a more ecologically valid inspection of how maternal and paternal input varies across events. When manually coding daylong recordings, we also recommend that future efforts focus on segments from a wider range of environments (e.g., those that contain the highest FAN, MAN, and CVC values), especially when quantifying maternal and paternal speech. In addition, it would be interesting to explore nonbinary annotations of parentese that are more sensitive to variation in infants' linguistic environments.

Relatedly, we used LENA's adult volubility measures to approximate parent speech (akin to Gilkerson & Richards, 2009). Future studies should validate what proportions of these word counts can be attributed to mothers and fathers, versus other individuals, in addition to evaluating their general accuracy. Such validation is vital, given the growing body of work that has found LENA to make systematic errors as a function of speaker gender (Bergelson et al., 2018; Bulgarelli & Bergelson, 2020; Cristia et al., 2020, 2021; Lehet et al., 2020). For example, Bergelson et al. (2018) noted that LENA was more likely to mislabel male speakers as female when they were using CDS and, conversely, female speakers as male when they were addressing adults. Similarly, studies have found that LENA has a harder time identifying female speakers when their speech is infant directed (Lehet et al., 2020), but, when recognizing female speech versus male speech overall, it exhibits greater precision (0.60 vs. 0.43) and comparably low recall (0.32 vs. 0.31 Cristia et al., 2021). These findings also emphasize the importance of supplementing automatic measurements with manual variables, as we have done in the present study with parentese. For instance, despite

the shortcomings of LENA's gender-specific tags, it is promising that our automatic and manual analyses yielded such similar results: that infants heard 46.8% fewer words and 51.9% less parentese from fathers.

With respect to SES, in none of our analyses did an effect of SES appear, contrary to previous literature that has linked lower SES to less parental input (e.g., Gilkerson et al., 2017; Hart & Risley, 1995; Hoff, 2003a, 2003b; Rowe, 2008). However, our sample did not include low SES families, who may exhibit more pronounced variation in parental speech. The lack of differentiation between the mid and high SES families could also reflect a genuine social shift towards the merger of parental language behaviors across certain SES groups (cf. Gilkerson et al., 2017). Future comparisons of parental language should sample families from a broader range of socioeconomic backgrounds. It would also be interesting to see how different dimensions of SES (e.g., education and income; cf. Rowe, 2018) might vary by parent in their associations to parentese.

Finally, it is important to acknowledge that the infants in our sample were each raised by predominantly White, English-speaking mothers and fathers. This narrow demographic may exhibit different patterns of language input compared to families who are not represented in the current study. Future research should address parental parentese differences amongst more diverse populations, such as non-White families, multilingual and non-English-speaking households, single-parent families, and families with same-sex parents. In particular, while we believe fathers in mother-father households have been largely sidelined by past research and deserve greater attention, this is even more true of LGBTQ+ parents. The language development literature on these families is virtually non-existent. While parent gender did predict parental input in the present study, future work that relates parental language to child language should also explore to what extent gender differences can be explained by family dynamics, differing responsibilities, and the diversity of beliefs surrounding family roles and child development. Future efforts might also consider modeling parent gender as a variable that is multifaceted, rather than binary (Cameron & Stinson, 2019).

2.5 Conclusion

This exploratory study has illustrated how maternal and paternal language can differentially predict infant volubility, highlighting the need to control for both parents when relating parental input to child language outcomes. In our sample of English-speaking families, maternal word counts and paternal

parentese predicted infant vocalizations during the first 2 years of life, whereas paternal word counts and maternal parentese did not. These asymmetries emerged even as infants heard considerably fewer words and less parentese from fathers. Quantifying these patterns is an important step towards better understanding paternal contributions to child language development. The observed paternal parentese gap also presents an opportunity to design culturally sensitive interventions that enhance father-infant interactions.

Acknowledgements

We thank Patricia K. Kuhl, Denise Padden, Julia Mizrahi, and Bo Woo for their valuable assistance during the data collection in the original intervention study. We are also grateful to Kevin Chen for his extensive feedback on the present article. Both the intervention study and current analysis were supported by the Overdeck Family Foundation and the University of Washington's Language Acquisition and Multilingualism Endowment.

3 Iconic artificial language learning: Distinguishing universality from transfer effects

3.1 Introduction

Artificial languages have emerged as an insightful tool for probing human language learning and processing. In the artificial language learning (ALL) paradigm, participants are taught miniature constructed languages in controlled settings to reveal properties of the human linguistic-cognitive system. Such investigations have explored topics fundamental to language and cognition, including first and subsequent language acquisition, communicative efficiency, statistical learning, hypothesized cognitive constraints on language learning, the origin of crosslinguistic patterns, and the complicated relationships therein. (For acquisition-focused reviews, see Culbertson & Schuler, 2019; Gómez & Gerken, 2000; Grey, 2020; Morgan-Short, 2020; for reviews on constraints and learning biases, see Culbertson, 2012, 2023; Fedzechkina, Newport, & Florian Jaeger, 2016; Folia, Uddén, De Vries, Forkstam, & Petersson, 2010.)

At the same time, these efforts have largely centered on speakers of English and other Indo-European languages—particularly members of Western, educated, industrialized, rich, and democratic (WEIRD) societies (Henrich, 2020; Henrich, Heine, & Norenzayan, 2010a, 2010b)—putting them at risk of overgeneralizing ALL findings to understudied communities or humans *en masse* (cf. Blasi, Henrich, et al., 2022; Majid, 2023; see also Evans & Levinson, 2009, and Levinson, 2012). That is where this chapter takes its cue: We introduce a novel *iconic* approach to ALL to help broaden the representation of diverse language communities in investigations of the human endowment for language.

3.1.1 Artificial language learning and language diversity

Researchers have gravitated towards ALL as a tool for detecting cognitive constraints that may be universal. When learners of an artificial language display patterns of preference that are consistent with typological asymmetries, it could signal that such biases, if innate, have shaped the evolution and structure of languages (Culbertson, 2012, 2023; Fedzechkina et al., 2016). The particular explanatory power of ALL comes from the access it grants to behaviors we might not otherwise be able to observe minus confounds. Compared to studying second language learners or language learning in the wild, ALL offers better experimental control over study variables, such as prior exposure to or familiarity with the target language, the precise features of the target language, and the amount of training input (overcoming "data sparsity").

Still, an ever-present concern in these investigations, especially in the pursuit of universal biases, is that participants may transfer their existing linguistic knowledge to the task of learning the artificial language. This is known as *crosslinguistic influence*: the ways in which language learners are influenced by their prior language experience (a phenomenon more actively studied in applied linguistics; for reviews, see Jarvis & Pavlenko, 2008; Lago et al., 2021). In ALL studies, unwanted transfer effects can confound result interpretations by making it unclear whether an observed behavior is genetically driven or reflects participants' codified experience.

For example, ALL studies and adjacent work have grappled with the origin of the prefixing-suffixing asymmetry (e.g., Bruening, Brooks, Alfieri, Kempe, & Dabašinskienė, 2012; Hupp, Sloutsky, & Culicover, 2009; Martin & Culbertson, 2020; St. Clair, Monaghan, & Ramscar, 2009), the observation that most languages favor suffixing over prefixing as a word-formation strategy (Bybee, Pagliuca, & Perkins, 1990; Cutler, Hawkins, & Gilligan, 1985; Greenberg, 1957; Hall, 1988; Hawkins & Cutler, 1988; Hawkins & Gilligan, 1988). Hupp et al. (2009) proposed that this pattern may be due to a universal, domain-general perceptual bias that privileges the beginning of sequences (see also Hawkins & Cutler, 1988), spurring languages to place content morphemes in this position of salience. As experimental evidence, Hupp et al. showed that English speakers maintain a sequence-initial bias when given strings of meaningless nonces, musical notes, and geometric shapes. However, the authors cautioned that these findings were also compatible with a perceptual bias driven by language experience, since English is a predominantly suffixing language.[1] In subsequent

[1] Levinson (2012) also foreshadowed this sampling issue: "So the variation is there, regimented by culture both

work, Martin and Culbertson (2020) showed that speakers of Kîîtharaka, a heavily-prefixing Bantu language, privilege sequence-final positions, calling into question whether suffixing preferences can be explained by a universal bias.[2]

In this regard, seeking out transfer effects can be deeply informative about the human endowment for language. In the context of ALL, the most compelling evidence of a universal bias would come from populations whose language experience does not predispose them to prefer or *disprefer* the phenomenon in question: Note that the prefixing-suffixing asymmetry might well be fed by a universal perceptual bias for sequence-initial positions, but that language experience has heightened this bias for speakers of suffixing languages (e.g., English) and reduced the bias for speakers of prefixing languages (e.g., Kîîtharaka; see Martin & Culbertson, 2020, for discussion). This ambiguity underscores the importance of juxtaposing typologically diverse communities for teasing apart effects that are innate versus experiential in origin.

Despite the epistemological (and ethical) imperative for broader language and community representation in the cognitive sciences (Blasi, Henrich, et al., 2022; Majid, 2023), ALL studies are typically limited in the populations they analyze. The choice of participants often comes down to convenience sampling, where researchers design studies for familiar or readily accessible language populations, tailoring the artificial languages to those communities. With the exception of "silent gesture" studies (e.g., Futrell et al., 2015; Goldin-Meadow, So, Özyürek, & Mylander, 2008; Schouwstra & de Swart, 2014), which do not involve researcher-designed languages, ALL experiments traditionally employ artificial languages that are grounded in the primary phonotactic and orthographic systems of the participating speakers, making them easier for the participants to learn but also limiting the usability of the stimuli across a diverse speaker pool. ALL studies, as a consequence, often focus on a single monolingual population (typically English speakers), putting them at risk of overlooking how participants' existing linguistic knowledge may bear on their hypotheses or modulate their findings.

To summarize, in the pursuit of explanations for crosslinguistic patterns, ALL studies have focused

within populations and across populations, and instead of treating it like noise we should be treating it like a major source of insight, in the same way that geneticists use mutations or knock-out mice. For example, suppose you have a theory of language acquisition that predicts that learning prefixes is harder than learning suffixes: Then you can study infants learning a sample of Mayan languages that naturally titrate all the possible affix orders (Pye, Pfeiler, De León, Brown, & Mateo, 2007)".

[2]Martin and Culbertson (2020) also pointed out that, up until their study with Kîîtharaka speakers, all experimental work on the prefixing-suffixing asymmetry had focused on English speakers.

on identifying universal biases that may shape language learning and structure. Constrained by methodological limitations and driven by a focus on what makes languages *similar*, these studies can overlook linguistic diversity and thus variation in how phenomena surface crosslinguistically. This may prove especially problematic for hypothesized cognitive universals and accounts of a language-specific endowment. However, by employing the ALL methodology in broader crosslinguistic investigations (i.e., experiments involving multiple language communities) and by actively seeking out transfer effects, we can arrive at a more holistic understanding of how language users both resemble and diverge from one another in their language processing and, in doing so, shed a clearer light on the aspects of cognition shared across humans.

3.1.2 The present study

In this study, we set out to design linguistic stimuli that endeavor towards the impossible goal of "language neutrality". In particular, we propose rooting artificial languages in *iconic lexicons*, replacing phonologically-realized stimuli with miniature pictographic systems, thereby enabling ALL scholars to work with a greater diversity of language populations. In principle, the same language neutral lexicon could be taught to participants from typologically diverse linguistic communities, facilitating crosslinguistic analyses and, by extension, the study of transfer effects with finer experimental control. This, in turn, could help broaden language representation in the ALL literature. Compared to written and auditory stimuli, pictographic languages may prove more accessible to children and to signing and Indigenous communities, in addition to enhancing experimental control when studying bi/multilingual populations.

To date, iconicity and non-orthographic symbols have been leveraged to a limited degree in ALL work.[3] Notably, in a recent artificial sign language experiment, Sato, Schouwstra, Flaherty, and Kirby (2020) showed that iconic gestures helped participants learn form-meaning mappings compared to non-iconic gestures.[4] Turning to the written modality, Saratsli, Bartell, and Papafragou (2020)

[3]It is important to clarify that considerable work has used non-orthographic symbols in artificial *grammar* learning experiments, where participants are taught grammars over *semantic-less* forms. For instance, Martin and Culbertson (2020) and Culbertson and Kirby (2022) used geometric shapes with no associated meanings to study possible domain-general biases impacting morpheme order.

[4]Relatedly, some research has argued that iconicity is an essential ingredient for language development and word learning, providing scaffolding that enables learners to link embodied experience to linguistic form and, ultimately, to a symbolic system that is mostly arbitrary. For discussion, see Perniss, Thompson, and Vigliocco (2010) and Lupyan and Winter (2018).

evidenced preferential marking of certain sources of information over others, even when substituting a nonce evidentiality morpheme with an arbitrary symbol (a black-filled circle). Bowerman and Smith (2022) also used iconic stimuli in an iterated communication game to study semantic extension. Finally, though not conceived of as an ALL study, Dautriche, Buccola, Berthet, Fagot, and Chemla (2022) taught simple iconic symbols to *Papio papio* baboons, finding evidence that baboons can entertain compositional representations, namely, negation structures.

To our knowledge, the present ALL study is the first to utilize a fully iconic artificial language in the written modality (with *humans*, that is). We test the viability of iconic stimuli by using miniature pictographic languages to replicate prior ALL findings pertaining to the linear order of nominal modifiers (Experiment 1; Culbertson & Adger, 2014; Martin et al., 2020, 2019) and inflectional morphemes (Experiment 2; Saldana et al., 2021). These studies focused on a hypothesized universal bias for semantic-mirroring (i.e., *scope-isomorphic*) structures, which we introduce further in the next section. Crucially, we use scope-isomorphism as an apparatus for reinforcing the importance of seeking out transfer effects when assessing the universality of cognitive constraints.

We conduct Experiments 1 and 2 with English speakers to replicate the scope-isomorphic preferences that were attested with English speakers in the aforementioned literature, using the earlier findings as a baseline.[5] These replications represent a proof-of-concept that iconic stimuli can be used successfully at multiple linguistic levels in ALL research. We then motivate our ongoing work, where we are re-running Experiment 2 with Polish speakers to showcase using the *same* pictographic language with multiple language communities to investigate transfer effects (Experiment 3). Future research should continue to validate using iconic artificial languages with linguistically diverse communities and continue to seek out insights from patterns of crosslinguistic influence—points we return to in the General Discussion.

3.2 Scope-isomorphism

Considerable work has sought to attribute frequent crosslinguistic patterns and hypothesized language universals to constraints on human cognition. One such proposed bias is the tendency to linearly order elements of an expression in a way that is *scope-isomorphic*, such that the expression's

[5]Our decision to focus initially on English speakers was based on interpretability: If, as our first step, we chose to study non-English speakers and did not observe the same bias, it would be unclear whether the discrepancy in findings was due to the iconic artificial language (a methodological issue) or the difference in population (a sampling issue).

underlying hierarchical structure mirrors the semantic scope relations between its individual parts. This section introduces two ordering phenomena—at the phrase and inflectional levels—that prior research has suggested may be explained by a universal bias for scope-isomorphism.

3.2.1 Modifier orders and Universal 20

Greenberg's (1963) *Universal 20* noted the tendency for languages to order demonstratives (Dem), numerals (Num), and adjectives (Adj) in a specific manner: "When any or all of the items (demonstrative, numeral, and descriptive adjective) precede the noun, they are always found in that order. If they follow, the order is either the same or its exact opposite". That is, languages favor the following linear orders within noun phrases (NPs): Dem-Num-Adj-N, N-Dem-Num-Adj, or N-Adj-Num-Dem. While the typological literature since Greenberg (1963) has in fact attested a wide variety of NP-internal word orders (at least 18 out of the 24 possible linearizations; Dryer, 2018), the Dem-Num-Adj-N and N-Adj-Num-Dem orders are estimated to account for nearly half of the world's languages, with the latter being the most prevalent (Cysouw, 2010; Dryer, 2018).[6]

Recent ALL work has attributed the dominance of these two patterns to a bias for syntactic structures that are scope-isomorphic, mirroring the underlying semantic scope relations—in this case, that demonstratives take scope over numerals and numerals over adjectives (Culbertson & Adger, 2014; Martin et al., 2020, 2019). To give an illustration, the English NP [*those* [*two* [*black feathers*]]] is considered scope-isomorphic because its constituent structure mirrors the semantics of the expression, with the demonstrative *those* picking out in space *two* entities that are *black feathers*. Importantly, the prevalent orders Dem-Num-Adj-N and N-Adj-Num-Dem are both scope-isomorphic, with the eight possible scope-conforming structures all being attested (Figure 3.1).

Using an extrapolation variant of the ALL paradigm, Culbertson and Adger (2014) and Martin et al. (2020, 2019) taught participants single-modifier NPs in an artificial language, then examined how the participants ordered *multiple* modifiers during a critical testing stage. As is key to the extrapolation paradigm, the linguistic stimuli taught to the participants were always ambiguous as to whether the artificial language adhered to scope-isomorphic or non-isomorphic linearizations, since the participants only ever saw one modifier within a given NP during training (e.g., N-Adj, N-Num,

[6]Much of the work cited in this chapter frames these structures as "noun phrases" (NPs) versus "determiner phrases" (DPs). In following suit, we are not making a strong claim about which item heads the projection, instead focusing on the hierarchical relationships therein.

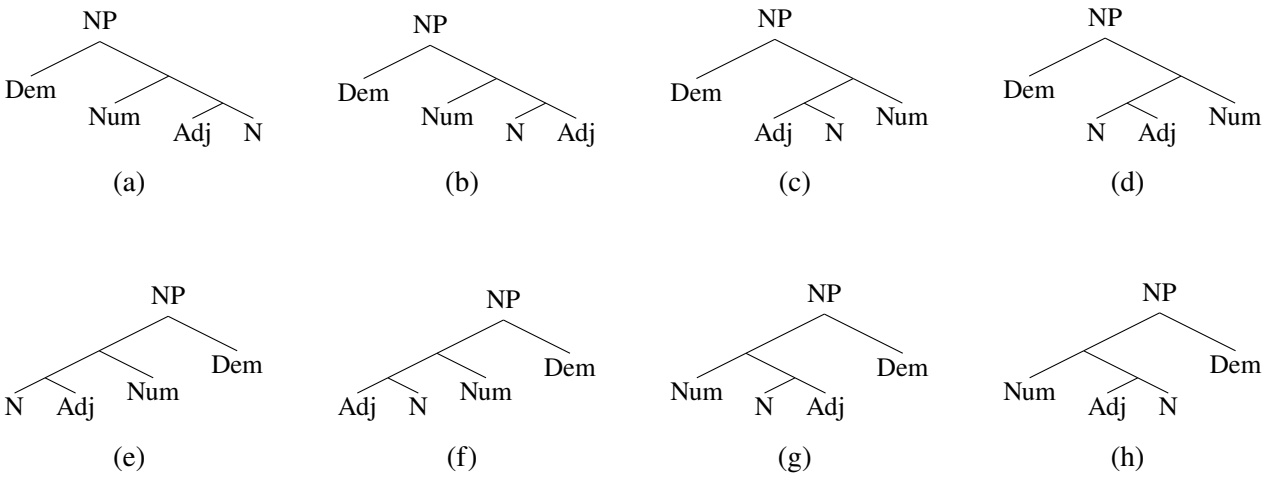


Figure 3.1: The eight possible scope-isomorphic word orders in NPs containing demonstrative determiners (Dem), numerals (Num), and descriptive adjectives (Adj). All eight orders are attested, with (a) and (e) being the most common (Cysouw, 2010; Dryer, 2018).

and N-Dem). Across the board, the studies observed a preference for scope-isomorphic word orders: When the participants had to produce novel, multi-modifier NPs, they favored ordering the noun and modifiers in a way that reflected semantic scope. However, a key question is whether this bias is universal. Though Culbertson and Adger (2014) acknowledged that their participants may have derived scope-isomorphic orders from their knowledge of English, the authors also suggested that these findings "may reflect a deep property of the human cognitive system". Notably, all three studies were conducted with English speakers, with Martin et al. (2019) additionally focusing on speakers of Thai, which has the NP structure in Figure 3.1e.

In Experiment 1 of the present study, we use iconic stimuli to replicate Martin et al. (2020) with English speakers. It is important to note that Martin et al. (2020) corrected for methodological issues in Culbertson and Adger (2014) and Martin et al. (2019). These issues may have led the participants in the first two studies to produce scope-conforming responses for reasons outside of a preference for scope-isomorphism. We revisit and expand on these confounds in our analysis of the pictographic responses in Experiment 1. We then return to the question of a universal bias for scope-isomorphic modifiers, foreshadowing Martin and colleagues' ongoing work with speakers of Kîîtharaka, a Bantu language spoken in rural parts of Kenya. Crucially, since Kîîtharaka modifiers are ordered N-Dem-Num-Adj, it's possible that the underlying modifier structure is non-isomorphic.

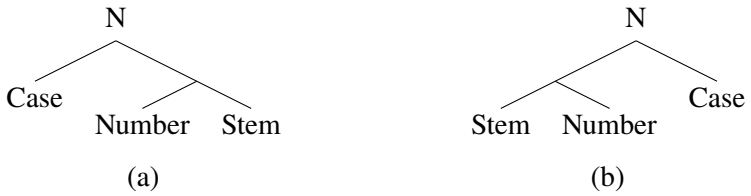


Figure 3.2: The structure of scope-isomorphic case-number inflections when both affixes appear on the same side of the noun stem. These inflections are depicted as affixes in (a) and as suffixes in (b).

3.2.2 Morpheme orders and Universal 39

Another universal proposed by Greenberg (1963) pertains to the linear order of case and number inflections when both bound morphemes are clearly delineated in a language. According to *Universal 39*, "Where morphemes of both number and case are present and both follow or both precede the noun base, the expression of number almost always comes between the noun base and the expression of case." In other words, the relevant agglutinative languages universally order these morphemes Case-Number-Stem or Stem-Number-Case.

In contrast to Universal 20, there are no known counterexamples to Universal 39—though some Sino-Tibetan languages like Chintang, spoken in Nepal, allow free permutations over other inflectional affixes (Bickel et al., 2007). Importantly, assuming that derived and inflected nouns can be decomposed hierarchically (illustrated in Figure 3.2), the Case-Number-Stem and Stem-Number-Case orders are both scope-isomorphic, since it is the noun (group) that fulfills the morphosyntactic role assigned by the case feature. Can a bias for scope-isomorphic structures explain Universal 39?

Tackling this question, Saldana et al. (2021) conducted extrapolative ALL experiments with English and Japanese speakers to show that they consistently favor scope-isomorphic morpheme orders in artificial languages. Despite never being taught how to order the number and case morphemes, the participants consistently ordered the number marker closest to the noun stem. This effect was robust both to the values of the features (plural~accusative or singular~nominative) and affix position (prenominal or postnominal), leading the authors to assert that a bias for scope-isomorphism may be a universal constraint that shapes morpheme orders in languages.

We replicate Saldana et al.'s (2021) findings with English speakers in Experiment 2. We then revisit their claim about the universality of this constraint, exploring how the prior language experience of English and Japanese speakers coulf also explain the observed scope-isomorphism

preferences. This discussion motivates Experiment 3, where we are conducting a revised version of the morpheme-ordering study with Polish speakers

3.3 Experiment 1: Modifier order with English speakers

To test the viability of a fully iconic artificial language at the word and phrase levels, we sought to reproduce the scope-isomorphism preferences that Culbertson and Adger (2014) and Martin et al. (2020, 2019) observed with English speakers. We paired a pictographic lexicon with the visual stimuli from Martin et al. (2020). Each scene from Martin et al. (2020) depicted a table with a girl standing behind it. The participants were tasked with describing objects that appeared on the table. For example, if two feathers were spread apart on the table and the girl was pointing to the nearest one, this was intended to solicit the interpretation "this feather".

Following Culbertson and Adger (2014) and Martin et al. (2020, 2019), we adopted the extrapolation paradigm, training the participants on bare and one-modifier NPs, then examining their ordering preferences when tasked with producing two-modifier NPs. We taught all three modifier types—Dem, Num, and Adj—to each participant (akin to Experiment 2a in Culbertson & Adger, 2014). To minimize transfer effects, the participants were taught noun-initial word orders (i.e., N-Dem, N-Num, and N-Adj), differentiating the artificial language from English (consistent with the earlier studies).

3.3.1 The iconic language

We created a pictographic lexicon to replace the lexical stimuli from Martin et al. (2020). Shown in Table 3.1, the pictographic lexicon was composed of three nouns (*ball*, *feather*, *mug*) and six modifiers—two demonstratives (*this*, *that*), two numerals (*two*, *three*), and two adjectives (*red*, *black*). These stimuli were never phonologically realized in our study. The individual icons—hereafter, *glyphs*—were adapted from SVG icons downloaded from Flaticon,[7] then cast into a font using IcoMoon.[8] By mapping each glyph to a character, the font made it simpler to render the linguistic stimuli in HTML/CSS and to analyze them afterward.

[7] https://www.flaticon.com/
[8] https://icomoon.io/

Noun	Adjective	Numeral	Demonstrative
 *ball*	 *red*	 *two*	 *this*
 *feather*	 *black*	 *three*	 *that*
 *mug*			

Table 3.1: The pictographic lexicon for Experiment 1.

We privileged within-category similarity in our design of the lexicon. For instance, each noun glyph was encased in a square frame and we used colored rhombuses for the adjectives. All of the modifier glyphs were approximately the same width and height to avoid visual biases that may lead participants to place modifiers closer to or farther away from the noun based on size.

3.3.2 Procedure

We constructed the experiment in jsPsych (de Leeuw, 2015) using custom plugins. The experiment was composed of forced-choice and production-style exercises. There were three training blocks (26 trials) and one testing block (24 trials), totaling four blocks altogether (50 trials).

Instructions and feedback

At the start of the experiment, the participants were told they would be learning a "pretend pictographic language". The instructions featured a mild deception that led the participants to believe they would be testing a new language learning app. This deception was done to help the participants buy into the language-learning exercise without overly analyzing the language itself; we disclosed the deception at the end of the study. In the instructions, we explicitly defined the Dem glyphs as *this* and *that* to encourage determiner interpretations of these items; we found during piloting that participants favored directional/prepositional interpretations such as *down* and *across* when they were not given the Dem meanings in advance. However, we refrained from giving English translations for the other vocabulary to minimize the participants' English awareness during the experiment.

Since one of our goals was to make the study as language neutral as possible, only the consent form and the instructions at the start of the study, as well as a post-experiment questionnaire, were presented in English (i.e., language-specific). No other English text appeared during the study, with the arguable exception of Arabic numerals in Blocks 3 and 4. In lieu of providing English-based instructions and feedback during the experiment trials, we used pictures and simple CSS animations to guide the participants through the study. For instance, at the start of each forced-choice training trial, the different options were surrounded by a neutral glowing border until an item was selected. When the correct answer was selected, its border turned green. If, on the other hand, an incorrect answer was selected, its border turned red; the correct answer would then glow green, bouncing up and down in a "pick me" animation until the participant clicked on it, advancing them to the next trial. Such animations can be used with diverse populations and have the added benefit of minimizing language-specific priming during the experiment.

Training blocks

We implemented an "active learning" design, where we refrained from *teaching* the participants the meaning of the lexical items upfront (e.g., we did not tell them ♦ means *red*). Instead, the training trials immediately quizzed the participants on the artificial language, requiring the participants to intuit the meanings of the glyphs—which was made easy by their iconicity. For each trial, animations informed the participants as to whether they had answered correctly; the participants were further required to correct any mistakes.

Block 1 of the experiment focused on noun learning. For each of the three nouns, the participants completed a glyph-selection task and a picture-selection task, totaling six trials. In the glyph-selection tasks, the participants were shown a picture of an object, then had to select the corresponding glyph from the set of three noun glyphs (Figure 3.3a). Conversely, in the picture-selection tasks, the participants were shown a noun glyph and had to select the corresponding picture from the set of three noun images. The trials were intermixed with respect to the noun and task type, and the order of the selection choices were shuffled from trial to trial.

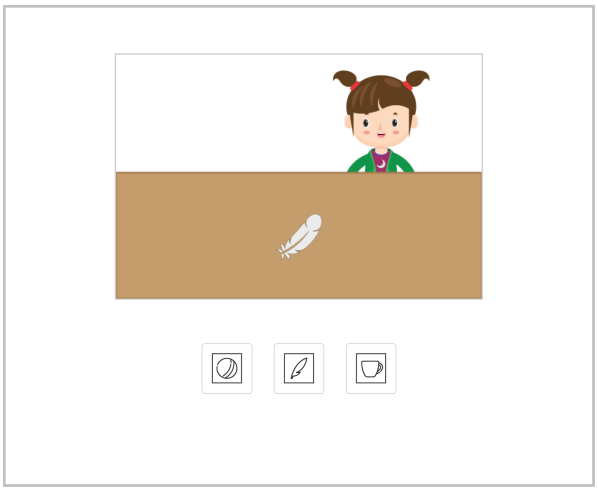
Block 2 introduced the six modifiers. For each modifier, the participants completed a picture-selection task and a cloze ("fill-in-the-blank") task, totaling 12 trials. In the picture-selection tasks (Figure 3.3b), the participants were shown a complete one-modifier NP and had to select the

corresponding picture from a set of two pictures; the foil image was always the same noun paired with the other modifier of the same category. In the cloze task (Figure 3.3c), the participants were shown a picture with an incomplete one-modifier NP caption, then had to select the missing modifier given the choices of the correct modifier and the other modifier of the same category. Note that both trial types further served to familiarize the participants with the noun-initial word order. Each noun appeared four times across the block. The trials were intermixed with respect to the noun, modifier, and task type, and the order of the selection choices were shuffled.

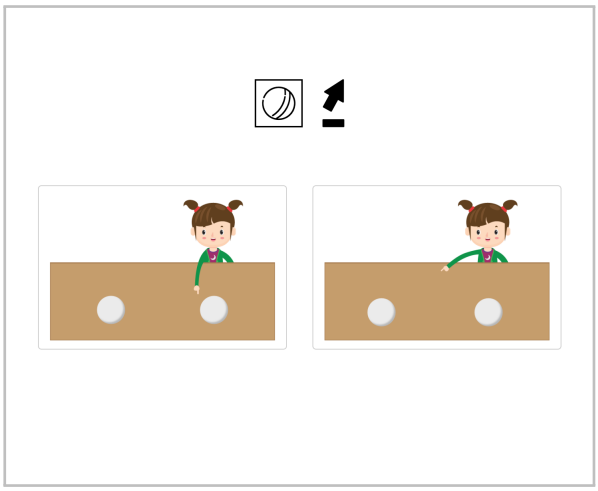
In Block 3, the final training block, the participants practiced producing zero- and one-modifier NP through eight "keyboard" trials. In addition to reinforcing the noun-initial word order, this block familiarized the participants with the production task format. In each trial, the participants were presented with a picture, then tasked with producing a caption for the image using a clickable keyboard provided on the screen (cf. Figure 3.3d). The keyboard contained all nine glyphs in the lexicon, shuffled within category *and* across categories from trial to trial; a "backspace" key appeared on the far right. The participants were only able to submit a response once they had entered the correct number of glyphs, as indicated by a glyph counter at the bottom of the screen (presented in Arabic numerals). If a participant entered the wrong caption, they were shown the correct caption and prompted to correct their answer, which then allowed them to proceed to the next trial. The block began with two randomly-selected bare noun trials, followed by six one-modifier NP trials. Note that, given the lexicon's three nouns and six modifiers, there were 18 possible one-modifier NPs; accordingly, the six one-modifier NPs presented in Block 3 were the ones held out from Block 2.

Testing block

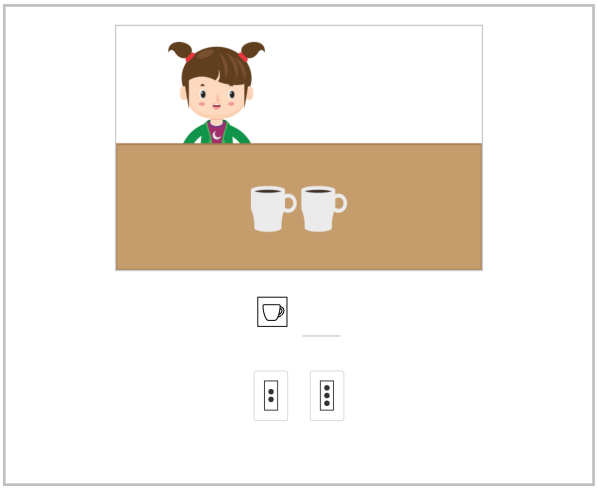
Block 4 tested the participants on one- and two-modifier NP productions using the same keyboard task design from Block 3 (Figure 3.3d). The block consisted of 12 *non-critical* one-modifier NP trials and 12 *critical* two-modifier NP trials, totaling 24 trials. The block began with four randomly-selected non-critical trials, followed by the remaining eight non-critical trials and the 12 critical trials intermixed. The 12 one-modifier NPs were the ones observed in Block 2, meaning that, across the experiment, the participants produced each one-modifier NP exactly once. The 12 two-modifier NPs included each possible combination of the six modifiers, with each noun appearing 4 times. As in Block 3, the participants could only submit a response once they had



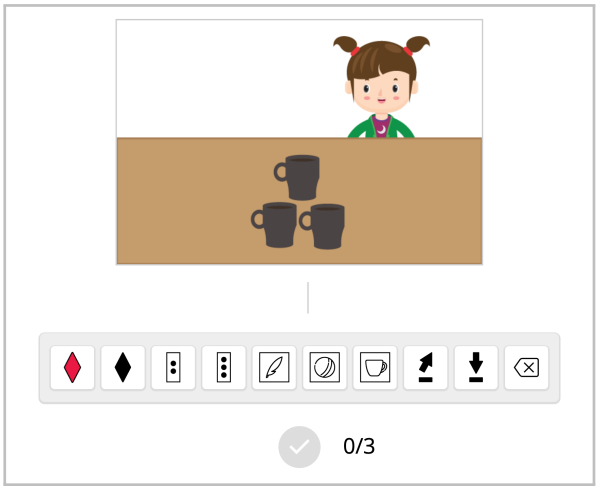
(a) Block 1 glyph-selection task



(b) Block 2 picture-selection task



(c) Block 2 cloze task



(d) Block 4 production task

Figure 3.3: Experiment 1 trials with visual stimuli from Martin, Holtz, Abels, Adger, and Culbertson (2020).

entered the correct number of glyphs. This was done to ensure that the participants provided *maximal* descriptions of the images.[9] In contrast to Block 3, the participants were not given feedback on their responses.

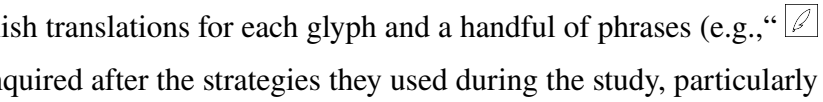
[9]For instance, if an image depicted a black feather on either end of the table with the girl pointing at the farthest one, both "that black feather" and "that feather" would be fair descriptions, since the modifier *black* is unnecessary for disambiguating the two feathers. We therefore included the glyph counter to solicit two-modifier NPs, overriding this pragmatic reasoning.

Counterbalancing

Within each block, half of the trials pictured image stimuli that were flipped horizontally, such that the girl appeared standing on the right side of image, rather than the left. This was done to prevent *left-right* interpretations of the demonstrative glyphs, as encountered by Martin et al. (2020). The flipped and unflipped images co-occurred equally with each demonstrative within each block.

In Blocks 3 and 4, the keyboards were counterbalanced such that the relevant keys appeared in matching order of the expected answer in half the trials, and in the reverse order in the remaining trials. This was done to control for effects of "key order". In the one-modifier NP trials, this meant that the noun appeared before the correct modifier in half of the trials and after the modifier in the other half. Likewise, in the two-modifier NP trials, the modifiers in question appeared in isomorphic order in half of the keyboards and in non-isomorphic order in the other half.

Post-experiment questionnaire

After the experiment, we asked the participants to complete a brief questionnaire. The questionnaire asked the participants to give English translations for each glyph and a handful of phrases (e.g.,"  "). The questionnaire then inquired after the strategies they used during the study, particularly with respect to how they ordered the modifiers during the testing block. We further asked the participants about the extent to which they "verbalized" the icons to themselves and in what languages. The questionnaire concluded with a language history form loosely inspired by the LEAP-Q (Marian, Blumenfeld, & Kaushanskaya, 2007).

3.3.3 Filtering criteria

Prior to data collection, we defined filtering criteria to ensure that our analysis included only high-quality responses from participants. Based on their Block 4 responses, we required participants to produce at least 10 "correct" non-critical one-modifier NPs (out of 12; 83%) and at least nine "analyzable" critical two-modifier NPs (out of 12; 75%), resembling the filtering criteria in Martin et al. (2020). For the non-critical NPs, a response was marked as incorrect if it included the wrong glyphs or if the glyphs appeared in the wrong order (i.e., noun-final rather than noun-initial). Likewise, for the critical NPs, a response was marked as un-analyzable if it contained the wrong

glyphs or if the noun did not appear phrase-initially. Participants whose responses did not qualify them for the analysis were still compensated for completing the study.

3.3.4 Participants

We recruited 55 participants online via the Prolific platform.[10] Using Prolific's pre-screening filters, we recruited participants who self-identified as monolingual English speakers and who reported having no language-related disorders or issues seeing colors. As required by Prolific, we only recruited participants who had previously indicated on the platform that they were comfortable with being deceived. All of the participants gave informed consent and received 4 USD in compensation. Based on our filtering criteria, we included 45 of the participants in our analysis. The participants who qualified for the analysis took on average 7.0 minutes to complete the experiment (range: 3.3-33.4) and 4.8 minutes to complete the post-experiment questionnaire (range: 1.2-24.8).

3.3.5 Results

After filtering the participant response data, we wound up with 508 analyzable two-modifier NPs (165 Dem-Adj, 170 Dem-Num, and 173 Num-Dem). Out of the analyzable two-modifier NPs, 82.3% were in scope-isomorphic order (Dem-Adj: 87.3%, Dem-Num: 90.6%; Num-Adj: 69.3%). *By participant*, the average percentage of the responses that were scope-isomorphic was 91.2% (Dem-Adj: 85.6%; Dem-Num: 89.8%; Num-Adj: 68.3%), visualized by modifier group in Figure 3.4.

We fit a logistic mixed-effects regression model to the critical NPs, predicting a response's modifier order (1 = *isomorphic*; 0 = *non-isomorphic*). We included random intercepts for participants, but did not do so for each noun (contra Culbertson & Adger, 2014; Martin & Culbertson, 2020; Martin et al., 2019), since there was zero variance. We found that the intercept was positive and significant (Table 3.2), indicating that the participants chose scope-isomorphic modifier orders at above chance level.

In a likelihood ratio test, we then compared the model to one that included a fixed effect for modifier group (Dem-Adj, Dem-Num, and Num-Adj) and found that the latter model better fit the data ($p < 0.001$), indicating that the preference for scope-isomorphic orders was stronger for some

[10] https://www.prolific.co

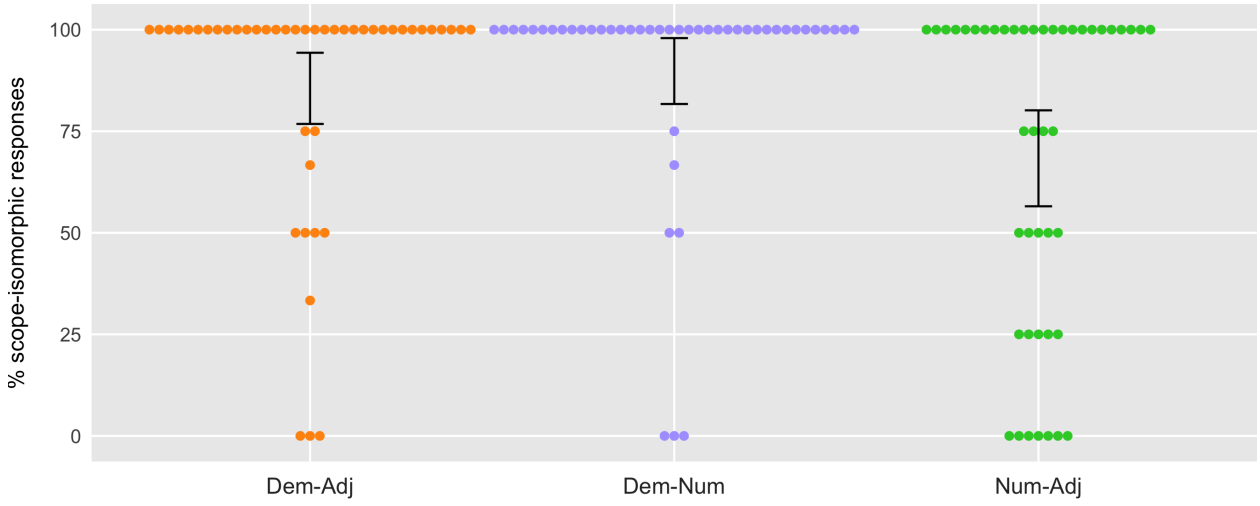


Figure 3.4: Percentage of responses that were scope-isomorphic *by participant* in Experiment 1 (with 95% confidence intervals). Each dot represents a participant. Note that this plot is somewhat misleading: Since the participants encountered only four items per modifier group and were only required to produce nine analyzable two-modifier NPs (out of 12), each percentage dot is calculated out of a total of 1–4 items.

(Intercept)	$\hat{\beta}$	SE	z	p
All data	2.274	0.364	6.240	<0.001
Dem-Adj	7.783	1.697	4.586	<0.001
Dem-Num	9.922	2.083	4.764	<0.001
Num-Adj	2.298	0.962	2.388	0.017

Table 3.2: Logistic regression results for Experiment 1. The top row conveys the results from the full model (all modifier groups) while the bottom three rows display the results per modifier group.

modifier pairs than for others. To facilitate interpretation, we then fit separate models for each group, confirming that the scope-isomorphism preference held between each pair of modifiers, albeit to different degrees (Table 3.2). This preference was strongest for the pairs involving demonstratives (Dem-Adj and Dem-Num), with the Dem-Num model presenting the largest intercept. A post hoc analysis excluding just the Num-Adj data revealed no significant difference between the Dem-Adj and Dem-Num groups (i.e., no significant effect of modifier group; $p = 0.096$).

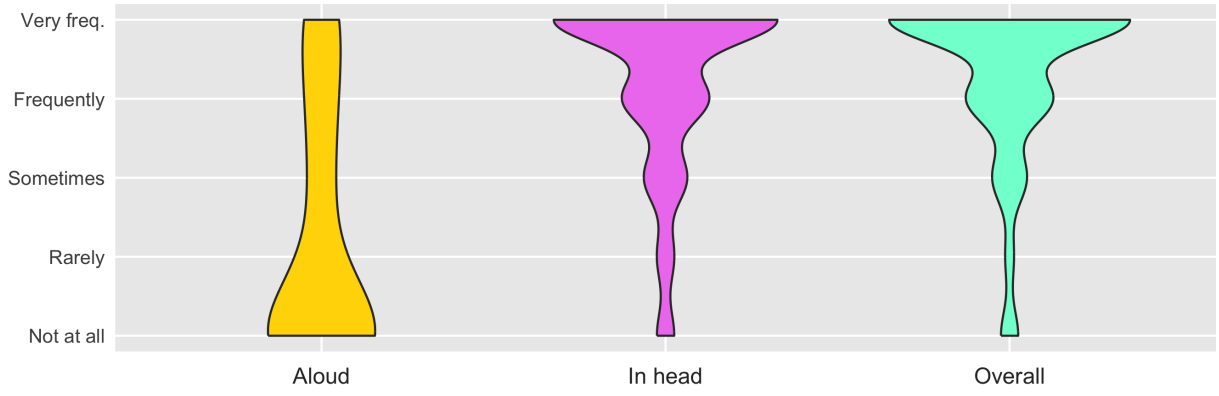


Figure 3.5: The frequency with which the participants "verbalized" the glyphs in Experiment 1. All of the participants provided ratings.

3.3.6 Discussion

In this experiment, we set out to test the viability of iconic artificial languages. We used a pictographic lexicon to replicate prior findings showing that English speakers favor noun and modifier orders that preserve semantic scope relations (Culbertson & Adger, 2014; Martin et al., 2020, 2019). Consistent with this work, our experiment showed that participants were significantly more likely to produce scope-isomorphic NPs, validating iconic artificial languages at the word and phrase levels.

Response analysis

Perhaps ironically, the absence of a phonologically realized lexicon may inadvertently lead to transfer effects if participants are *verbalizing* the iconic stimuli to themselves in their own languages. Indeed, Likert scales in our post-experiment questionnaire revealed that the participants often 'said the symbols out loud' or 'in their head', as summarized in Figure 3.5. This could have resulted in unwanted transfer effects from English.

However, it is worth highlighting that the experiments in Culbertson and Adger (2014) and Martin et al. (2019) involved similar noun-initial NPs, but used English words for English speakers and Thai words for Thai speakers (e.g., teaching English-speaking participants *ball that* to mean "that ball"). As explored in Martin et al. (2020), this led many participants to adopt a "flipping" strategy, where they arrived at scope-isomorphic NPs simply by reversing the English and Thai word orders. Martin et al.. therefore used an artificial lexicon with nonce nouns and modifiers (e.g.,

puku), finding still a scope-isomorphic preference, but without the confound of the participants consciously transferring their English knowledge to the task.

Similarly, only one participant in the present study reported using a flipping strategy, with the majority of the participants reporting that they just picked a word order and stuck with it. This shows that, despite the participants articulating the iconic lexical items to themselves, this did not trigger the same level of crosslinguistic influence encountered by Culbertson and Adger (2014) and Martin et al. (2019). Future work can compare this form of metalinguistic awareness across iconic and traditional ALL tasks. Yet, if verbalizing iconic stimuli *can* elicit transfer effects, if exploited strategically, this could provide an interesting avenue for studying and comparing crosslinguistic influence across different languages, while holding the artificial language constant.

Semantic scope

Experiment 1 further supports a preference for scope-isomorphic word orders within the noun phrase, at least among English speakers. Furthermore, we found that the strength of this bias was modulated by modifier pair, with comparably strong preferences observed among the Dem-Adj and Dem-Num trials and a weaker preference among the Num-Adj trials.

On the surface, these findings differ modestly from those of Culbertson and Adger (2014) and Martin et al. (2020): When running separate *conditions* for each modifier pair (i.e., only teaching and testing participants on two modifier types), both sets of authors saw significantly higher proportions of scope-isomorphic orders in the Dem-Adj condition, with the Dem-Num condition instead patterning more similarly to the Num-Adj condition. This led Culbertson and Adger (2014) to hypothesize that a preference for scope-isomorphism for a subset of modifiers (e.g., Dem-Adj) may lead to a stronger scope-conforming mapping for *all* of the modifiers. In follow-up experiments, the authors taught all three modifiers to the same set of participants—akin to the present study. When they did so, the proportion of scope-isomorphic responses no longer differed significantly between the groups.

Returning to the present experiment, one possibility is that teaching all three modifier categories to our participants similarly resulted in a stronger scope-conforming preference in the Dem-Num group. At the same, since our experiment was much shorter than Culbertson and Adger's, it may be that the participants in our study were not "immersed" enough in our artificial language to develop

an equally strong scope-isomorphism preference in the Adj-Num group.

Interestingly, Martin et al. (2020) suggested that the observed between-condition effects may reflect differences in the "strength of associations" held between the lexical categories. Citing Culbertson, Schouwstra, and Kirby (2020), who quantified these associations with pointwise-mutual information, Martin et al. speculated that "this kind of dependency determines how likely speakers are to separate a head and dependent". Notably, Culbertson et al. (2020) had found that N~Adj, on average, were most closely associated, followed by N~Num and then N~Dem. Building on Martin et al.'s proposal, it's possible that when taught all three modifier categories, the amount of exposure required to form a strong scope-conforming mapping between two particular modifiers is relative to the same strength of associations. We leave this question for future work. However, if true, the present findings would further demonstrate the ability of iconic artificial languages to test fine-grained linguistic hypotheses.

Finally, a remaining question is whether this bias for scope-isomorphism is universal or instead reflects structural transfer from the participants' prior language experience, since English NPs are scope-isomorphic. As hinted by Martin et al. (2020) and confirmed in personal communication, Martin and colleagues are currently addressing this question with speakers of Kîîtharaka, a Bantu language where the word order is not straightforwardly scope-conforming: N-Dem-Adj-Num (Kanampiu & Muriungi, 2019).[11] If Kîîtharaka speakers display no ordering preferences or prefer non-isomorphic modifiers, it's possible that the bias is not universal or, alternatively, that familiarity with non-isomorphic modifiers can dampen or reverse it (cf. discussion by Martin & Culbertson, 2020, on the prefixing-suffixing asymmetry).

Conversely, if Kîîtharaka speakers do prefer scope-isomorphic modifiers in artificial languages, this would be the strongest evidence yet of the universality of this bias. In such an event, it would be especially interesting to compare their preferences with speakers of additional non-isomorphic languages (e.g., Stanfard Tibetan) and scope-isomorphic languages (e.g., English, Thai) to see if the *strength* of the bias is modulated by the NP linearizations in their respective languages. Is there is

[11] Situated within the generative tradition (and adopting the DP Hypothesis), Kanampiu and Muriungi (2019) argue that Kîîtharaka DPs are base-generated in scope-isomorphic order—Dem-Num-Adj-N—but that the NP moves into Spec-DP to check agreement features, deriving the non-isomorphic order N-Dem-Adj-Num. It is a deep question how we would expect a scope-isomorphism bias to bear on pre- versus post-movement representations. Assuming the psychological reality of this analysis, could *underlying* scope-isomorphism present a confound for confirming the universality of a scope-isomorphism bias?

still an effect of crosslinguistic influence? On that note, iconic ALL provides a promising direction for measuring and juxtaposing transfer effects in this way across diverse populations in future work.

3.4 Experiment 2: Morpheme order with English speakers

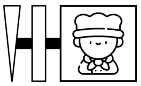
While Experiment 1 demonstrated the utility of iconic artificial languages at the word and phrase levels, Experiment 2 seeks to test pictographic stimuli at the inflectional and sentence levels. We therefore attempt to replicate the findings of Saldana et al. (2021). Across a series of experiments, the authors showed that English and Japanese speakers prefer scope-isomorphic orders between number and case morphemes, consistent with Universal 39. In their study, participants learned verb-subject-object (VSO) sentences to describe scenes containing both agents and patients (e.g., a waitress pointing at two burglars). Pairing Saldana et al.'s visual (scene) stimuli with a new pictographic lexicon, we taught plural and accusative markers to English speakers. Crucially, the participants never encountered plural objects during training, allowing us to observe their ordering preferences when shown a scene with multiple patients at test time. Following Saldana et al. (2021), we hypothesized that the participants would predominantly order the plural marker closest to the noun stems, without any meaningful differences between prenominal and postnominal conditions.

3.4.1 The semi-iconic language

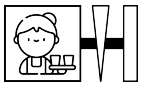
Table 3.3 shows the pictographic lexicon for the lexical items from Saldana et al. (2021): three verbs (*kick*, *point*, *punch*), four nouns (*burglar*, *chef*, *cowboy*, *waitress*), and two affixes (plural, accusative). We again privileged within-category similarity, for instance, by enclosing the verbs in circles and the nouns in squares. In contrast to Experiment 1, the lexicon for Experiment 2 was only semi-iconic, with the two affixes being of arbitrary design. Likewise, the two affix glyphs were approximately the same width and height. To encourage *inflectional* analyses of these markers, we further depicted the glyphs as "bound" to the noun stems (cf. Saldana et al., 2021), as illustrated in Figure 3.6. When casting the lexicon into a font, we implemented the stem+marker bigrams (and subsequent trigrams) as ligatures. Finally, though we adopted VSO word orders across the experiment, we deviate from Saldana et al. somewhat by also including VOS word orders (discussed below).

Verb	Noun	Prefix	Suffix
 *kick*	 *burglar*	 PLURAL	 PLURAL
 *point*	 *chef*	 ACCUSATIVE (object marker)	 ACCUSATIVE (object marker)
 *punch*	 *cowboy*		
	 *waitress*		

Table 3.3: The pictographic lexicon for Experiment 2. The prefixes appeared in the prenominal condition and the suffixes in the postnominal condition.



(a) ACC-PL-chef



(b) waitress-ACC-PL

Figure 3.6: Examples of inflected (two-marker) nouns in Experiment 2. The plural and accusative markers are depicted as prefixes in the prenominal condition, shown in (a), and as suffixes in the postnominal condition, shown in (b). Example (a) is scope-isomorphic and example (b) is not.

3.4.2 Procedure

Slightly longer than the first experiment, Experiment 2 consisted of four training blocks (36 trials) and one testing block (24 trials), totaling five blocks (60 trials). All of the blocks were again composed of forced-choice and production-style exercises, and featured an active learning design and simple animations. Only the consent form, instructions, and post-experiment questionnaire were language-specific (i.e., in English). Lastly, it is important to note that in the scene stimuli from Saldana et al. (2021), the plural nouns were always depicted with exactly *two* characters (i.e., two agents or two patients). We discuss the implications of this at the end of the section.

Instructions

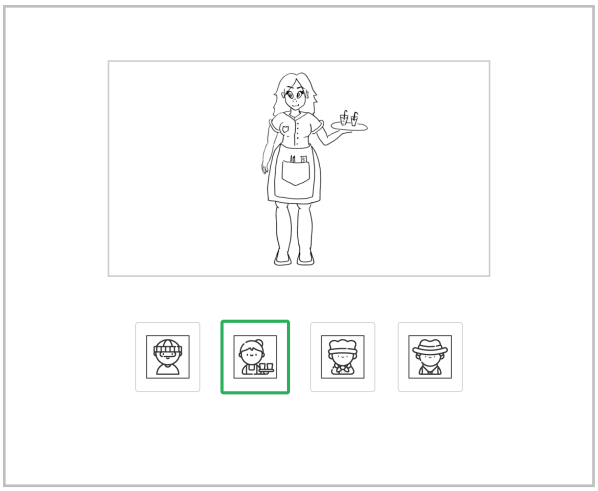
The experiment began with a small deception, where the participants were told they would be learning a real language that had been transcribed into pictograms. Similar to Experiment 1, this deception was done to help the participants accept the validity of the artificial language and was disclosed at the end of the study. The participants were further told that they would be learning either prefixes or suffixes (depending on the condition) that mark "plural nouns and characters who receive a verb action".

Training blocks

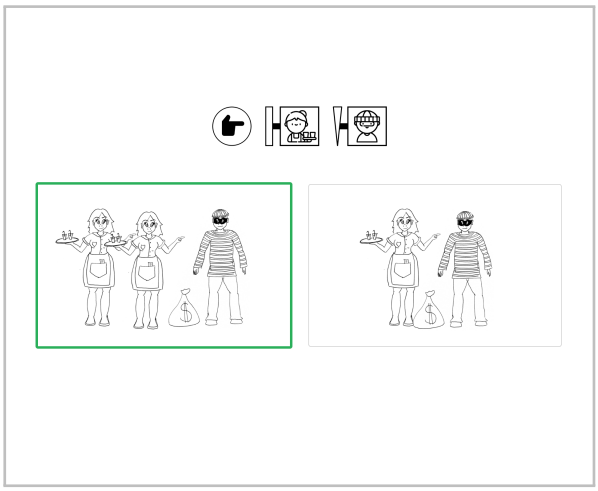
Blocks 1 and 2 focused on noun learning and verb learning respectively. Block 1 consisted of four glyph-selection trials, one per noun. The participants were shown a picture of a single character (a bare noun) and had to select the corresponding glyph from the set of four nouns, which were shuffled from trial to trial (Figure 3.7a). Block 2 then introduced the verb glyphs in three cloze ("fill-in-the-blank") trials, one per verb. The participants were shown a scene with an incomplete caption, then had to select the missing verb from the set of three verb glyphs (shuffled). These trials also served to implicitly familiarize the participants with the VSO word order. Out of the three verbs trials, the subject was singular once and plural twice.

Block 3 trained the participants on one-marker inflections through 21 picture-selection trials, where the participants were shown a complete sentence and had to select the corresponding picture from a set of two pictures (shuffled). Thirteen of the trials focused on "Number Learning" (Figure 3.7b), with the foil image depicting the wrong number of agents; the subject was singular in five trials and plural in eight. The remaining eight trials focused on "Case Learning" (Figure 3.7c), where the foil image swapped the subject and object (shuffled and both singular). In contrast to similar trials in Saldana et al. (2021), we also inverted the subject and object *in the sentence*, such that the word order was VOS in these trials. This was done to further associate the case marker with the sentential object, since participants might otherwise interpret it as some kind of end-of-sentence marker, especially in the postnominal condition (cf. Saldana et al., 2021). Across the block, the trials were intermixed with respect to the Number and Case Learning exercises.

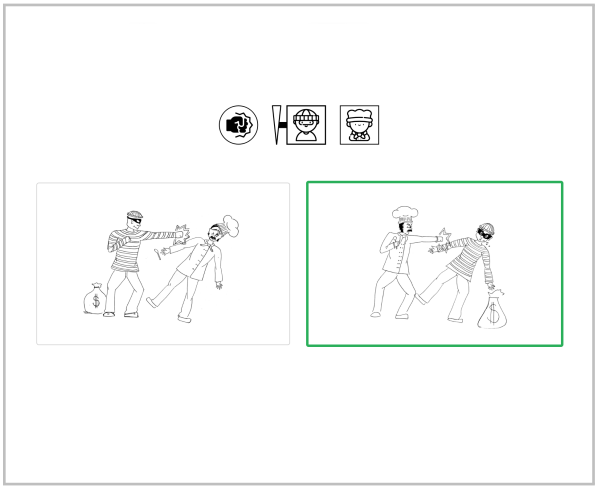
In Block 4, the final training block, the participants practiced producing zero- and one-marker inflections through eight keyboard trials (similar to Experiment 1). Like in the previous blocks,



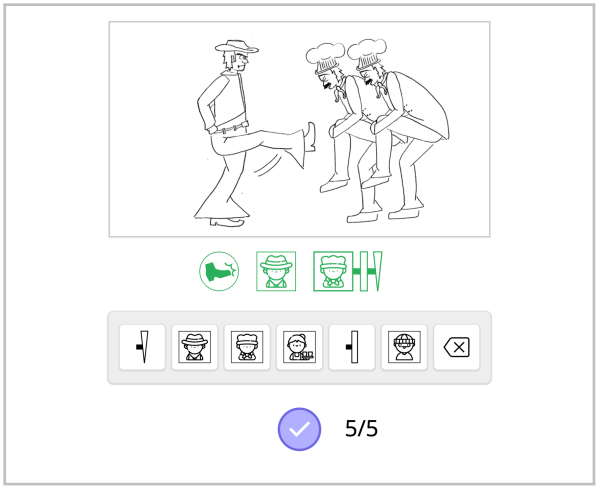
(a) Block 1 glyph-selection task



(b) Block 3 “Number Learning” task



(c) Block 3 “Case Learning” task



(d) Block 5 production task

Figure 3.7: Experiment 2 trials with visual stimuli from Saldana, Oseki, and Culbertson (2021). Correct responses are shown in green.

the two case markers never appeared on the same nouns, since only the subjects were pluralized. Following Saldana et al. (2021), we provided the participants with the sentence-initial verb, so that they only needed to provide the two inflected nouns. The keyboard included a key for each of the four noun glyphs and the two morpheme glyphs (shuffled), in addition to the “backspace” key on the far right. The participants could only submit a response once they had entered the correct number of glyphs and were required to fix any incorrect captions before they could proceed to the next trial. The block began with two randomly-selected bare noun trials (i.e., not embedded in a sentence).

These were followed by six full-sentence trials; the images featured one agent in two trials and two agents in the rest (intermixed). Both VSO and VOS word orders were accepted.

Testing block

Block 5 tested the participants on sentence productions with one- and two-marker nouns via 24 keyboard trials (Figure 3.7d). The 12 *non-critical* sentences had singular objects (one-marker nouns) and the 12 *critical* sentences had plural objects (two-marker nouns). Both the critical and non-critical trials appeared evenly with singular and plural subjects. In the critical trials, the keyboards were counterbalanced such that the two markers always appeared on either side of and equidistant from the object noun, with the remaining keys otherwise randomized; furthermore, the critical trials alternated between which affix appeared to the left and right of the noun. The block began with four randomly-selected non-critical trials, followed by eight non-critical and 12 critical trials intermixed. As in Experiment 1, the participants were required to submit the correct number of glyphs but were not given feedback on their responses.

Post-experiment questionnaire

After the testing block, the post-experiment questionnaire asked the participants to translate the individual glyphs, followed by a few sentences into English (e.g., "  "). Afterwards, the participants were asked about the strategies they used during the training blocks and how they approached ordering the two markers during the testing block. The verbalization Likert scales and language history questions were the same as in Experiment 1.

3.4.3 Filtering criteria

Similar to Experiment 1 (cf. Saldana et al., 2021), we required participants' Block 5 responses to include at least 10 correct one-marker NPs (out of 12; 83%) and at least nine analyzable two-marker NPs (out of 12; 75%). A response was marked as analyzable only if it contained the correct nouns (regardless of order) and the correct markers on each noun (regardless of order).

3.4.4 Participants

We recruited 64 participants online via Prolific. Using the platform's pre-screening filters, we only recruited participants who self-identified as monolingual English speakers and who reported having no language-related disorders. All of the participants had previously indicated on Prolific that they were comfortable with being deceived. Each participant gave informed consent at the start of the study and received 5.50 USD in compensation, based on an estimation that the study would take 15-20 minutes to complete.

Out of the recruited participants, 32 were assigned to the prenominal condition and 32 to the postnominal condition; however, one participant in the latter condition did not finish the study. Based on the remaining participants' Block 5 responses, we included 41 of the participants in our analysis (21 prenominal, 20 postnominal). The participants who qualified for the analysis took on average 11.0 minutes to complete the experiment (range: 6.9-23.8) and 4.8 minutes to complete the post-experiment questionnaire (range: 1.6-10.8).

3.4.5 Results

Given the filtered response data, there were 575 analyzable two-marker nouns (240 in the prenominal condition + 235 in the postnominal condition). Out of the analyzable two-marker NPs, 86.1% ordered the morphemes scope-isomorphically (prenominal: 83.3%; postnominal: 88.0%). The average percentage of scope-isomorphic responses *by participant* was 86.3% (prenominal: 83.5%; postnominal: 89.2%). The by-participant results are visualized in Figure 3.8, divided by condition.

We fit another logistic mixed-effects regression model to the critical nouns, predicting a response's marker order. We included a fixed effect for marker position (i.e., prenominal vs. postnominal condition), as well as by-participant random intercepts. Similar to Experiment 1, we did not include a random effect for the marked nouns, which displayed zero variance; the same was observed by Saldana et al. (2021). We found that the intercept was positive and significant, indicating that the participants chose scope-isomorphic morpheme orders at above chance level, while the fixed effect for marker position did not come out significant (Table 3.4).

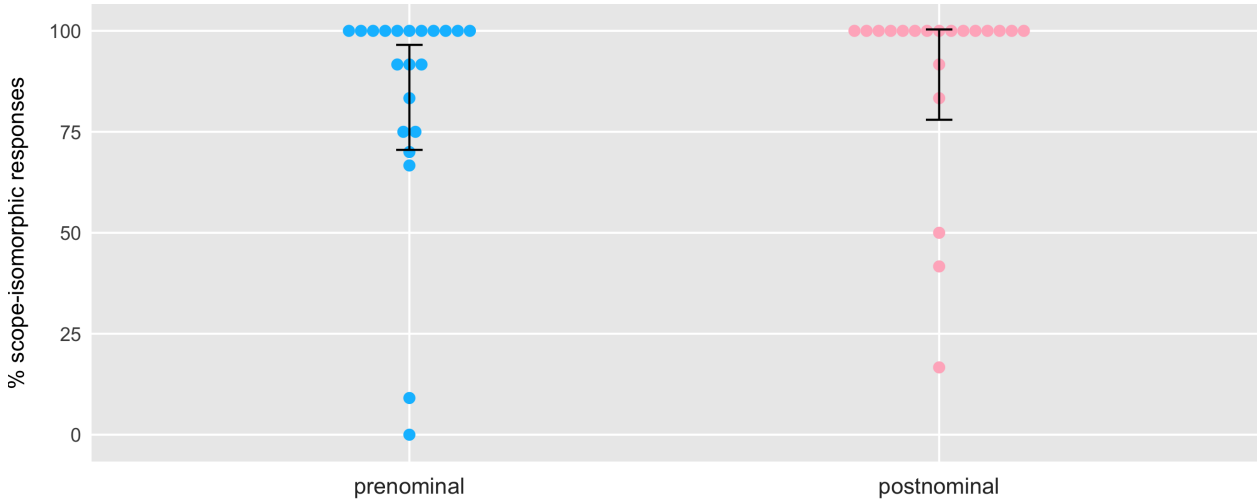


Figure 3.8: Percentage of responses that were scope-isomorphic *by participant* in Experiment 2 (with 95% confidence intervals). Each dot represents a participant and was calculated out of a total of 9–12 analyzable two-marker nouns.

	$\hat{\beta}$	SE	z	p
(Intercept)	4.297	1.578	2.723	0.006
Marker Position	1.889	1.467	1.287	0.198

Table 3.4: Logistic regression results for Experiment 2.

3.4.6 Discussion

In Experiment 2, we aimed to replicate the findings of Saldana et al. (2021). Consistent with their work, we found that English speakers were strongly inclined to place the number marker closer to noun stems than the case marker—that is, to order the markers scope-isomorphically. This bias was evident, regardless of whether the markers appeared prenominally or postnominally. These findings further validate the potential for using pictographic stimuli, this time at the sentence level.

However, due to ambiguity surrounding how the participants analyzed the number marker, especially with respect to its boundedness, it is unclear whether Experiment 2 demonstrates the ability to use pictographic *affixes* in ALL stimuli and, likewise, whether the current results bear on Universal 39. We unpack these issues below and discuss factors outside of semantics that may be driving the scope-isomorphic responses in Experiment 2 as well as in Saldana et al. (2021). In doing so, we motivate Experiment 3, where we seek to elucidate these questions.

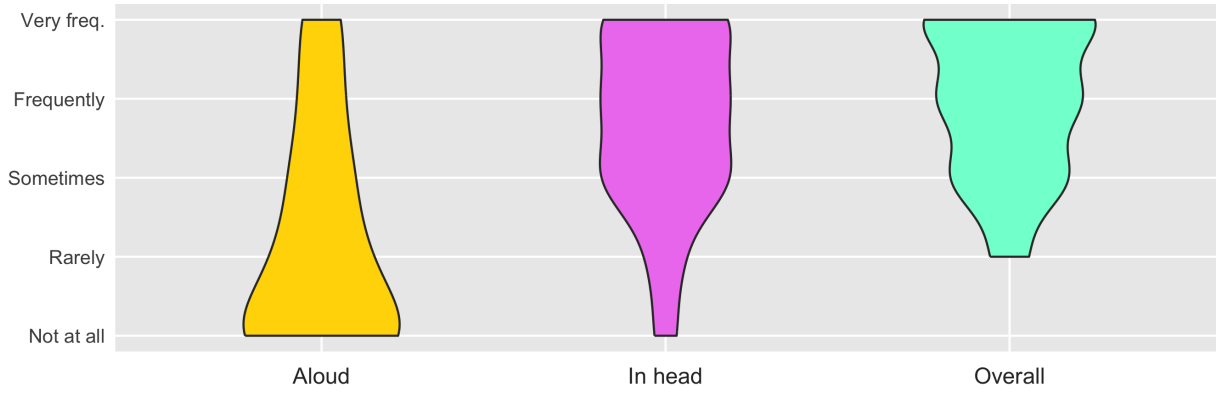


Figure 3.9: The frequency with which the participants "verbalized" the glyphs in Experiment 2. All but one of the participants provided ratings.

Response analysis

According to their responses in the post-experiment questionnaire, the participants largely noted articulating the glyphs to themselves as they moved through the experiment (Figure 3.9), though to a lesser degree than in Experiment 1. In the glyph translations, 70.7% (29/41) of the participants—61.9% (13/21) in the prenominal condition and 80.0% (16/20) in the postnominal condition—reported analyzing the case marker as signifying patients, using descriptors like *receiver*, *recipient*, *object*, *target*, and *victim*. Furthermore, only 9.8% (4/41) of the participants interpreted the case marker as an adposition (two per condition). As pointed out by Saldana et al. (2021), analyzing the case marker as an adposition may lead participants to place it outside of the number marker for syntactic reasons, rather than for reasons of semantic scope. Notably, far more of our participants arrived at non-adpositional *patient* interpretations of the case marker than in Saldana et al.'s equivalent experiment. We can likely credit this to the subject-object inversion in the Block 3 "Case Learning" trials.

Turning to the number marker, participants similarly favored specific descriptors in the glyph translations, with 63.4% (26/41) of the participants describing the marker as *two* and only 14.6% (6/41) using the word *plural* specifically (despite being given this information in the instructions).[12] Crucially, the numeral *two* is a free morpheme. At first, it may seem unproblematic to assume

[12]In the prenominal condition, 66.6% (14/21) described the number marker as *two* while only 9.5% (2/21) described it as *plural*. By modest contrast, in the postnominal condition, 60.0% (12/20) described it as *two* and 20.0% (4/20) as *plural*. While our data is inconclusive, it could be that English speakers are more inclined to analyze the number marker as a plural affix in the postnominal condition, given their familiarity with the English plural suffix *-(e)s*.

that the participants—for whom the word *plural* may be less accessible or productive—used *two* to denote pluralization. (Similarly, the remaining participants had used descriptors like *double*, *multiple*, and *more than one*.) However, in the sentence translation portion of the questionnaire, the majority of the participants (78.0%; 32/41) consistently added the numeral *two* to quantify more than one character (e.g., "the chef punched the two cowboys"). In contrast, only four participants relied exclusively on English plural morphology to indicate the number of characters (e.g., "the chef punched the cowboys"), while two additional participants switched between using pluralization exclusively and pairing it with a numeral across the different sentence translations.

One possibility is that the participants mostly analyzed the markers as affixes, but in an effort to make their translations more informative—think Grice's Maxim of Quality—they chose to insert the numeral *two*. On the other hand, the translation questions did not include scene stimuli, so the participants did not know *a priori* that a number-marked subject, for instance, necessarily meant *two* agents, unless they assumed that the sentences would not refer to scene types they had not seen previously (e.g., scenes with *three* agents). Alternatively, it could be that the participants analyzed the number marker as something akin to a *dual* number morpheme, though this would be surprising given the absence of this feature in English inflectional morphology.

If the participants genuinely analyzed the number marker as a numeral versus a bound morpheme, it is unclear whether Saldana et al. (2021) encountered similar interpretations in their experiments. The authors reported that their participants unanimously assigned a meaning of plurality to the number markers, but they did not share how exactly the participants defined the markers in their post-experiment questionnaire (e.g., with words such as *plural*, *two*, or *double*). Regardless, a prevalence of numeral interpretations in either study would likely arise from the scene stimuli never featuring more than two agents or patients—an issue we remedy in Experiment 3. In the interim, given the ambiguity surrounding whether the number marker was analyzed as a free or bound morpheme, we cannot conclude that Experiment 2 demonstrates the ability to portray bound morphology using pictographs.

Semantic scope

If the number marker was largely analyzed as the numeral *two*, this could have led the participants to interpret the case marker as a clitic or particle that sits at the edge of the entire phrase (such as

the accusative particle -を *-o* in Japanese—discussed further in the coming pages). In the event that the markers were not analyzed as nominal affixes, our results would still evidence a bias for scope-isomorphism, just not one that operates over inflectional morphemes. The present experiment would thus not bear on Universal 39, which is specific to bound morphology. Similarly, Saldana et al. (2021) did not restrict their analyses to this domain and included free case and number markers in their experiments. They ultimately concluded that the scope-isomorphism preference they observed held independent of "degree of boundedness"; we return to this idea in the General Discussion.

Setting aside the boundedness of the participants' marker interpretations, it is worth exploring other explanations beyond scope-isomorphism that may account for the current results. In particular, frequency effects could have played a role in the participants' ordering preferences: During training, every sentence contained an accusative marker, while only a subset of the sentences contained a plural marker. This had the overall effect of skewing the morpheme frequency toward the case marker, where the stem + case bigrams were roughly twice as frequent as the stem + number bigrams (15:7), following Saldana et al. (2021).[13] This could have led participants in both studies to place one marker closer to the noun stem than the other.

For instance, we might expect participants to preserve the more frequent bigrams (stem + case), rather than inject the less frequent (number) marker in the middle. Citing Hay's (2001) parsability principle, Saldana et al. similarly conjectured that the lower frequency of the number marker *relative* to the frequency of the stems by themselves might make it more parsable (decomposable) than the case marker, prompting participants to place it farther away from the noun stem than the case marker. However, it's worth noting that Hay's parsability principle, also known as Complexity-Based Ordering (Hay, 2002; Hay & Plag, 2004), was specific to English derivational morphology.[14] Even so, if were to extend the parsability prediction to inflectional morphemes in an artificial language, it would still seemthat our scope-isomorphic findings, like Saldana et al.'s, were robust to both

[13] In their initial experiments, Saldana et al. (2021) exposed the participants to both markers with equal frequency during training by including 'Number Only' trials, where singular and plural nouns appeared in isolation (i.e., outside of a sentential context). However, the authors removed these trials in subsequent experiments out of concern that they unduly led participants to privilege number-inflected nouns as a unit before marking them for case. For the same reason, we did not include such trials in our experiment.

[14] Hay (2001) closed with, "How does inflection fit into the picture—a question beyond the scope of this paper, but one that will clearly need to be addressed. And, perhaps the biggest question of all, how well can the account suggested generalize to other languages—languages both with different parsing and segmentation strategies in speech perception, and with different morphological systems?" Regarding the principle's crosslinguistic generalizability, consider Rice's (2011) discussion of Athabaskan.

absolute and relative frequencies.

Beyond distributional effects, other factors could have led the participants to place the number marker closest to the noun stem, both in the present experiment and in the original study. The familiarity of English speakers with number but not case marking may lead them to more closely associate number markers with noun stems compared to case markers. In particular, English speakers are accustomed to both (i) plural affixes and (ii) few inflections being able to intervene between a noun stem and plural affix (gerunds are a possible exception). This prior language experience could bias participants away from allowing an unfamiliar affix to intervene between the stem and a number marker in an artificial language.

To control for this issue, Saldana et al. (2021) taught novel plural~accusative markers to Japanese speakers who happened to speak English as a second language. Importantly, Japanese is also an agglutinative language that, opposite to English, includes case marking but does not inflect for number. The participants were thus familiar with both case marking (from Japanese) and plural inflectional morphology (from English). That the Japanese speakers largely produced scope-isomorphic responses, Saldana et al. argued, was even stronger evidence of a universal bias for scope-isomorphism, especially since case marking would have been more prominent for those participants.

However, as alluded to earlier, though the participants significantly favored scope-isomorphic responses, their competence in Japanese could have additionally led to an ordering confound, since Japanese case marking is expressed through *particles*. Crucially, it has long been noted that Japanese particles occupy a special space between bound morphemes and independent words (Hattori, 1950; Kageyama, 2020; Tsujimura, 2013; Vance, 1993; *inter alia*; see also Dobashi, 2003)—what Hattori (1950) termed 附属語 *fuzoku-go* or "attached words"—distinguishing them from other case systems.[15] Vance (1993), in particular, classified Japanese particles as clitics according to several criteria put forth by Zwicky and Pullum (1983) and Zwicky (1985). For example, he remarked that two coordinated nouns in Russian will both be inflected for accusative case (1a), whereas the Japanese accusative particle will appear only after the full noun phrase (1b), evidencing the latter's non-status as an inflectional affix:

(1) a. *Ivan chitayet gazyet-**u** i knig-**u**.*
is-reading newspaper-ACC and book-ACC

[15] Vance (1993) translated *fuzoku-go* as "non-independent, free forms", but here I adopt my mom's translation.

'Ivan is reading a newspaper and a book.'

b. *Tarô wa shinbun to hon* ***o*** *yonde iru.*
TOP newspaper and book ACC reading is
'Taro is reading a newspaper and a book.'

(Vance, 1993; emphasis added)

Likewise, Kageyama (2020) described the ability of Japanese particles to scope over entire phrases, showing also that parenthetical clauses can separate them from their hosts (2):

(2) *Watasi no yuuzin (Kanada-zin no isya desu ga)* ***wa*** *Tokyoo to Kyooto*
I GEN friend (Canada-person COP doctor POL.COP CONJ) TOP Tokyo and Kyoto
(dotiramo Nihon no yuumei na mati desu) ***o*** *otozure-masi-ta.*
(both Japan GEN famous COP cities POL.COP) ACC visit-POL-PST
'My friend (who is a Canadian doctor) visited Tokyo and Kyoto (both are famous Japanese cities).'

(Kageyama, 2020)

In other words, Japanese speakers are accustomed both to the scoping abilities of case particles and to material intervening between case particles and the nouns they mark. Moreover, the Japanese-speaking participants in Saldana et al. (2021) were familiar with the tight coupling of stem + number morphemes in English. These combined facts could have influenced the participants' ordering preferences, prompting them to place the number marker closest to the stem and the case marker farther away—congruent with scope-isomorphism.

To summarize, Experiment 2 is consistent with Universal 39 insofar as our participants analyzed the number and case markers as *affixes*—which remains ambiguous. Regardless, our findings may still evidence a preference for scope-isomorphic orders of case and number *morphemes*. Still, it is unclear whether this bias stems from a universal cognitive constraint or instead emerged from our participants' deep familiarity with English. The same applies to Saldana et al.'s (2021) findings with English and Japanese speakers. In Experiment 3, we seek to address these issues by conducting a modified version of the experiment with speakers of a fusional language: Polish.

3.5 Experiment 3: Morpheme order with Polish speakers (ongoing)

Since the English- and Japanese-speaking participants in Experiment 2 and in (Saldana et al., 2021) may have produced scope-isomorphic orders for reasons of crosslinguistic influence, it remains undetermined whether there is a universal bias for scope-isomorphism that constrains the linearization of bound morphemes. Accordingly, in Experiment 3, we have set out to conduct the morpheme-ordering experiment with Polish speakers. This ongoing work serves to showcase using the same pictographic language with multiple language populations as a means to control for prior language experience in ALL.

True of most Slavic languages, Polish exhibits rich fusional morphology, where case, number, and gender are fused into a single inflectional morpheme, also known as a polyexponential formative (Bickel & Nichols, 2013):

(3) a. *nauczyciel*
teacher-MASC.SG.NOM

b. *nauczyciel-e*
teacher-MASC.SG.ACC

c. *nauczyciel-a*
teacher-MASC.PL.NOM

d. *nauczyciel-i*
teacher-MASC.PL.ACC

(Adapted from Bielec, 1998)

Typologically distinct from agglutinative languages (the domain of Universal 39), languages with fused case + number morphemes are well-suited for testing the universality of an ordering bias that operates over bound morphemes, since monolingual speakers of such a language would have roughly equal familiarity with case and number marking. Furthermore, we would not expect them to have an ordering preference between concatenative case and number morphemes, at least not one that is informed by the structure of nominal morphology in their language.[16] If monolingual Polish speakers produce largely scope-isomorphic responses, this could provide even stronger evidence of

[16]Likewise, language users who are equally *unfamiliar* with case and number inflections—such as speakers of certain analytic languages (e.g., Thai, Vietnamese) or speakers of synthetic languages without the relevant nominal markers (e.g., Tu'un Savi, spoken in Oaxaca, Mexico)—would make strong participants for testing this hypothesis.

(a) (b)

Figure 3.10: Revised image stimuli for Experiment 3. The stimulus in (a) depicts three waitresses to foster plural (vs. numeral) interpretations of the number marker. The stimulus in (b) replaces the dollar sign with coins that had spilt out of the burglar's money sack (see Figure 3.7 for comparison); this was done to render the depiction more cross-culturally relevant.

this bias being universal. Conversely, if they do not favor scope-isomorphic orders, it would suggest that the bias is not universal or does not universally apply at the level of bound morphology.

In Experiment 3, we are teaching Polish speakers the same artificial language taught in Experiment 2, including the semi-iconic lexicon in Table 3.3 and VSO ~ VOS word order. While the experiment procedure has also remained the same, we made two modifications to the visual stimuli: First, we adjusted half of the images containing multiple subjects to depict three agents instead of two (Figure 3.10a). By doing so, we hope to encourage "plural" interpretations of the number marker. As discussed under Experiment 2, it was unclear whether the participants had analyzed the number marker as a plural affix or the free numeral *two*. Second, in the images featuring the burglar character, we removed the dollar sign (*$*) from their loot (Figure 3.10b) to make the stimuli more culturally relevant to Polish participants. (In Poland, the unit of currency is the złoty (*zł*).)

As in Experiments 1 and 2, we are using Prolific to recruit participants. All of the participants reside in Poland and self-identified as monolingual Polish speakers in Prolific's pre-screening filters. Nevertheless, since Prolific is an English-based platform, we do anticipate that they will have some familiarity with English. We thus translated all of the peripheral language-specific materials into Polish (i.e., the consent form, instructions screen, and questionnaire), first to prime Polish processing but also to signal to the participants that the study was designed for Polish speakers (i.e., that they don't do need to adjust their responses to what they think an English speaker would do).

That the participants will inevitably vary in their English familiarity also presents an opportunity

to model crosslinguistic influence from *secondary* languages. Based on the participants' responses to the language history portion of the post-experiment questionnaire, we plan to quantify their familiarity with (i) scope-isomorphic languages and (ii) languages with unequal case and number marking as continuous variables in our logistic regression model. If this improves the model's fit and the participants' preference for scope-isomorphic morphemes scales with their familiarity with scope-isomorphic languages or their familiarity with languages that have number but not case marking, this would be a strong indication that their prior language experience is factoring into their responses and, moreover, that their bias for scope-isomorphic structures would be weaker or nonexistent without this influence.

Equation 3.1 presents an initial sketch of quantifying a participant's exposure to their non-primary (i.e, non-Polish) languages. Moving forward, we will refer to this idea as an NPE (Non-Primary Exposure) score:

$$NPE = \max_{i \in L} \left(\frac{\max(S_i, C_i, R_i)}{\max(S_p, C_p, R_p)} \right) \tag{3.1}$$

Above, L is the set of relevant languages listed by the participant, while S_i, C_i, and R_i index the participant's respective scores for (i) speaking/signing, (ii) spoken/sign comprehension, and (iii) reading comprehension in language L_i. Essentially, the metric reflects the participant's self-reported *maximal* proficiency in an outside language, normalized by their self-reported maximal proficiency in Polish (taken over S_p, C_p, and R_p). Note that this metric uses proficiency as a loose proxy for experience.

Importantly, we can explore separate NPE scores for different identities of L. For instance, NPE_{iso} could operate over languages with scope-isomorphic case and number marking, NPE_{num} over languages with number but not case marking, and so forth. Building on this sketch, we can also explore *cumulative* effects for participants with familiarity in 3+ languages. Future work might also evaluate separate NPE scores for S, C, and R. More broadly, NPE scores might carve a path forward to quantifying L2–L*n* influence in ALL studies.

3.6 General discussion

The present study tests the viability of using pictographic writing systems in ALL studies. With continued validation, iconic artificial languages have the potential to facilitate crosslinguistic investigations by making it easier to design linguistic stimuli for diverse language populations. By not requiring a phone-to-meaning mapping, they also increase opportunities for reusing linguistic stimuli across different communities, leading to more tightly controlled comparisons across groups. If applied accordingly, iconic artificial languages can enhance scholarly mindfulness of language diversity and thus advance our understanding of linguistic cognition.

We have used scope-isomorphism as a vehicle to explore the effectiveness of pictographic languages at addressing questions of language universals and crosslinguistic influence. We first set out to reproduce ordering biases that were previously observed with English speakers. In Experiment 1, we performed a successful conceptual replication of the study by Martin et al. (2020), verifying that English speakers prefer modifier orders that comport with semantic scope relations. Likewise, Experiment 2 successfully reproduced the findings of Saldana et al. (2021), where English speakers favored scope-isomorphic orders of number and case markers (both prenominally and postnominally), though it remains unclear whether the speakers analyzed the markers as free or bound morphemes. Finally, we outlined a third, ongoing experiment, where we are using the *same* artificial language from Experiment 2 to examine ordering preferences amongst Polish speakers and, ultimately, to weigh in on the universality of the scope-isomorphism bias.

In the remainder of this discussion, we delve into the implications, limitations, and future directions of iconic artificial learning. We then return to the question of whether the attested bias for scope-isomorphism is the result of an innate feature of cognition or instead stems from crosslinguistic influence. Crucially, we highlight the importance of seeking out patterns of crosslinguistic influence, arguing to reframe the search for universals as a search for transfer effects.

3.6.1 Iconic artificial language learning

The iconic artificial languages in the present study have encompassed multiple lexical categories and linguistic levels: Across Experiments 1 and 2, we deployed pictographs ("glyphs") for nouns, verbs, adjectives, numerals, demonstratives, and case marking. In addition, we are continuing to

verify that pictographs can signify bound morphology, such as pluralization and case inflections (Experiments 2 and 3). Notably, the pictographic language in Experiments 2 was only semi-iconic, such that the glyphs for case and number were arbitrary, demonstrating the potential of mixing iconic and non-iconic glyphs in artificial lexicons. Overall, we have shown that these lexicons can be used to form complex phrases (Experiment 1) and full sentences (Experiments 2) in ALL studies.

Our hope is that pictographic stimuli can help broaden broaden language representation in the ALL literature. Future work should continue to vet using iconic artificial languages with typologically and orthographically diverse populations. In particular, these efforts should validate pictographic systems among individuals that are traditionally left out of ALL studies, such as children, deaf and Indigenous communities, and speakers of languages that don't have writing systems. Notably, pictographic stimuli may prove more accessible to pre- and non-literate individuals than written, phonologically-based stimuli, while being easier to learn than auditory stimuli. (They're also just more fun, which can come in handy with children.) Moreover, the relative "language neutrality" of pictographic lexicons may make them easier to design with multiple language groups in mind. This reusability feature can aid crosslinguistic comparisons and studies of crosslinguistic influence.

With that said, we should acknowledge that a truly neutral artificial language is likely impossible. Depending on the goal of a study, an iconic artificial language may need to be adjusted from one population to the next, particularly at the level of word order. For instance, since Thai NPs are already noun-initial, replicating a bias for scope-isomorphic modifiers with Thai speakers would call for reversing the word order taught in Experiment 1, training the participants on noun-final NPs (Martin et al., 2019), though the same pictographic lexicon could be used. Furthermore, *language neutral* does not mean *culture neutral*. Pictographic stimuli can still invoke culturally-grounded semiotics. Future work should thus be mindful of sociocultural differences when designing pictographic lexicons and might benefit from conducting norming studies within the desired communities beforehand. Similarly, language neutrality does not ensure the neutrality of companion image stimuli (e.g., the burglar depictions in Experiment 2) and animated feedback. Nevertheless, pictographic systems place us in a stronger position to study language diversity while making it easier to use the same lexicon with multiple populations.

Some may ponder that, because the artificial systems are pictographic, they can only tap into the general cognitive system, not bearing on language learning or processing *per se*. In addition

to being vague, this "Not Language" concern makes several assumptions. First, it assumes that the processes prompted in ALL experiments only involve language mechanisms when the stimuli are phonologically realized. It's unclear, however, why these mechanisms would be contingent on the artificial languages having a sound-to-meaning mapping. For instance, prelingually deaf children can learn to read spoken languages (see, e.g., Emmorey, 2020; Emmorey & Lee, 2021; Goldin-Meadow & Mayberry, 2001; Hoffmeister & Caldwell-Harris, 2014; Morford, Wilkinson, Villwock, Piñar, & Kroll, 2011; Musselman, 2000), especially when they are already skilled signers (e.g., Andrew, Hoshooley, & Joanisse, 2014; Chamberlain & Mayberry, 2000).[17] In these cases, the morphosyntactic units of the spoken language are acquired in the absence of acoustic reinforcement. Second, the "Not Language" concern assumes that an analogous, phonologically-realized artificial language would elicit language mechanisms where the pictographic language couldn't. This may be true to whatever extent a more canonically language-like system can *prime* language-like processing.[18] However, it seems to be a common offhand critique of ALL *in general* that they don't involve "real" language learning or processing, but instead pattern recognition and domain-general sequence processing.

Notably, neurolinguistic studies on second language acquisition have shown participants to display neural correlates of language when tested on artificial languages and grammars (for review, see Morgan-Short, 2020), including particular event-related potentials (e.g., N400 and P600 effect) and hemodynamic activity (e.g., in Broca's area). However, the neuroimaging in these studies typically occurred after lengthy training sessions—spanning hours, if not days or weeks—and only once the participants had reached a certain level of proficiency in the artificial system. Complicating things further is that many of the neural correlates of language *have* been connected to domain-general processing (see Osterhout, Kim, & Kuperberg, 2012, for discussion). For example, recent work has begun to characterize the P600 effect—first associated with grammatical processing (Osterhout & Holcomb, 1992)—as a domain-general consequence of encountering unexpected sequences (e.g., Christiansen, Conway, & Onnis, 2012; Hoen & Dominey, 2004; Sassenhagen & Fiebach, 2019).

Taken together, short single-session ALL experiments, such as the tasks in this chapter and those that typify typology-related investigations, may not be immersive enough for the artificial languages to evoke the neural hallmarks of language. However, this issue is not specific to pictographic

[17] Likewise, adults participating in ALL studies might benefit from their previous linguistic experience...

[18] Similarly, telling participants in advance that the system is a language might have similar priming effects.

stimuli but rather to any ALL experiment with a short runtime (relative to the complexity of the artificial system). Moreover, even if an artificial language, pictographic or otherwise, elicits the desired neural activity, we still might not be in a position to conclude that the observed mechanisms are language-specific. This nebulous terrain only underscores the continued need for research to explicate the neurobiological mechanisms underlying language, both in real and contrived contexts.

It is also worth emphasizing that an ALL study need not elicit language-specific processes to be informative about language and cognition. For instance, the present work is situated within a tradition of using artificial grammars and languages to detect general cognitive constraints that may have shaped how languages have evolved. For such research questions, perhaps the better methodological concern is whether the recruited community is appropriate for discerning the universality of the hypothesized constraint.[19] Towards this end, pictographic stimuli provide an avenue for studying more diverse language populations and exposing transfer effects in ALL experiments, even if the biases under investigation are domain-general.

3.6.2 Scope-isomorphism and mining for transfer effects

Prior work (Culbertson & Adger, 2014; Martin et al., 2020, 2019; Saldana et al., 2021), as well as the present study, has explored whether humans possess an innate bias towards morphosyntactic structures that mirror semantic scope relations. Although this preference has been recovered across numerous ALL experiments, a recurring question has emerged: Do the observed scope-isomorphic responses reflect a soft universal constraint or instead prior language experience? Put in gradable terms, how much of this bias is innate versus informed by familiarity with scope-isomorphic structures?

Martin et al. (2020) showed that English speakers prefer scope-isomorphic modifiers in artificial languages, at least with respect to demonstratives, numerals, and adjectives (cf. Experiment 1 of the current chapter; see also Culbertson & Adger, 2014 and Martin et al., 2019). As detailed by the authors, it's possible that this bias was driven by the participants' familiarity with the strict scope-isomorphism of English modifiers. Martin and colleagues have thus set out to conduct similar ALL tasks with speakers of Kîîtharaka,[20] a Bantu language where the modifier orders are not straightforwardly scope-isomorphic (N-Dem-Adj-Num), to see if they display similar preferences.

[19] Let's call this the "Appropriate Population" concern.

[20] Confirmed in personal communication.

At a more granular level, Saldana et al. (2021) found that English and Japanese speakers prefer scope-isomorphic orders of number- and case-marking morphemes (cf. Experiment 2 in the present study). That both groups of speakers displayed this preference, despite their unequal familiarity with the two markers, led the authors to suggest this bias as a "universal feature of cognition" that holds independent of boundedness. However, earlier in the chapter, we considered whether this conclusion may be premature. In particular, we noted the close-knit relationship between noun stems and number morphemes in English, which is contrasted by the special peripheral status held by Japanese case particles (Hattori, 1950; Kageyama, 2020; Tsujimura, 2013; Vance, 1993). Therefore, familiarity with the distribution of either marker in either language could have contributed to the scope-isomorphic responses in Saldana et al. (2021) and the present study.

The argument that scope-isomorphism is a universal feature would, again, be strengthened by demonstrating scope-isomorphic preferences across more diverse, theoretically-motivated groups of language users (Blasi, Henrich, et al., 2022; Majid, 2023). For instance, languages with fused nominal inflections (e.g., Finnish, Polish, Yaqui) or without the relevant markers (e.g., Thai, Tu'un Savi, Vietnamese) are well-suited for examining innate ordering biases in bound morphology. Considering modifiers and morphemes in tandem, the ongoing work with Kîîtharaka and Polish speakers could yield different but complementary insights. In the modifier-ordering study with Kîîtharaka speakers, if the scope-isomorphism bias doesn't hold, it may still be that it is universal, but that the bias has been dampened or reversed by the speakers' lifelong experiences in a language with non-isomorphic modifiers. In contrast, if the bias doesn't hold in a study on morpheme orders with Polish speakers, it would suggest that the bias is *not* universal or, at least, does not universally apply over certain morphemes.

What would be especially interesting is if Kîîtharaka speakers do not favor scope-isomorphic orders for modifiers but Polish speakers do for morphemes. Together, it could suggest that the bias is indeed universal but modulated by language experience. After all, what is under investigation in these studies is likely a *singular* bias (whatever this means neurobiologically) that can operate over multiple lexical categories and levels of expression. Moreover, this bias is clearly violable and can interact with other constraints (including those that are culturally transmitted) that may apply only at specific linguistic levels. For instance, Saldana et al. (2021) showed that a preference for scope-isomorphic morphemes could be overridden in the presence of allomorphy where only the

case marker was highly dependent on the noun stem. Recall also that languages like Chintang allow some inflectional morphemes to appear in free variation (Bickel et al., 2007).

Now to throw a wrench in the gears: In the abstract, languages that entail no scope-isomorphic ordering preferences *at any linguistic level* would be the ideal test bed for verifying the universality of a scope-isomorphism bias. If Polish speakers do favor scope-isomorphic morphemes, it is still possible that this bias is one acquired through experience—namely, since Polish does have scope-isomorphic *modifiers*. Perhaps problematically for testing the universality of scope-isomorphism, English, Japanese, and Polish are all Dem-Num-Adj-N languages (Figure 3.1a; Dryer 2018). All else being equal, an interesting question is whether speakers who have been entrenched in a language with scope-isomorphic modifiers are more likely to order artificial morphemes scope-isomorphically, and vice versa. In general, do languages with more scope-conforming structures lead to stronger preferences for scope-isomorphic expressions across different linguistics levels?

Crucially, it is only by juxtaposing behaviors across diverse language communities that we can begin to disentangle experience from genetics and get at the (non-)universality of a scope-isomorphism bias. In particular, it behooves us to explore how this bias might vary in strength across different populations and how this variation correlates with the typological makeup of the languages (e.g., Martin et al., 2019). For instance, future work should contrast languages that differ in *where* scope-isomorphism appears and the overall *degree* to which the languages encode scope-isomorphic structures. Moreover, these efforts should examine ordering preferences at multiple linguistic levels (e.g., modifiers *and* morphemes) within the same sample or population. Notably, iconic ALL can help with these endeavors.

In short, it remains unknown whether scope-isomorphism is a universal feature of cognition that shapes language. We need to look beyond English, Japanese, and Polish speakers to answer this question. Importantly, a bias for scope-isomorphic structures may be attenuated—if not completely informed—by language experience, leading to variation in how it surfaces crosslinguistically (perhaps correlating with the strength of the bias) and where in languages it appears most prominently (e.g., operating over words vs. bound morphemes). If there is an innate, general bias for scope-isomorphic expressions, we might expect it to influence language structure at different linguistic levels (e.g., morphology, syntax), interacting with other constraints along the way (e.g, morphophonological dependencies; historical artifacts, by way of language experience). On the other hand, if humans can

develop a structural bias for scope-isomorphism entirely from language experience, it's a fascinating question how we can generalize such a bias from one linguistic level to the next. What cognitive mechanisms would make this possible?

Beyond scope-isomorphism, future work would benefit from interrogating proposed universals, treating them as "transfer questions" and actively seeking out evidence that would disprove rather than evidence the universality of common patterns. While demonstrating the universality of a bias may be infeasible in many cases due to transfer effects, we can make meaningful progress on understanding crosslinguistic influence and, by extension, the pathways through which we come to internalize language experience. Even when constraints come out of a combination of biology and experience, a complex tapestry of transfer effects can be more informative for human language processing than the stale fact that we are genetically predisposed towards particular behaviors. Thus, contrasting transfer behaviors across diverse groups can enable us to arrive at a more holistic understanding of the cognitive mechanisms underlying language. When, after a robust and careful search, we can't find an experiential explanation for a prevalent linguistic pattern, perhaps then we will have the strongest evidence of a universal constraint on language.

3.7 Conclusion

We have shown the viability of replacing phonologically-realized writing systems with iconic, non-orthographic symbols. Through successful conceptual replications of experiments by Martin et al. (2020) and Saldana et al. (2021), we verified that English speakers prefer modifier and morpheme orders that comport with semantic scope relations ("scope-isomorphism")—even when learning artificial languages composed entirely of pictographs. We further used scope-isomorphism as a case study to explore confounds posed by language experience in the pursuit of universal constraints. While it's fully possible that our participants' ordering choices reflect a universal bias for scope-isomorphism, it's also possible that this bias is informed or at least attenuated by the degree to which their existing linguistic knowledge encodes scope-isomorphic structures.

Crucially, iconic artificial languages make it easier to design linguistic stimuli for diverse language populations, and can thereby facilitate investigations of universality and crosslinguistic influence. Given the relative language neutrality of pictographic stimuli, it may be possible to use the same artificial lexicon with speaking and signing populations, typologically and orthographically diverse

languages, children and adults, as well as monolinguals, bilinguals, and multilingual individuals. Such studies should delve into how these communities both *resemble* and *vary* from one another in their language learning and processing.

4 Probing mBERT: Variation in crosslingual morphosyntactic representations

4.1 Introduction

The question of *nature vs. nurture* isn't specific to humans. In the rapidly evolving field of natural language processing (NLP), engineers and researchers seek to develop language technologies that leverage statistical learning. Every such pursuit requires specifying a model architecture (explicit and implicit decisions regarding the model's nature) and amassing training data (what will "nurture" the model). In recent years, large language models (LLMs) have dominated this landscape (e.g., *ELMo*, Peters et al., 2018; *BERT*, Devlin et al., 2019; *RoBERTa*, Y. Liu et al., 2019; *GPT 1–4*, Radford, Narasimhan, Salimans, & Sutskever, 2018; Radford et al., 2019; Brown et al., 2020; OpenAI, 2023), leading to groundbreaking advances in artificial intelligence. In turn, the "black box" nature of LLMs has revived interest in what can be learned about human languages from data alone, prompting new comparisons of human and machine linguistic intelligence. We contribute to these efforts by proposing a method for probing the morphosyntactic capacity of *multilingual* LLMs.

4.1.1 Contemporary language modeling

The architectures of LLMs pull from a family of supervised, biologically-inspired machine learning algorithms called *artificial neural networks*. Conceptually, neural networks are composed of layers of "artificial neurons". Each neuron computes the sum over an input vector $x = [x_1, ..., x_n]$, which it scales using an assigned weight; this weighted sum is then transformed by a non-linear "activation" function, producing a single output value. The weights of a neural network, also known as *parameters*, are randomly initialized, then iteratively adapted during the model's training regime to perform a specific task (e.g., labeling words for their lexical categories or translating sentences

from one language to another). If a neural network has multiple hidden layers between its input and output layers, training the model is said to involve "deep learning".

Inspired by their *n*-gram predecessors, LLMs are neural networks that have been trained to perform cloze-style tasks, such as predicting a missing word or subword unit in a sentence (e.g., that *roses* can fill the blank in "Moses supposes his toeses are ____"). Due to the cloze-style training objective, the intermediate layers of LLMs learn to produce context-dependent numeric vector representations of the (sub)word tokens, also called *contextualized word embeddings* (for review, see Smith, 2020). Each embedding captures local information (e.g, the meaning and linguistic features of the token), as well as information about how the token relates to other (sub)words in the context (e.g., modification, agreement, and long-distance dependencies). Crucially, LLMs are intended as *pre-trained* models, where they are typically "fine-tuned" and collaged into broader systems to perform new tasks, such as machine translation, question answering, and human-machine dialogue. These downstream systems benefit from *transfer learning*, where they can tap into the rich information already captured by the LLMs. Numerous studies have shown that initializing a system with a pre-trained LLM offers considerable performance gains over training an analogous system from scratch (e.g., Devlin et al., 2019; Y. Liu et al., 2019; Peters et al., 2018; Radford et al., 2018, 2019). Accordingly, pre-trained LLMs underpin a growing number of cutting-edge language technologies.

Fueling their uptake is that LLMs do not require *annotated* training data, let alone data that is annotated by an expert. Instead, the language modeling objective seemingly allows these models to recover linguistic information from raw text that historically needed to be annotated first (e.g., lexical categories, syntactic parses, and semantic roles; Tenney et al., 2019), in addition to general information about the world. However, this doesn't come cheap: Modern LLMs require increasingly exorbitant amounts of training data (see Bender et al., 2021, for discussion).

In this way, linguistic diversity continues to pose significant challenges to language modeling. In particular, languages vary substantially in the amount and quality of training data available for them, resulting in *high-* and *low-resource* languages (Joshi et al., 2020). Moreover, typological diversity means that design decisions that work well for some languages don't always extend well to others (Bender, 2011). Such factors, in combination with the NLP field's preoccupation with a handful of high-resource languages (e.g., English, German, and Chinese), has widened the technology gaps

between language communities (cf. Bender et al., 2021).

These disparities have led to the rise of *multilingual* LLMs that have been trained to make cloze-style predictions for multiple languages (for review, see Doddapaneni, Ramesh, Khapra, Kunchukuttan, & Kumar, 2021). Notably, multilingual language modeling does not require aligned corpora (i.e., parallel sentences for each language represented in the training data); therefore, in an attempt to prevent high-resource languages from overpowering low-resource languages in the model, the training data will sometimes downsample high-resource languages and upsample low-resource languages. At its foundation, work on multilingual language modeling aims for the models to learn and exploit similarities across languages without direct supervision. This implicit goal is accomplished through the individual weights of the model coming to encode information that is applicable to multiple languages (e.g., morphosyntactic features and "commonsense" reasoning). In the ideal, this *crosslingual transferability* would improve the linguistic functionality for all of the languages represented in the model, achieving performance parity between the high- and low-resource languages.

It remains a tricky and persistent feat, however, to design LLMs that can maximally identify and leverage connections between languages. Whether a model will recognize crosslinguistic similarities is highly dependent on the particular languages involved, the quality of the training data (e.g., the balance of languages; the domain-specificity versus generality of the data), the model's architecture (e.g., the number and configuration of parameters; the mathematical operations performed), as well as some randomness. As a consequence, multilingual LLMs can exhibit considerable variation in how they learn, represent, and process linguistic features.

4.1.2 Morphological complexity

Morphologically rich languages, in particular, present unique challenges to multilingual language modeling. These languages typically exhibit complex agreement patterns and their high diversity of inflected forms can lead to sparse examples of vocabulary words in training data, even in large corpora (Blevins & Zettlemoyer, 2019; Gerz et al., 2018). It is therefore worthwhile to explore how LLMs, which serve as the foundation of many state-of-the-art systems, handle the morphological complexity of diverse languages.

Morphosyntactic features of natural languages bear meaningful information that is useful for

downstream tasks, such as machine translation, question answering, and language generation. Hence, adding morphological supervision through multi-task training regimes (Blevins & Zettlemoyer, 2019) or morphologically-informed tokenization (Klein & Tsarfaty, 2020; Park et al., 2021) can improve the quality of multilingual language models. Nonetheless, recent work has shown that language models trained without explicit morphological supervision can still produce useful representations that capture morphosyntactic phenomena (e.g., Bacon & Regier, 2019; Dufter & Schütze, 2020; Pires, Schlinger, & Garrette, 2019).

More broadly, numerous studies have sought to study the linguistic properties captured by language models (e.g., Bacon & Regier, 2019; Chi et al., 2020; Conneau et al., 2018; Futrell & Levy, 2019; Gulordava et al., 2018; Hewitt & Manning, 2019; Hupkes et al., 2018; Jawahar et al., 2019; N. F. Liu et al., 2019; Marvin & Linzen, 2018; Tenney et al., 2019; Zhang & Bowman, 2018). In the morphology domain, the LINSPECTOR suite by Şahin, Vania, Kuznetsov, and Gurevych (2020) probes 24 languages via 15 linguistic tasks, including multiple tasks to identify morphological features. In a similar vein, Edmiston (2020) uses several morphological prediction tasks to inspect embeddings from five monolingual Transformer-based language models, focusing exclusively on Indo-European languages.

4.1.3 The present study

To enhance investigations of the morphosyntactic and crosslingual abilities of multilingual language models, we propose using a multilabel probing task to assess the morphosyntactic representations of word embeddings. This work is premised on the intuition that, if a simple model (a "probe") can easily extract linguistic properties from embeddings, this indicates that the language model has learned to encode those features in some fashion Conneau et al. (2018); Hupkes et al. (2018); N. F. Liu et al. (2019). The probing paradigm proposed in this study builds on earlier works, such as Şahin et al. (2020) and Edmiston (2020), by consolidating multiple morphosyntactic feature prediction under a single task that leads more naturally to the study of feature co-occurrence patterns. We further show how multilabel probing can shed light on the morphosyntactic representations of multilingual language models, both holistically and at the level of individual features.

Our contributions are twofold:

1. We introduce a paradigm for analyzing the morphosyntactic features captured by language

models. We demonstrate this methodology with Multilingual BERT (henceforth, *mBERT*; Devlin et al., 2019), training probes for seven typologically diverse languages: Afrikaans, Croatian, Finnish, Hebrew, Korean, Spanish, and Turkish.

2. By evaluating the probes on six "held-out" languages—Arabic, Chinese, Marathi, Slovenian, Tagalog, and Yorùbá—we show how this paradigm can be used in a zero-shot manner to study crosslingual representations, illuminating the features that multilingual language models represent *similarly* crosslinguistically.

This rest of this chapter is structured accordingly: We first introduce our multilabel approach to morphosyntactic probing, followed by the datasets and models we use to probe mBERT. To demonstrate our probing paradigm, we then conduct a series of **monolingual experiments**, training and evaluating separate probes for each of the seven aforementioned languages, and providing an example feature-level analysis of Hebrew determiners. Next, in a set of **multilingual experiments**, we train probes on an aggregated subset of the languages, showing that the multilingual probes *for the most part* yield comparable insights to the monolingual probes and delving into the implications of their subtle differences in performance. Finally, in a set of **crosslingual experiments**, we evaluate how the monolingual and multilingual probes handle the six held-out languages, finding considerable variation in how mBERT captures the crosslinguistic status of features. We conclude the chapter with a discussion, considering the ethical dimensions of the present study before briefly entertaining the implications of our crosslingual findings for *human* language processing.

4.2 Multilabel morphosyntactic probing

We propose using multilabel morphosyntactic tagging to assess the morphosyntactic representations of neural LMs. In this diagnostic task, we hold contextualized word embeddings constant, then train linear classifiers on top of them (cf. Hupkes et al., 2018; N. F. Liu et al., 2019) to perform morphosyntactic tagging. In its objective, morphosyntactic tagging resembles the second SIGMORPHON 2019 shared task, which called for labeling words in a sentence with their morphosyntactic descriptions McCarthy et al. (2019).

It is easy to imagine doing morphosyntactic tagging in a traditional multiclass fashion, where we train separate probes to identify different features, such as part of speech (POS), gender, or

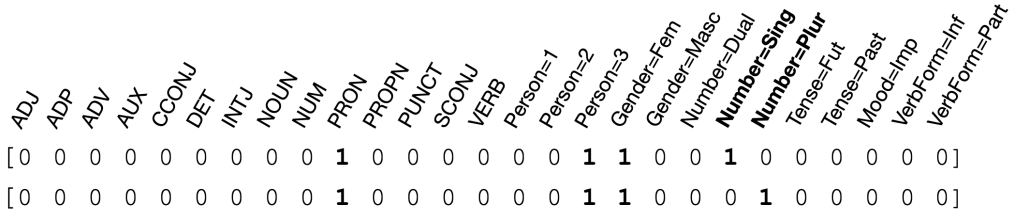


Figure 4.1: Hypothetical multi-hot encoded vectors for the Hebrew *3.sing.fem* pronoun הִיא *hi* (top) and *3.plur.fem* pronoun הֵן *hen* (bottom). The two vectors differ only with respect to the two cells indicating "singular" and "plural", reflecting their otherwise similar grammatical characteristics.

number (cf. Edmiston, 2020; Şahin et al., 2020). However, this style of probing is more likely to prompt narrow analyses that consider morphological properties in isolation. Alternatively, we could train a single probe to extract complex labels like `3rd.plur.masc.past.verb` and `def.sing.masc.noun`. Thus, each word would have a single correct label and a final softmax layer would output the probability of each class being the correct one. However, a drawback to this approach is that, depending on the number of properties we would like to identify, this can result in a combinatoric nightmare 👻, with few training examples per class.

To overcome these limitations, we frame morphosyntactic tagging as a word-level multilabel task, allowing for a token to receive multiple feature labels (e.g., both `Number=Sing` and `Person=1`) that are multi-hot encoded, where a cell is 1 if a specific feature label applies to a word and 0 otherwise. Such a paradigm allows us to encode features with multiple or ambiguous values (e.g., `Gender=Fem,Masc`; a.k.a. multi-valued features) and enables a closer inspection of learnt agreement and feature co-occurrence patterns. Figure 4.1 illustrates hypothetical gold vectors for two Hebrew pronouns that differ only in number.

4.2.1 Notation and nomenclature

We define a feature label as the conjunction of a linguistic feature (e.g., *number*) and a possible realized value of that feature (e.g., *singular*), as depicted in Figure 4.2. Multiple feature labels can correspond to the same feature (e.g., `Number=Sing` and `Number=Plur`).[1] F is the set of feature labels $\{f_1, \ldots, f_{|F|}\}$ that we use to identify morphosyntactic properties from word embeddings. (It is due to the inclusion of parts of speech that we refer to the task as "morpho*syntactic* tagging".)

[1] We drop `POS=` from part-of-speech labels, conforming to UPOS notation (e.g., `NOUN` instead of `POS=NOUN`).

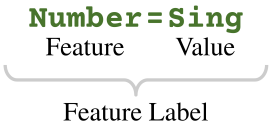


Figure 4.2: Anatomy of a feature label.

Assuming a vocabulary of word types V, let $\mathbf{s} = s_1 \dots s_{|\mathbf{s}|}$ denote a specific sentence and r_i denote the contextualized representation of each token s_i, such that $s_i \in V$. The inputs to the probe are therefore the embeddings $r_i \in \mathbb{R}^d$. In the multilabel morphosyntactic tagging task, we define the target output of each embedding r_i as a multi-hot encoded vector $\mathbf{y}^i = y^i_1 \dots y^i_{|F|}$, where F is the aforementioned set of feature labels. We encode y^i_j as 1 if the feature label $f_j \in F$ describes the token s_i and 0 otherwise.

4.2.2 Multilabel evaluation

The multilabel paradigm lends itself well to analyzing features both holistically and at a granular level. We can analyze individual features by first counting the true positives (TP), false positives (FP), and false negatives (FN) for each feature label f, then calculating each feature's respective precision (P), recall (R), and F_1 scores:

$$\mathrm{P}_f = \frac{TP_f}{TP_f + FP_f}$$

$$\mathrm{R}_f = \frac{TP_f}{TP_f + FN_f} \tag{4.1}$$

$$\mathrm{F}_{1f} = 2 \times \frac{\mathrm{P}_f \times \mathrm{R}_f}{\mathrm{P}_f + \mathrm{R}_f}$$

Furthermore, we can glean the overall or *micro-averaged* performance of a probe by tallying TPs, FPs, and FNs across the features:

$$\mathrm{P}_{micro} = \frac{\sum_f TP_f}{\sum_f TP_f + FP_f}$$

$$\mathrm{R}_{micro} = \frac{\sum_f TP_f}{\sum_f TP_f + FN_f} \tag{4.2}$$

$$\mathrm{F}_{1micro} = 2 \times \frac{\mathrm{P}_{micro} \times \mathrm{R}_{micro}}{\mathrm{P}_{micro} + \mathrm{R}_{micro}}$$

It is also possible to take the *macro-average* over feature label scores (e.g., averaging the F_1 scores across the labels). For simplicity, we focus on micro-averages in this study, since they are more impervious to one-off failures to extract a specific feature label.

4.3 Experiment setup

We demonstrate multilabel morphosyntactic probing with Multilingual BERT (mBERT; Devlin et al., 2019), using morphologically annotated corpora from Universal Dependencies (UD; Nivre et al., 2016, 2020).[2]

4.3.1 Crosslinguistic data

Our target vectors draw on UD part of speech and morphological feature annotations. In a set of monolingual experiments, we trained separate probes to predict morphosyntactic features from corpora for seven languages of varying morphological complexity: Afrikaans (AfriBooms; cf. Dirix, Augustinus, van Niekerk, & Van Eynde, 2017), Croatian (SET; cf. Agić & Ljubešić, 2015), Finnish (TDT; cf. Haverinen et al., 2014; Pyysalo, Kanerva, Missilä, Laippala, & Ginter, 2015), Hebrew (HTB; cf. McDonald et al., 2013; Sadde, Seker, & Tsarfaty, 2018; Tsarfaty, 2013), Korean (PUD; cf. Zeman et al., 2017), Spanish (AnCora; cf. Alonso & Zeman, 2016), and Turkish (IMST; cf. Sulubacak et al., 2016; Türk et al., 2019; Tyers, Washington, Çöltekin, & Makazhanov, 2017). With the exception of the Korean data, all of the corpora came pre-split into training, validation, and

[2] https://universaldependencies.org

Language	Genus	$\|F\|$	Train		Dev		Test	
			Sentences	Tokens	Sentences	Tokens	Sentences	Tokens
Afrikaans	Germanic	53	800	21,160	194	5,317	425	10,065
Croatian	Slavic	66	800	17,811	960	22,292	1,136	24,260
Finnish	Finnic	89	800	10,786	1,363	18,311	1,553	21,069
Hebrew	Semitic	53	800	16,061	484	8,358	491	8,829
Korean	Korean	35	800	13,177	100	1,679	100	1,728
Spanish	Romance	63	800	24,345	1,654	52,161	1,719	52,429
Turkish	Turkic	64	800	8,244	983	9,768	981	9,794
Multilingual	n/a	72	4,800	98,297	5,638	116,207	n/a	n/a

Table 4.1: Composition of the training and evaluation data for the monolingual and multilingual probes.

test sets. We performed an 80-10-10 split on the 1,000-sentence Korean PUD corpus. To throttle the probes' training data (cf. Zhang & Bowman, 2018), we reduced the other training sets to 800 sentences as well.

Next, in a set of multilingual experiments, we trained probes on a shuffled combination of the training sentences from the monolingual probes. However, we excluded the Korean dataset from this analysis, due to the lack of documentation on its construction. The monolingual and multilingual datasets are summarized in Table 4.1.

Finally, in a set of crosslingual transfer experiments, we evaluated the monolingual and multilingual probes on six held-out languages: Arabic (PADT; cf. Hajič et al., 2009; Smrž et al., 2008; Smrž, Šnaidauf, & Zemánek, 2002), Chinese (PUD; cf. Zeman et al., 2017), Marathi (UFAL; cf. Ravishankar, 2017), Slovenian (SST; cf. Dobrovoljc & Nivre, 2016), Tagalog (TRG), and Yorùbá (YTB; cf. Ishola & Zeman, 2020). This data is summarized in Table 4.2. To clarify, mBERT *was* pre-trained on these languages; we consider them "held-out" in that we never train probes to extract linguistic properties from these corpora (i.e., the experiments are zero-shot).

All of the probes were trained to extract multiple features, such as POS, number, gender, case, and tense, as well as language-specific features, such as Finnish infinitive forms. Since the languages vary in their linguistic properties, we used different label sets for each language and a semi-aggregated set for the multilingual probes. Across our experiments, we extracted 166 different feature labels in total, as listed in Appendix A.

Language	Genus	Test	
		Sentences	Tokens
Arabic	Semitic	675	24,195
Chinese	Chinese	1,000	21,415
Korean	Korean	1,000	16,584
Marathi	Indic	47	376
Slovenian	Slavic	995	9,880
Tagalog	Greater Central Philippine	55	292
Yorùbá	Defoid	318	8,198

Table 4.2: Composition of the "held-out" language data in the crosslingual transfer experiments.

4.3.2 Models, training, and implementation

Linear classifiers For our experiments, we instantiated a "BERT-Base, Multilingual Cased" model using HuggingFace's *Transformers* library Wolf et al. (2019). This BERT variant contains 110M parameters across 12 Transformer layers, each with 12 attention heads and a hidden size of 768. The model was pre-trained on Wikipedia dumps from 104 languages. The authors over-sampled the smaller Wikipedia corpora to create a more crosslinguistic vocabulary, consisting of 100K wordpieces.

We froze mBERT and trained linear classifiers on top of embeddings produced by mBERT's initial embedding layer and its successive Transformer layers (cf. Hupkes et al., 2018; N. F. Liu et al., 2019). Preliminary experiments showed that the even-numbered layers (mBERT-0, mBERT-2, mBERT-4, etc.) faithfully captured the layer-by-layer trends across mBERT, so we opted to cut down on computation by focusing exclusively on these layers. The classifiers used sigmoid activation and were trained with mean binary cross-entropy loss to perform the multilabel tagging task. We trained each classifier for 50 epochs, selecting the model from the epoch that achieved the best validation loss. Courtesy of PyTorch Paszke et al. (2019), the classifiers were optimized using Adam (learning rate = 0.001, β_1=0.9, β_2=0.999, ϵ=1e-08; Kingma & Ba, 2015). No dropout was used.

Word-level predictions Despite mBERT's word-piece vocabulary, we performed word-level predictions of morphosyntactic properties. To do so, we first passed the raw corpus sentences through mBERT, then aggregated the contextualized word embeddings on a word-by-word basis. In small exploratory experiments, we found that summing the subword embeddings performed the best;

we thus used this aggregation strategy throughout our experiments. Notably, summing the subword representations achieved comparable F_1 scores but better "selectivity" than taking their average (cf. Hewitt & Liang, 2019, and Section 4.3.3). The summation and averaging strategies also performed better than representing each word by the embedding for its word-initial or word-final word piece.

The UD corpora include decompositions of multiword tokens and separate annotations for their respective components. To keep the input to the probes faithful to naturalistic text, we embedded the multiword tokens themselves, but aggregated the feature labels from their components (e.g., the Hebrew multiword token הספר *hasefer* 'the book' is marked as both a determiner and noun).

Caching and batching Prior to training, we cached the aggregated word representations; these stored embeddings then served as the inputs to the probes. This was done in lieu of passing a batch of input sentences through mBERT and doing the aggregation on the fly at each training step. Since the probes themselves are simple linear layers and therefore non-contextual, we were able to batch the embeddings at the token level: We dispensed with the sequence length dimension and skipped padding. In all of the experiments, we opted for a batch size of 512 tokens (i.e., the batches had a dimensionality of 512×768). This "cache and batch" approach allowed each monolingual probe to train in ∼1 minute and each multilingual probe in ∼4 minutes on a Tesla K80 GPU. (In fact, it took longer to embed the initial inputs and to obtain test predictions than it did to train the probes.)

4.3.3 Vying for control

Prior probing work has sought to curtail the amount probes memorize about linguistic tasks to ensure that they *reflect* information available in their input embeddings. In other words, probes should be extractive rather than learnèd themselves. Efforts to minimize memorization have included reducing the training data to probes Zhang and Bowman (2018) and limiting probe complexity, such as through dropout (e.g., Belinkov, Durrani, Dalvi, Sajjad, & Glass, 2017; Belinkov, Màrquez, et al., 2017; Şahin et al., 2020) and the use of simpler architectures (e.g., a linear layer instead of a multilayer perceptron, as in Alain and Bengio 2018 and N. F. Liu et al. 2019).

To guide the design and interpretation of probes, Hewitt and Liang (2019) proposed supplementing diagnostic tasks with *control tasks*, where a probe is trained to predict random outputs within the same output space as the diagnostic task, given the same embeddings. If the probe performs well

on the control task, they caution that it has the capacity to memorize the linguistic features under consideration; conversely, if the probe does well on the diagnostic task but poorly on the control task, then it is a reliable diagnostic of linguistic representations in the embeddings (though see Pimentel, Saphra, Williams, and Cotterell, 2020, and Pimentel, Valvoda, et al., 2020, for interesting discussions). Hewitt and Liang operationalize this comparison as *selectivity*, the difference in performance on the diagnostic and control tasks. The greater the selectivity, the more the probe is said to "express" the information encoded in its input.

Following Hewitt and Liang (2019), we constructed a control task to complement the multilabel tagging task, whereby each word type in the task vocabulary was assigned a multi-hot output vector that was randomly generated according to the true distribution of feature labels in the training data. Deviating from Hewitt and Liang's notation, we generated a control output vector $\mathbf{c^i}$ for each word type $v_i \in V$, such that $\mathbf{c}^i = c^i_1 \ldots c^i_{|F|}$, where c^i_j was sampled from the true distribution of feature f_j in the training data.[3] For instance, if f_j was a feature of 4% of the tokens in the training set, then c^i_j has a 0.04 probability of being 1 for any word type v_i (or, conversely, a 0.96 probability of being 0). To help ensure the presence of controlled counterparts for low-frequency feature labels, each feature label had a minimum probability threshold of 0.001. For each language, the probes for the various mBERT layers were trained to predict the same set of random output vectors.

Note also that UD corpora assume an open vocabulary; many of the word types in the validation and test sets do not appear during training. This allows us to evaluate the effectiveness of the probes on out-of-vocabulary words. If the probes truly extract features rather than memorize the task, we would expect them to perform similarly on in-vocabulary and out-of-vocabulary words. We perform such an analysis in Section 4.5.1.

4.4 Monolingual experiments

In a set of monolingual experiments, we trained and evaluated individual diagnostic probes on Afrikaans, Croatian, Finnish, Hebrew, Korean, Spanish, and Turkish, given representations from the even-numbered mBERT layers. Their micro-averaged F_1 scores are conveyed in Figure 4.3, along with their results on the analogous control tasks.

[3] V is based on the word types across the training, validation, and test sets, since UD corpora use an open vocabulary.

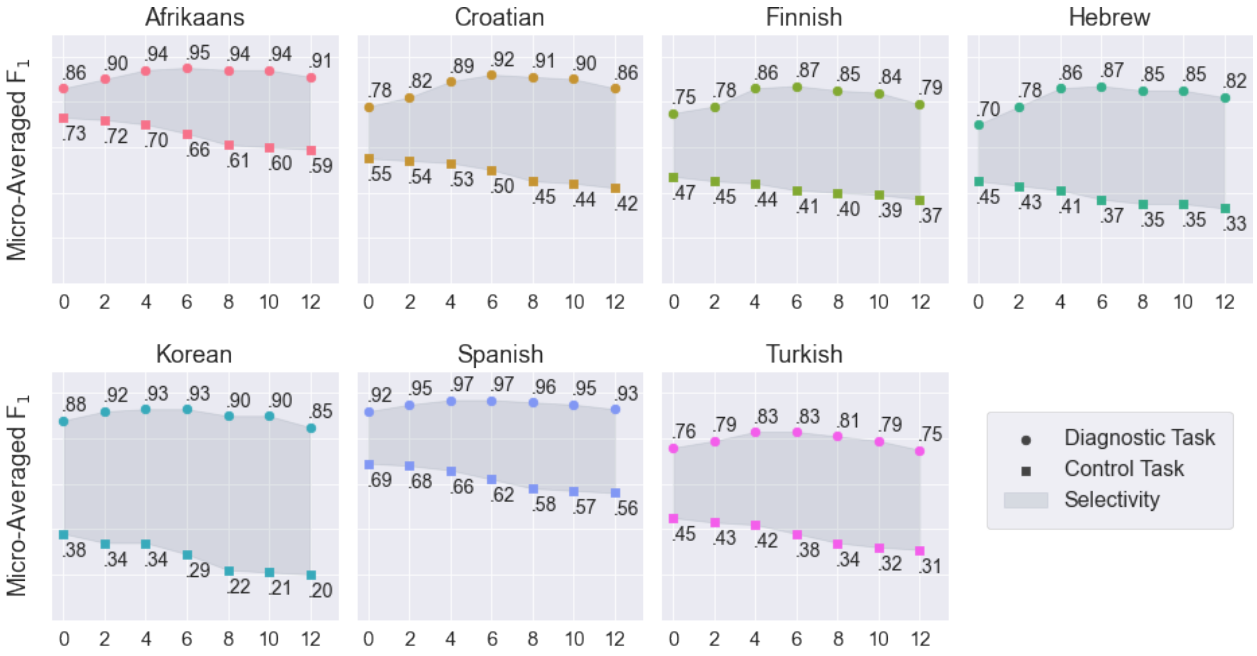


Figure 4.3: Micro-averaged F_1 results from the monolingual probes on the diagnostic and control tasks. The *x*-axes indicate the mBERT layer.

4.4.1 Monolingual performance at a glance

The micro-averaged F_1 scores confirm that mBERT renders many morphosyntactic properties easily extractable, with the best performing probes for each language achieving scores between 0.83 and 0.97. We find that mBERT-6 scored the highest across the languages. This is consistent with prior work that has shown English BERT's interior layers to perform best on similar linguistic tasks N. F. Liu et al. (2019); Tenney et al. (2019). Once mBERT has encoded morphologically relevant information, it seems that probe performance steadily declines as the topmost layers gear up for cloze predictions.

Notably, the Afrikaans and Spanish probes performed the best and the Turkish probes the worst. It is tempting to conclude that 'mBERT knows Afrikaans and Spanish better than Turkish'. However, we should refrain from comparing global probe performance across languages, as each language differed in the sets of features that were extracted. Furthermore, although each of the probes were trained on 800 sentences, they were ultimately trained on varying numbers of tokens. It may be that the Afrikaans and Spanish probes performed the best because they had the largest training sets *token*-wise, whereas Turkish had the smallest training set and lowest F_1 scores.

Language	Train Size	Selectivity	mBERT-Layer
Spanish	24,345	0.38	8, 10
Afrikaans	21,160	0.34	10
Croatian	17,811	0.46	8, 10
Hebrew	16,061	0.50	6, 8, 10
Korean	13,177	0.69	10
Finnish	10,786	0.46	6
Turkish	9,768	0.47	8, 10

Table 4.3: Training set size (in tokens) and the highest monolingual selectivity score achieved per language.

4.4.2 Monolingual selectivity

While the diagnostic probes drastically outperformed their controlled counterparts, we do see a trend of selectivity improving with the deeper mBERT layers in Figure 4.3. This reinforces the findings of Hewitt and Liang (2019), who posit that classifiers trained on top of lower layers are better equipped to memorize input-output mappings, due to their proximity to the initial vocabulary representations of the embedding layer. Nevertheless, the high selectivity scores across the probes show that a multilabel probing classifier offers a promising diagnostic of morphosyntactic representations.

From a crosslinguistic standpoint, it is interesting that the probes for Afrikaans—the one morphologically *impoverished* language in the bunch—exhibited the worst selectivity. This suggests that, perhaps, it is easier for probes to memorize mappings for analytic languages (i.e., languages that lack rich inflectional systems). However, as the the Afrikaans probes were trained on the second largest number of tokens, they may have had more opportunity to memorize the control task. (Similarly, the Spanish probes, which had the largest training set, displayed the second best performance on the control task.)

With the exception of Korean, we see that selectivity generally decreases as the amount of training data increases (see Table 4.3). Although this effect may seem somewhat surprising, given that each probe was trained on only 800 sentences, it is still consistent with the idea that more training data will lead to more memorization (Zhang & Bowman, 2018) and higher F_1 scores on both the diagnostic and control tasks and, therefore, lower selectivity (cf. Hewitt & Liang, 2019): Our findings suggest that token-level probes are highly sensitive to their number of training tokens, and that the traditional method of controlling for dataset size by limiting the number of

sentences may be insufficient for fair comparisons between probes. Future work should explore controlling for the number of *tokens* that appear during training and its effects on selectivity. This is made straightforward by caching the embeddings and de-sequencing the inputs, as described in Section 4.3.2.

4.4.3 Case study: Hebrew covert determiners

The micro-averaged scores in Figure 4.3 show that mBERT has indeed learned *some* linguistic system or portion thereof. However, these scores do not give much insight into which aspects of morphosyntax mBERT has come to represent, the interplay between these properties, nor how much mBERT varies in capturing each feature value. Crucially, a key strength of multilabel probing is that it makes it easy to mine fine-grained morphosyntactic observations that implicate multiple features. In this section, we present such an analysis with Hebrew determiners, inspired by Klein and Tsarfaty (2020). We focus on the predictions from mBERT-6, since it displayed the highest F_1 and selectivity scores out of the Hebrew probes.

Ambiguous orthographies as well as multiword tokens (MWTs) are ubiquitous in Hebrew. As stated previously, we represented MWTs by flattening their structure and labeling each MWT with the feature labels of its components. A common structure of MWTs in Hebrew is ADP-(DET)-NOUN, where the determiner is the definite article -ה *ha* 'the'. Depending on the preposition, the definite article is represented orthographically (e.g., -מה *miha* 'from the') or as a vowel change on the preposition that is not represented orthographically (e.g., -ל can be either *le* 'to a' or *la* 'to the'). When the article is absent from the orthography, we refer to it as being *covert*.[4]

The definite article is one type of determiner in the HTB corpus, but is uniquely identified by the label `PronType=Art`. We thus extracted all of the ADP-(DET)-NOUN cases from the Hebrew test set (234 in total) and examined how well mBERT-6 captured this property. We found that it was less able to recognize `PronType=Art` when the article was not overt (Table 4.4).

Yet, we also found that agreement patterns facilitated recognition of the covert definite article. In particular, Hebrew adjectival modifiers agree with the nouns they modify in gender, number, and definiteness (e.g., in the noun phrase הַבית הַקטן *habayit hakatan* 'the small house', בית *bayit* is 'house.sing.masc', קטן *katan* is 'small.sing.masc', and -ה *ha* is the definite article). Based on UD's

[4]Since the article is (optionally) audible, this usage of *covert* differs slightly from its usage in linguistic theory.

PronType=Art	P	R	F_1
Overt determiner	0.93	0.56	0.70
Covert determiner	0.69	0.40	0.50

Table 4.4: Recognition of the feature PronType=Art in ADP-DET-NOUN multiword tokens, given the Hebrew mBERT-6 probe.

amod annotations, the MWTs that appeared in these constructions constituted 44.3% of TPs, 19.4% of FPs, and 26.2% of FNs when identifying the covert definite article. Moreover, the majority of the FNs involved additional erroneous predictions, where either PronType=Art was not captured on the modifier, the parts of speech were misidentified, or the modifier and the noun were mis-predicted to disagree along an additional feature (i.e., gender or number). These concomitant errors were largely missing from the TPs.

It seems that mBERT-6 has learned that Hebrew nouns and their modifiers agree along multiple features, and that it is able to use the presence of an overt definite article on a modifier to help infer the presence of a covert article in a MWT. When not all of the grammatical features that participate in agreement are captured, this can attenuate recognition of the covert article (and vice versa).

4.5 Multilingual experiments

We used monolingual probes to assess the linguistic representations from mBERT on a language-by-language basis; however, can we replace the individual monolingual probes with a single multilingual probe and derive comparable insights? To address this question, we trained multilingual probes on a shuffled combination of the training sets for Afrikaans, Croatian, Finnish, Hebrew, Spanish, and Turkish. The multilingual probes extracted an aggregated subset of the features captured by the monolingual probes. We then assessed the multilingual probes' performance on each language independently. Overall, the multilingual probes exhibited slight dips in performance, but better selectivity, compared to their monolingual counterparts (Figure 4.4). These trends occurred despite all of the multilingual models converging before they reached epoch 50 during training.

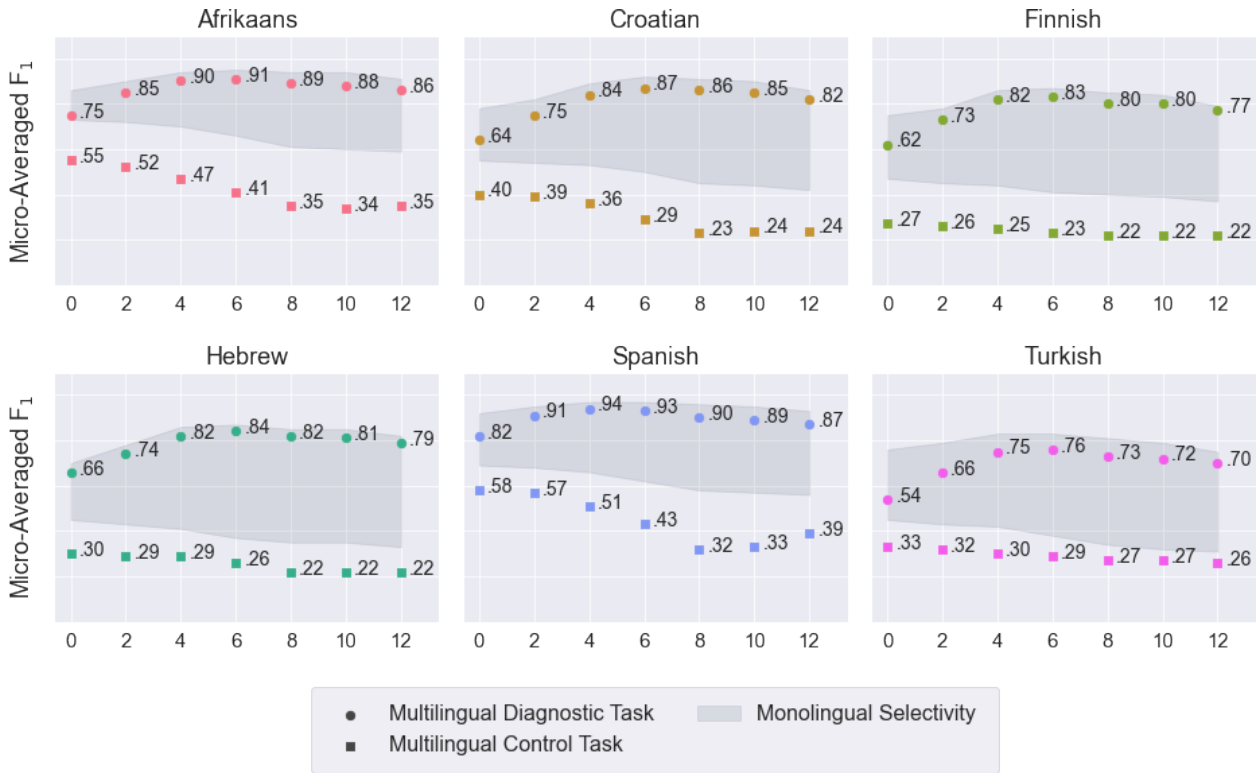


Figure 4.4: Micro-averaged F_1 results from the multilingual probes on the diagnostic and control tasks for each language. The x-axes indicate the mBERT layer. The shaded monolingual selectivity region (for comparison) assumes the same feature label subsets as the multilingual models; incidentally, these monolingual diagnostic task scores are equivalent to the scores reported in Figure 4.3, while the control task scores differ by ±2 points.

4.5.1 Hints of memorization

Recall that the multilingual probes were trained on substantially more data than the monolingual probes. The larger challenge posed by multilingual tagging signals the possibility that a simple linear probe is not complex enough to accommodate the task. If true, this might offer an explanation as to why the monolingual probes outperformed the multilingual probes.

In a small set of post hoc experiments that replaced the linear classifier with a multilayer perceptron, we found that this boosted the multilingual probes' performance—but at the expense of selectivity: We trained multilayer perceptrons with a single hidden layer (MLP-1s) to perform the multilingual morphosyntactic tagging task. As we increased the dimensionality of the hidden layer, the micro-averaged F_1 performance approached that of the monolingual probes, but with comparable or worse selectivity (Table 4.5). In contrast, the linear multilingual probes consistently exhibited the

	Mono.	*Multi.*	$h = 16$	$h = 32$	$h = 64$	$h = 128$
Afrikaans	**0.95**	0.91	0.89	0.91	0.93	0.94
Croatian	**0.92**	0.87	0.83	0.88	0.90	0.91
Finnish	**0.87**	0.83	0.77	0.83	0.85	**0.87**
Hebrew	**0.87**	0.84	0.81	0.84	0.86	**0.87**
Spanish	**0.97**	0.93	0.91	0.94	0.95	0.96
Turkish	**0.83**	0.76	0.71	0.77	0.80	0.82

Table 4.5: Micro-averaged F_1 scores from the linear monolingual and multilingual probes (*Mono.* & *Multi.*) and the multilingual MLP-1 probes with $h = \{16, 32, 64, 128\}$ hidden dimensions.

	Mono.	*Multi.*	$h = 16$	$h = 32$	$h = 64$	$h = 128$
Afrikaans	0.29	**0.50**	0.37	0.29	0.27	0.27
Croatian	0.42	**0.58**	0.42	0.39	0.39	0.39
Finnish	0.46	**0.60**	0.51	0.50	0.50	0.50
Hebrew	0.49	**0.58**	0.52	0.50	0.49	0.48
Spanish	0.35	**0.50**	0.35	0.31	0.30	0.30
Turkish	0.46	**0.47**	0.39	0.38	0.39	0.40

Table 4.6: Selectivity scores from the linear monolingual and multilingual probes (*Mono.* & *Multi.*) and the multilingual MLP-1 probes with $h = \{16, 32, 64, 128\}$ hidden dimensions.

best selectivity (Table 4.6). Together, these findings suggest that the improvements observed by the more complex probes resulted from them having an increased capacity for memorizing the task, rather than from being more "expressive" of the representations encoded within mBERT (cf. Hewitt & Liang, 2019). Thus, the advantage of the monolingual probes over the multilingual probes cannot be reduced to a linear layer not being sufficient enough to extract features from a larger amount of training data.

Indeed, a potential explanation for the contrast in monolingual and multilingual performance is that the simpler task affords the monolingual probes more opportunity to memorize the feature labels. This explanation is supported by how the linear multilingual probes generally exhibit greater selectivity (see Figure 4.4) and accounts for why their performance deficit is, for the most part, spread evenly across the feature labels (see Appendix B for the full feature-level results).

If the monolingual probes do rely more on memorization, this would predict that the multilingual probes are better able to generalize to new data. Recall that the UD corpora assume an open vocabulary; many of the word types in the validation and test sets do not appear during training.

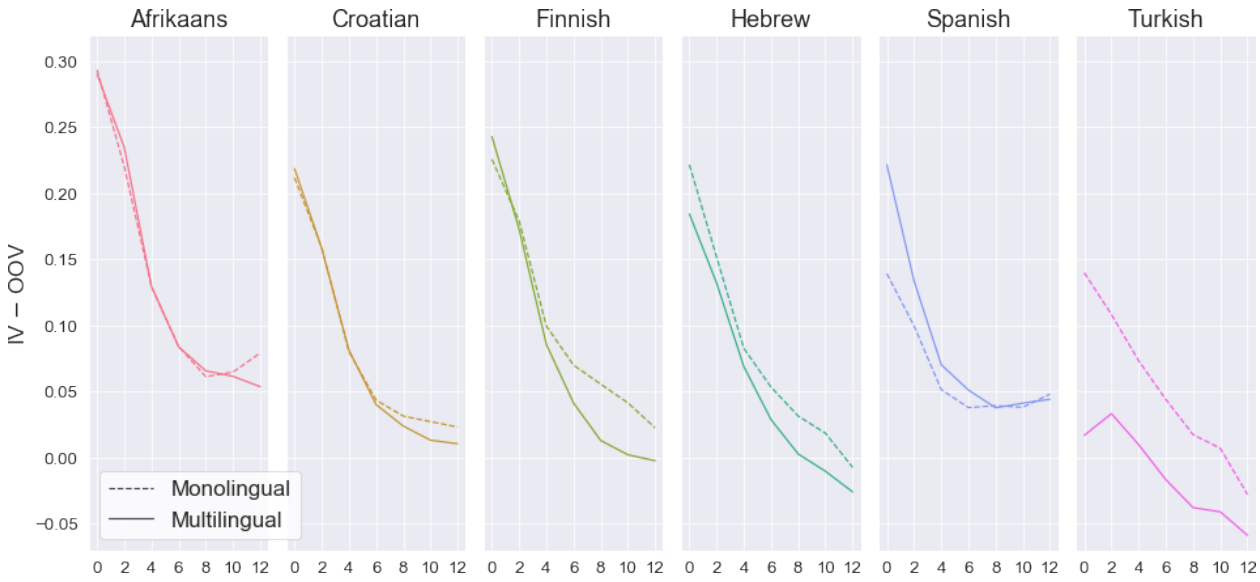


Figure 4.5: Generalizability of the monolingual and multilingual probes; calculated by subtracting the micro-averaged OOV F_1 scores from the micro-averaged IV F_1 scores. The x-axes indicate the mBERT layer. Negative IV-OOV scores indicate instances where the probes performed *better* on OOV tokens than IV tokens.

This allows us to evaluate the effectiveness of the probes on out-of-vocabulary (OOV) words. If both sets of probes comparably extract features versus memorizing the task, we would expect them to perform similarly on in-vocabulary (IV) and OOV words. Conversely, if the monolingual probes rely more heavily on memorization, this would predict that the multilingual probes are better able to generalize to new data.

This prediction is largely borne out by the OOV tokens: We micro-averaged separate F_1 scores for the words that were seen during training and those that weren't. Since the intuition is that a probe that generalizes better will exhibit smaller gaps in performance between OOV and IV words, we subtracted the OOV scores from the IV scores to quantify how well the probes generalized to unseen words, as visualized in Figure 4.5. For Croatian, Finnish, Hebrew, and Turkish, we observed that the gaps between IV and OOV performance tended to be smaller for the multilingual probes than the monolingual ones, especially in later layers.[5] Impressively, some of the Hebrew and Turkish probes

[5]In contrast, for Spanish, the multilingual probes generally exhibited greater IV-OOV gaps than the monolingual models, though this trend diminished with the deeper mBERT layers. Likewise, for Afrikaans, the IV-OOV gaps were very similar between the monolingual and multilingual probes. Crucially, relative to the other languages, the IV-OOV gaps were greatest for Spanish and Afrikaans (where IV performance was better) in both the monolingual and multilingual settings. This reversal of trends is likely due to their substantially larger training sets: The increased number of training tokens (and training steps) may have lured the multilingual probes to memorize the word-to-label

Probe	Af	Hr	Fi	Es	Tr
Mono.	0.89	0.88	0.89	0.96	0.79
Multi.	0.71	0.76	0.80	0.14	0.65

Table 4.7: F_1 results for nominative case (`Case=Nom`) in Afrikaans (Af), Croatian (Hr), Finnish (Fi), Spanish (Es), and Turkish (Tr), given the monolingual and multilingual mBERT-6 probes.

performed *better* amongst OOV tokens than with IV tokens. These trends further suggest that the monolingual probes are more inclined towards memorization than the multilingual probes.

A final piece of evidence comes from language-specific features. For the multilingual experiments, we included two sets of language-specific features: Finnish infinitive forms and Hebrew verb classes (*binyanim*). While the monolingual probes generally outperformed their multilingual counterparts at the feature level, the opposite tended to be true for language-specific features (see Appendices B.3 and B.4). If the multilingual probes are more extractive, especially with respect to crosslinguistic features, this might leave the probe with more "space" to memorize language-specific features.

4.5.2 Multilingual task complexity

Even though the multilingual experiments merely combine the monolingual training data, the multilingual task is just larger but inherently more complex than the monolingual task. Namely, the probes must balance the needs of multiple languages and extract features from a broader diversity of data. This may additional contribute to the slightly lower F_1 performance of the multilingual probes.

Consider the nominative case (subject marking) feature label. When focusing on predictions from mBERT-6, we see that the `Case=Nom` scores for each language dipped with the multilingual probe (Table 4.7). Importantly, the distribution of nominative morphology differs crosslinguistically, as reflected in the UD corpora. For instance, the `Case=Nom` feature label is marked on nouns, verbs, and adjectives in the Turkish corpus, but only on pronouns in the Spanish corpus (remnants from Latin). This variation may result in "conflicting" training signals to the probe, causing the performance of the multilingual probes to dip. Furthermore, it suggests that, although mBERT renders nominative case easily extractable for each language independently, mBERT has not recognized their nominative morphology to correspond to the same nominative notion.

mappings for these languages.

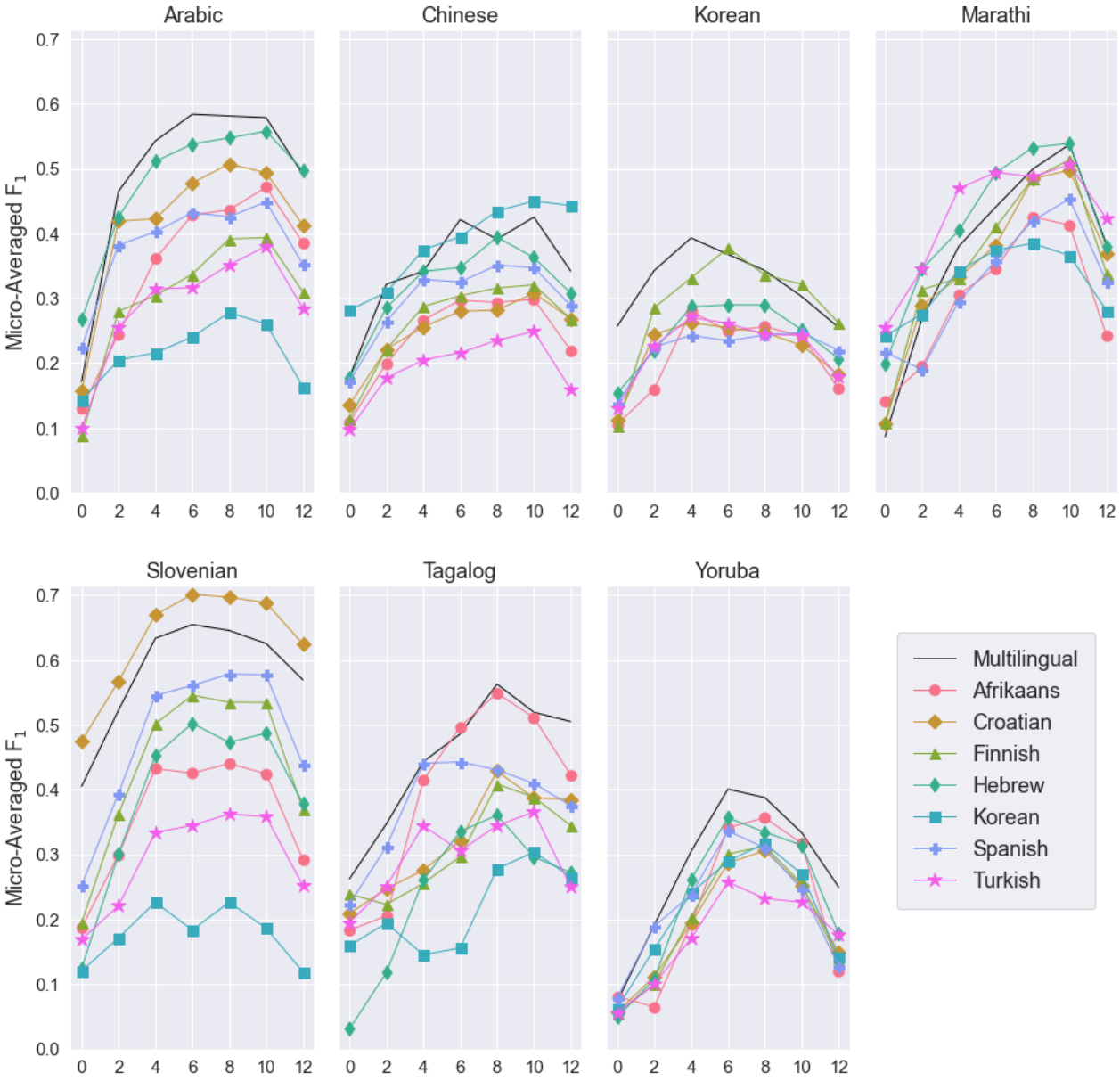


Figure 4.6: Micro-averaged F_1 results from evaluating the monolingual and multilingual probes on the "held-out" languages (plus Korean). The x-axes indicate the mBERT layer.

4.6 Crosslingual experiments

The multilabel probing paradigm can further be used to study which morphosyntactic features are encoded similarly crosslinguistically: If a monolingual probe can successfully extract a feature label given a *held-out* language, this suggests that the language model has come to recognize that property

as being shared by the two languages.

In this section, we evaluate the monolingual and multilingual probes on UD test sets for Arabic, Chinese, Marathi, Slovenian, Tagalog, and Yorùbá. These experiments are akin to prior work on zero-shot crosslingual transfer (Conneau, Wu, Li, Zettlemoyer, & Stoyanov, 2020; K, Wang, Mayhew, & Roth, 2020; Pires et al., 2019; Wu & Dredze, 2019), though we differ from these efforts in that we never fine-tune mBERT. Figure 4.6 shows the monolingual and multilingual micro-averaged F_1 performance across the held-out languages. Focusing once more on mBERT-6, we also examine a small subset of labels, presented in Figure 4.7.

4.6.1 Towards crosslinguistic categories

Getting at the heart of this dissertation: The probes performed relatively well on extracting nouns and verbs across the held-out languages. This suggests that mBERT encodes *noun*-hood and *verb*-hood in a crosslinguistic fashion—that it has some conception of nouns and verbs that transcends individual languages. *Adjective*-hood, in contrast, seems to be represented less cohesively, as indicated by the lower F_1 scores for adjectives in Figure 4.7. For example, the probes struggled to identify adjectives in Chinese, and even more so in Tagalog and Yorùbá. This is not to say that mBERT does not capture adjectives in these languages, but, rather, that it has not connected them to their counterparts in other languages—instead, representing the lexical category in various subspaces of the model. This may be especially true for low-resource languages like Tagalog and Yorùbá.[6] Even though mBERT's training involved over-sampling smaller corpora, it might be the case that the model required exposure to Tagalog and Yorùbá adjectives in a wider array of contexts in order to relate them to their counterparts crosslinguistically (see Conneau, Khandelwal, et al., 2020, for an interesting discussion).

Crosslinguistic variation in a feature's distribution in natural languages might also lead a language model not to recognize when a property is shared by multiple languages. In Section 4.5, we cited such variation as the reason the multilingual probes struggled with nominative case. We posited that mBERT had not recognized nominative morphology in different languages to correspond to the

[6]Recall that mBERT was trained the languages with the top 100 largest Wikipedias. Based on Wikimedia's *List of Wikipedias*, it seems that the Wikipedia dumps for Tagalog and Yoruba were among the smallest corpora that mBERT was trained on, ranking 92 and 106 at present, respectively. Note also that, in the "language resource race", Joshi et al. (2020) give Tagalog and Yoruba scores of 3/5 and 2/5.

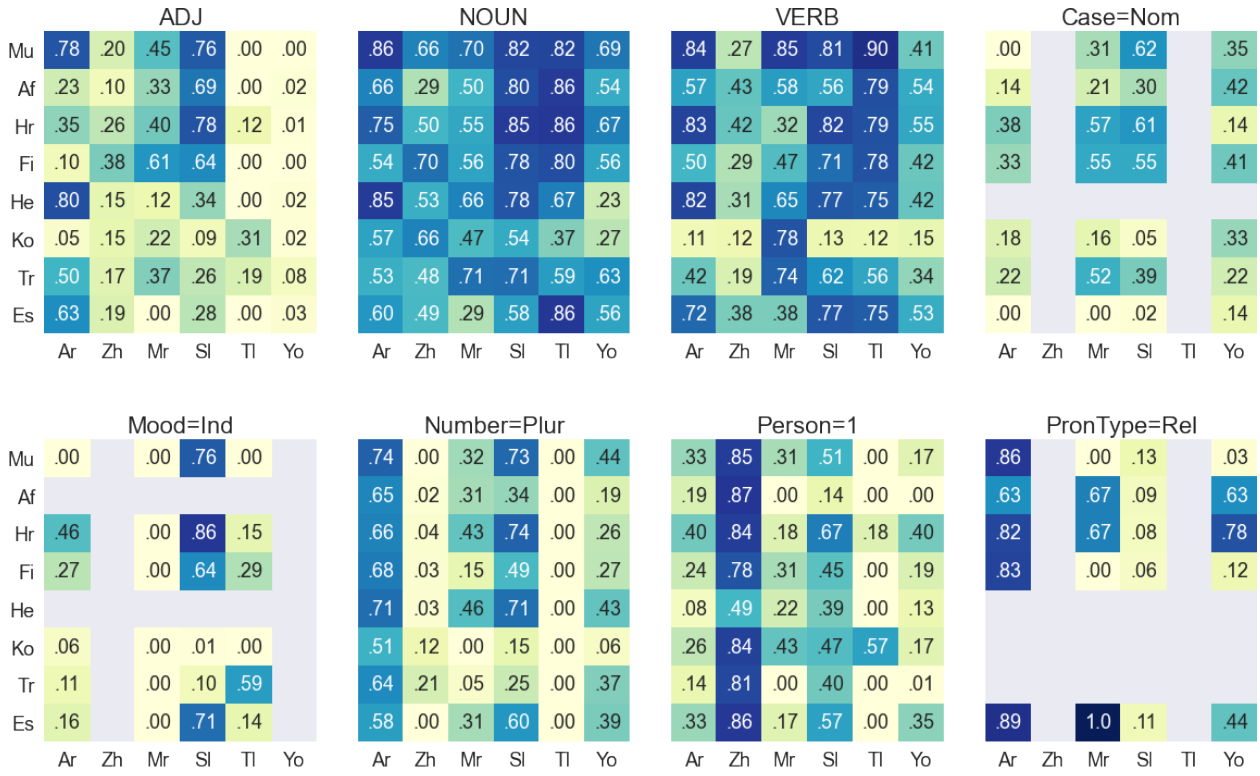


Figure 4.7: A handful of feature-level F_1 results from evaluating the monolingual and multilingual mBERT-6 probes on the"held-out" languages. The *x*-axes indicate the held-out language (Ar=Arabic, Zh=Chinese, Mr=Marathi, Sl=Slovenian, Tl=Tagalog, and Yo=Yoruba), while the *y*-axes indicate the probe (Mu=Multilingual, Af=Afrikaans, Hr=Croatian, Fi=Finnish, He=Hebrew, Ko=Korean, Es=Spanish, and Tr=Turkish). Grayed-out regions indicate where the feature is not applicable to the language or annotated in the language's corpus. In a fun example, the Marathi test set contained exactly one relative pronoun (`PronType=Rel`), which was perfectly identified by the Spanish probe ($F_1 = 1.0$).

same nominative notion. We see this suspicion further borne out in Figure 4.7, where predictions of `Case=Nom` in the held-out languages ranged from 0 to 0.62 F_1. As evidenced by this lack of transfer, it seems that crosslinguistic variation in the distribution of nominative morphology led to a decentralized encoding of nominative case in mBERT. Consequently, this made it more challenging for the probes to capture nominative case in the held-out languages and for the multilingual probes to identify nominative case in general.

Yet, there are also cases where the multilingual probes performed better than the monolingual probes with the held-out languages. Most strikingly, the mBERT-6 multilingual probe obtained 0.90 F_1 on Tagalog verbs, whereas none of the monolingual probes got over 0.79 F_1. This

suggests that, with crosslinguistic properties that are encoded more cohesively, such as *verb*-hood, exposure to multiple languages can lead a probe to forge more replete connections within mBERT's representational space.

4.6.2 Family ties

In the absence of *perfect* crosslingual transfer, we generally find that a monolingual probe will extend equivalently or better to a held-out language than the multilingual probe. In particular, the monolingual probes often did well with *related* languages, showing that family resemblances may have aided mBERT's ability to recognize a feature as being shared by those languages (cf. Conneau, Wu, et al., 2020; Pires et al., 2019; Wu & Dredze, 2019). Compared to the other monolingual probes, for instance, the Hebrew probes fared best with Arabic, another Semitic language, topping out at a micro-averaged F_1 score of 0.56 (see Appendix C). This was also the case at the feature level with nouns, verbs, and adjectives, as shown in Figure 4.7. Notably, Hebrew and Arabic use different scripts. If mBERT has come to represent them similarly, this likely falls out of the structural similarities between the two languages.

Likewise, the Croatian mBERT-6 probe achieved a micro-averaged F_1 score of 0.70 on Slovenian. (For comparison, the Turkish mBERT-6 probe scored 0.76 F_1 *on Turkish*.) In Figure 4.7, the Croatian probe also performed the best on Slovenian nouns, verbs, and adjectives, as well as with words inflected for first person, plurality, or indicative mood. This success seems due to both structural and surface similarities (e.g., cognates) between Croatian and Slovenian. For example, Croatian achieved 0.95 F_1 on conditional mood (`Mood=Cnd`; see Appendix C) and 0.86 F_1 on indicative mood (`Mood=Ind`) in Slovenian because the two languages share several auxiliaries that mark mood (e.g., *bi* for conditional, *je* for indicative).

4.6.3 Revisiting memorization

Note that, with the exception of shared morphemes, the successful instances of crosslingual transfer cannot be reduced to memorization. If the probes merely memorized their monolingual training data, one would expect chance performance and less variability when evaluating them on the held-out languages. These evaluations further verify that the multilabel probes extracted meaningful representations from mBERT. When applied to held-out languages, they also provide a supplementary

method for gauging the complexity of a probe and its ability to memorize a linguistic task.

4.7 Discussion

Studies have highlighted a wealth of linguistic information that can be extracted from language models (e.g., Bacon & Regier, 2019; Chi et al., 2020; Conneau et al., 2018; Edmiston, 2020; Futrell & Levy, 2019; Gulordava et al., 2018; Hewitt & Manning, 2019; Hupkes et al., 2018; Jawahar et al., 2019; N. F. Liu et al., 2019; Marvin & Linzen, 2018; Şahin et al., 2020; Tenney et al., 2019; Zhang & Bowman, 2018). Contributing to this effort, we propose using a multilabel probing task to analyze the morphosyntactic representations of multilingual word embeddings. We demonstrate this probing paradigm with mBERT Devlin et al. (2019).

In a set of monolingual experiments, we trained individual probes for Afrikaans, Croatian, Finnish, Hebrew, Korean, Spanish, and Turkish. We found that mBERT-6 holds the most morphosyntactic information (cf. N. F. Liu et al., 2019; Tenney et al., 2019), with the probes obtaining micro-averaged F_1 scores between 0.83 and 0.97. In a small case study of Hebrew determiners, we illustrated an analysis that implicates *multiple* features (i.e., lexical category, pronominal type, number, and gender). Crucially, traditional single-label efforts would require training multiple models to arrive at such an analysis and, in general, run the risk of overlooking relevant features. In this way, the multilabel training objective may be more efficient, though we leave this comparison for future work.

Next, in a set of multilingual experiments, we saw that the multilingual probes marginally underperformed their monolingual counterparts, while largely upholding the same trends and exhibiting better selectivity. We attributed this contrast in performance to the monolingual probes relying more on memorization, given a simpler task. These findings indicate that the multilingual probes may be more "expressive" diagnostics of linguistic representations (cf. Hewitt & Liang, 2019). However, since our goal is to probe embeddings rather than to perform state-of-the-art morphosyntactic tagging, the monolingual and multilingual probes offer the same *insights* to the extent that they exhibit comparable trends and lend themselves to the same generalizations.

Furthermore, while the multilingual probes are more efficient and less inclined towards memorization, they are also more susceptible to crosslingual interference when the language model has encoded the same property differently across languages. In a set of crosslingual experiments, we further evaluated the monolingual and multilingual probes on data from six "held-out" languages:

Arabic, Chinese, Marathi, Slovenian, Tagalog, and Yorùbá. We showed that applying the probes accordingly can help illuminate which linguistic properties a language model recognizes as being shared by multiple languages and what factors might lead a language model not to encode cohesive representations of a particular crosslinguistic feature.

For instance, we observed that the mBERT was more likely encode shared features cohesively for related languages (e.g., Hebrew and Arabic; Croatian and Slovenian), regardless of whether the languages shared the same script. On the other hand, the model maintained more dispersed encodings of nominative case, as evidenced by the probes' dip in monolingual-to-multilingual performance and their failure to detect the feature in the held-out languages. Together, these results indicate that the model did not always learn the "crosslinguistic status" of features, with structural similarity being a mediating factor. We thus conjectured that crosslinguistic variation in the distribution of nominative morphology is what led the model to form scattered representations of nominative case. In turn, this made it more challenging for the probes to extract nominative case in the held-out languages. Put differently, crosslinguistic variation in nominative morphology—presumably reflected in mBERT's training data (the model's *language experience*)—led the model not to recognize the crosslinguistic status of the feature, resulting in subject marking being encoded in multiple subareas of mBERT's vector space.

4.7.1 Ethical considerations

Numerous studies have highlighted an overblown focus on English as well other key issues in language technology research (Bender, 2011; Bender et al., 2021; Blasi, Anastasopoulos, & Neubig, 2022; Joshi et al., 2020; Ranathunga & de Silva, 2022; Schwartz, Dodge, Smith, & Etzioni, 2020; Søgaard, 2022; Strubell, Ganesh, & McCallum, 2019; *inter alia*). For instance, the vast majority of the world's languages are underrepresented and under-resourced in the language technology arena (Joshi et al., 2020). Even when multilingual systems like mBERT do provide coverage for marginalized languages, they often perform significantly worse for these languages than they do for English (shown, e.g., in Wu & Dredze, 2019). As a result, LLMs tend to be most beneficial for the communities that are already the most advantaged and underserve communities that are already disadvantaged, amplifying and perpetuating power imbalances (Bender et al., 2021). As an added layer of concern, training these systems is also known to emit considerable carbon emissions

(Strubell et al., 2019; see also Schwartz et al., 2020).

From an engineering standpoint, it may be inefficient for machine models not to recognize when properties are shared by multiple languages, because it means that encoding and processing these features is taking up more space and requiring more energy than is theoretically necessary. However, if we were to compare crosslingual behaviors across different machine multilingual models to see which design decisions facilitate improved crosslingualism (i.e., better recognition of the crosslinguistic status of features) and perhaps better alignment with *human* crosslingual patterns, this could inform the design of more streamlined models. Likewise, the more a model is able to draw connections across languages, the more effective it may be for a greater diversity of languages. Thus, this research could prove meaningful for the design of multilingual technologies. Future work should compare crosslingual behaviors across different multilingual models and try to trace disparities to (i) architectural and training data decisions and (ii) performance differences on downstream tasks.

While our proposed probing paradigm is intended for analyzing multilingual LLMs, which are computationally (and monetarily) expensive to produce (cf. Bender et al., 2021; Strubell et al., 2019), our probes are lightweight and quick to train. To help minimize our use of computational resources, we skipped fine-tuning and deployed a "cache and batch" approach to pre-processing our data. Our hope is that conscientious probing efforts will help lead to language model designs that are more efficient and equitable, functioning better for a greater diversity of language populations.

In our experiments, we prioritized working with data from typologically diverse languages, many of which are understudied in NLP (cf. Joshi et al., 2020). We further strove to select datasets that are well documented (cf. Bender & Friedman, 2018; McMillan-Major, Bender, & Friedman, 2023). Drawing on data from Universal Dependencies (Dobrovoljc & Nivre, 2016; Nivre et al., 2016), we worked with morphologically-annotated corpora for 13 different languages: Afrikaans (AfriBooms; cf. Dirix et al., 2017), Arabic (PADT; cf. Hajič et al., 2009; Smrž et al., 2008, 2002), Chinese (PUD; cf. Zeman et al., 2017), Croatian (SET; cf. Agić & Ljubešić, 2015), Finnish (TDT; cf. Haverinen et al., 2014; Pyysalo et al., 2015), Hebrew (HTB; cf. McDonald et al., 2013; Sadde et al., 2018; Tsarfaty, 2013), Korean (PUD; cf. Zeman et al., 2017), Marathi (UFAL; cf. Ravishankar, 2017), Slovenian (SST; cf. Dobrovoljc & Nivre, 2016), Spanish (AnCora; cf. Alonso & Zeman, 2016), Tagalog (TRG), Turkish (IMST; cf. Sulubacak et al., 2016; Türk et al., 2019; Tyers et al., 2017), and Yorùbá (YTB; cf. Ishola & Zeman, 2020).

In general, the datasets in the Universal Dependencies repository are accompanied by documentation of varying quality. Moreover, though Universal Dependencies is an incredible resource—rich with morphosyntactic and dependency annotations—it is important to remember that many of the datasets source texts from narrow domains (e.g., news articles and Bible passages). Thus, they may be limited in the linguistic phenomena they portray. Such domain-specificity, as a consequence, could have hampered our probes' ability to detect full scope of linguistic information captured by mBERT.

4.7.2 Implications of crosslingual transfer for human language processing

Our results showed that mBERT recognizes many morphosyntactic features in different languages, but only occasionally uncovers when a property is shared by those languages—for example, connecting pronouns in the related languages Hebrew and Arabic, but not nominative case in Turkish and Spanish. Moreover, mBERT represented certain types of properties more similarly crosslinguistically than others (e.g., lexical categories vs. case features). Taken together, these patterns indicate that models like mBERT can differentially encode crosslinguistic features with the same weights, exhibiting variation in how much "neuronal overlap" exists per feature between the model's languages (cf. subsequent work by Stańczak, Ponti, Hennigen, Cotterell, & Augenstein, 2022). These machine processing behaviors give rise to two questions: How much do they reflect artifacts of the model's design and training data? And to what extent are they indeed (un-)humanlike?

Fleshing out this latter question: Are humans, like machines, prone to representing certain types of features more similarly? Future work should examine whether certain kinds of properties are more "fundamental" or more likely to be encoded cohesively than others. Lexical categories (e.g., nouns and verbs), for instance, may be processed more similarly crosslinguistically than, say, case morphology. This variation may further exist along multiple continuums, such that members of the same language community might hold more similar feature representations than non-members, whereas whole language communities might resemble one another depending on their languages' structural similarities. If such variation does exist across feature representations, one hypothesis might be that the features that are most representationally similar are the ones that are either typologically frequent (i.e., closer to being universal) or follow a narrower distribution (less variation) crosslinguistically. To probe representational similarities in humans, future efforts

might look towards neuroimaging of morphological processes (e.g., Hauptman, Blanco-Elorrieta, & Pylkkänen, 2022) and patterns of crosslinguistic influence in artificial language learning experiments (cf. Chapter 3 of this dissertation).

Returning to language experience, variation in how common linguistic properties surface crosslinguistically may have led mBERT to differentially encode the crosslinguistic status of features, resulting in variable crosslingual transfer effects. If humans likewise vary in their encodings of crosslinguistic features, it is likely that crosslinguistic variation in our languages experiences both reflects and reinforces these underlying processing differences. Therefore, future research should explore how variation in human language processing might relate to feature-level distributional (dis)similarities across languages. It would also be interesting to see if variable machine crosslingual behaviors can predict such processing differences in humans, and vice versa.

4.8 Conclusion

Emerging studies on interpretability have highlighted a wealth of linguistic information that can be extracted from LLMs. Contributing to this effort, we proposed using a multilabel probing task to analyze the morphosyntactic representations of multilingual word embeddings. We demonstrated this probing paradigm with mBERT (Devlin et al., 2019). Multilabel probe predictions can be used to perform holistic analyses of a model's ability to encode systems of morphology, as well as more fine-grained analyses of individual features, agreement phenomena, and how shared properties are represented crosslinguistically. Future efforts should probe different multilingual language models using the multilabel paradigm and examine how these models might vary in their crosslingual morphosyntactic representations. Finally, future research should explore how global and feature-level morphosyntactic probe performance—especially with respect to crosslingual transfer—corresponds to the performance of downstream systems, especially amongst morphologically rich and low-resource languages.

5 Conclusion

From humans to machines, this dissertation has explored how variation in language experience can relate to variation in behavior—at times, complicating the *nature vs. nurture* debate in overlooked ways. Chapter 1 contextualized this exploration by outlining a disproportionate focus on English and the experiences of English speakers in the language-related sciences. This narrow lens limits our understanding of human cognition, particularly when it comes to distinguishing genetic influences from experiential ones, and hinders the development of equitable language technologies. Each chapter of this dissertation thus told the story of a learner, delving into the effects of prior and concurrent experience on the learner's language usage. Along the way, the chapters showcased different methodologies for examining diverse language behaviors.

Beginning with humans, Chapter 2—*How chatty are daddies?*—zoomed in on infants learning their first language within a regional community (i.e., the Seattle area of the US Pacific Northwest, where this dissertation was also written). Drawing on LENA technology, we showed that differences in maternal and paternal input in local English-speaking families can lead to variation in infant volubility. On average, the infants in our sample of 23 families were exposed to 46.8% fewer words and 51.9% less parentese from their fathers, even though paternal parentese grew at a 2.8× faster rate as the infants aged from 6 to 24 months. An interesting asymmetry also emerged where only maternal word counts and paternal parentese predicted child vocalizations—a microcosm of the complex relationships between language experience and linguistic behavior. We encourage future research to study parent-infant interactions among more socioculturally and linguistically diverse families. Crucially, these efforts should explicate gender differences in terms of family dynamics and parental beliefs. Doing so may help mitigate gender biases while clarifying the complex relationships between environment variables and child language outcomes.

Shifting to adults learners of artificial languages, we devised a novel, iconicity-based approach to artificial language learning (ALL) in Chapter 3—*Iconic artificial language learning*. Teaching picto-

graphic languages to our participants, we showed that English speakers prefer word and morpheme orders that are scope-isomorphic, conceptually replicating previously attested observations. We then questioned whether these findings can be attributed to a universal bias or, instead, participants' prior language experiences, detailing ongoing work with Polish speakers. Moving forward, the same pictographic languages may be used with diverse language populations, including adults and children, speakers and signers, members of Indigenous communities, as well as monolinguals, bilinguals, and multilinguals. This methodological shift can facilitate controlled crosslinguistic investigations in the ALL space, deepening our understanding of crosslinguistic influence and universal cognitive constraints.

Turning to machines, Chapter 4—*Probing mBERT*—dipped into artificial agents learning human languages. Using a multilabel probing task to analyze mBERT embeddings, we examined how the model captured 166 morphosyntactic features across 13 typologically diverse languages: Afrikaans, Arabic, Chinese, Croatian, Finnish, Hebrew, Korean, Marathi, Spanish, Slovenian, Tagalog, Turkish, and Yorùbá. We found that mBERT recognized many morphosyntactic features in different languages, but only occasionally uncovered when a property is shared by those languages. Taken together, these patterns indicate that multilingual language models can encode a single feature in multiple neuronal subspaces. Future work should explore how scattered crosslingualism correlates with downstream performance, especially for low-resource languages. Notably, differential feature encodings likely result from crosslinguistic variation, as capsulized in multilingual training data. Finally, an intriguing question is to what extent such variation in feature representations is humanlike... In this sense, language models can be used to *inspire*, rather than test, hypotheses of human language processing.

Coalescing the different elements of this dissertation, it seems we can ask the same question of both humans and machines: How do our "neurons" come to internalize and reflect variation in language environments? Answering this question invites us—*calls* for us—to explore the remarkable linguistic and sociocultural diversity that characterizes the human experience.

Bibliography

Agić, Ž., & Ljubešić, N. (2015). Universal Dependencies for Croatian (that work for Serbian, too). In *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing* (pp. 1–8). Retrieved from https://www.aclweb.org/anthology/W15-5301

Alain, G., & Bengio, Y. (2018). Understanding intermediate layers using linear classifier probes. *arXiv:1610.01644*. doi: 10.48550/arXiv.1610.01644

Alonso, H. M., & Zeman, D. (2016). Universal Dependencies for the AnCora Treebanks. In *Procesamiento del Lenguaje Natural* (pp. 91–98). Retrieved from http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5341

Andrew, K. N., Hoshooley, J., & Joanisse, M. F. (2014). Sign language ability in young deaf signers predicts comprehension of written sentences in English. *PLOS ONE*, *9*(2), e89994. doi: 10.1371/journal.pone.0089994

Bacon, G., & Regier, T. (2019). Does BERT agree? Evaluating knowledge of structure dependence through agreement relations. *arXiv:1908.09892*. doi: 10.48550/arXiv.1908.09892

Bagner, D. M. (2013). Father's role in parent training for children with developmental delay. *Journal of Family Psychology*, *27*(4), 650–657. doi: 10.1037/a0033465

Bagner, D. M., & Eyberg, S. M. (2003). Father involvement in parent training: When does it matter? *Journal of Clinical Child and Adolescent Psychology*, *32*(4), 599–605. doi: 10.1207/S15374424JCCP3204_13

Baker, C. E., & Vernon-Feagans, L. (2015). Fathers' language input during shared book activities: Links to children's kindergarten achievement. *Journal of Applied Developmental Psychology*, 53–59. doi: 10.1016/j.appdev.2014.11.009

Baker, M. C. (2001). *The Atoms of Language: The Mind's Hidden Rules of Grammar*. Oxford University of Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using

lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bates, E., & MacWhinney, B. (1982). Functionalist approaches to grammar. In E. Wanner & L. R. Gleitman (Eds.), *Language Acquisition: The State of the Art* (pp. 173–218). Cambridge University Press.

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., & Glass, J. (2017). What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 861–872). doi: 10.18653/v1/P17-1080

Belinkov, Y., Màrquez, L., Sajjad, H., Durrani, N., Dalvi, F., & Glass, J. (2017). Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1–10). Retrieved from https://www.aclweb.org/anthology/I17-1001

Bender, E. M. (2011). On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, *6*, 1–26. doi: 10.33011/lilt.v6i.1239

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (p. 610–623). doi: 10.1145/3442188.3445922

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2018). What do North American babies hear? A large-scale cross-corpus analysis. *Developmental Science*, *22*(e12724), 1–12. doi: 10.1111/desc.12724

Bickel, B., Banjade, G., Gaenszle, M., Lieven, E., Paudyal, N. P., Rai, I. P., . . . Stoll, S. (2007). Free prefix ordering in Chintang. *Language*, *83*(1), 43–73. doi: 10.1353/lan.2007.0002

Bickel, B., & Nichols, J. (2013). Fusion of selected inflectional formatives [Data set]. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online (v2020.3).* Zenodo. doi: 10.5281/zenodo.7385533

Bielec, D. (1998). *Polish: An Essential Grammar*. Routledge.

Bingham, G. E., Kwon, K. A., & Jeon, H.-J. (2013). Examining relations among mothers', fathers', and children's language use in a dyadic and triadic context. *Early Child Development and Care*, *183*(3-4), 394–414. doi: 10.1080/03004430.2012.711590

Blasi, D. E., Anastasopoulos, A., & Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5486–5505). doi: 10.18653/v1/2022.acl-long.376

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153–1170. doi: 10.1016/j.tics.2022.09.015

Blevins, T., & Zettlemoyer, L. (2019). Better character language modeling through morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1606–1613). doi: 10.18653/v1/P19-1156

Boland, J. E., Kaan, E., Valdés Kroff, J., & Wulff, S. (2016). Psycholinguistics and variation in language processing. *Linguistics Vanguard*, *2*(s1), 20160064. doi: 10.1515/lingvan-2016-0064

Bowerman, J., & Smith, K. (2022). An experimental study of semantic extension in a novel communication system. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44). Retrieved from https://escholarship.org/uc/item/3x3398ct

Broesch, T., & Bryant, G. A. (2018). Fathers' infant-directed speech in a small-scale society. *Child Development*, *89*(2), e29–e41. doi: 10.1111/cdev.12768

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Bruening, P. R., Brooks, P. J., Alfieri, L., Kempe, V., & Dabašinskienė, I. (2012). Children's tolerance of word-form variation. *Child Development Research*, *2012*(401680), 1–12. doi: 10.1155/2012/401680

Bruner, J. (1981). The social context of language acquisition. *Language and Communication*, *1*(2-3), 155–178. doi: 10.1016/0271-5309(81)90010-0

Bulgarelli, F., & Bergelson, E. (2020). Look who's talking: A comparison of automated and human-generated speaker tags in naturalistic day-long recordings. *Behavior Research Methods*, *52*(2), 641–653. doi: 10.3758/s13428-019-01265-7

Bybee, J. L., Pagliuca, W., & Perkins, R. D. (1990). On the asymmetries in the affixation of grammatical material. In W. A. Croft, S. Kemmer, & K. Denning (Eds.), *Studies in Typology and Diachrony: Papers Presented to Joseph H. Greenberg on his 75th Birthday* (pp. 1–42). doi: 10.1075/tsl.20.04byb

Byers-Heinlein, K., Esposito, A. G., Winsler, A., Marian, V., Castro, D. C., & Luk, G. (2019). The case for measuring and reporting bilingualism in developmental research. *Collabra: Psychology*, *5*(1), 37. doi: 10.1525/collabra.233

Bylund, E., Khafif, Z., & Berghoff, R. (2023). Linguistic and geographic diversity in research on second language acquisition and multilingualism: An analysis of selected journals. *Applied Linguistics*, *XX*(XX), 1–26. doi: 10.1093/applin/amad022

Cabrera, N. J., Tamis-LeMonda, C. S., Bradley, R. H., Hofferth, S., & Lamb, M. E. (2000). Fatherhood in the twenty-first century. *Child Development*, *71*(1), 127–136. doi: 10.1111/1467-8624.00126

Cabrera, N. J., Volling, B. L., & Barr, R. (2018). Fathers are parents, too! Widening the lens on parenting for children's development. *Child Development Perspectives*, *12*(3), 152–157. doi: 10.1111/cdep.12275

Cameron, J. J., & Stinson, D. A. (2019). Gender (mis)measurement: Guidelines for respecting gender diversity in psychological research. *Social and Personality Psychology Compass*, *13*(11), 1–14. doi: 10.1111/spc3.12506

Chamberlain, C., & Mayberry, R. I. (2000). Theorizing about the relation between American Sign Language and reading. In C. Chamberlain, J. P. Morford, & R. I. Mayberry (Eds.), *Language Acquisition by Eye* (pp. 221–259). doi: 10.4324/9781410601766

Cheng, L. S., Burgess, D., Vernooij, N., Solís-Barroso, C., McDermott, A., & Namboodiripad, S. (2021). The problematic concept of native speaker in psycholinguistics: Replacing vague and harmful terminology with inclusive and accurate measures. *Frontiers in Psychology*, *12*, 715843. doi: 10.3389/fpsyg.2021.715843

Chi, E. A., Hewitt, J., & Manning, C. D. (2020). Finding universal grammatical relations in

Multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5564–5577). doi: 10.18653/v1/2020.acl-main.493

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

Chomsky, N. (1980). On cognitive structures and their development: A reply to Piaget. In M. Piatelli-Palmarini (Ed.), *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky* (pp. 35–54). Harvard University Press.

Chomsky, N., Roberts, I., & Watumull, J. (2023, March 8). Noam Chomsky: The False Promise of ChatGPT. *The New York Times*. Retrieved from https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Christakis, D. A., Gilkerson, J., Richards, J. A., Zimmerman, F. J., Garrison, M. M., Xu, D., ... Yapanel, U. (2009). Audible television and decreased adult words, infant vocalizations, and conversational turns: A population-based study. *Archives of Pediatrics and Adolescent Medicine*, *163*(6), 554–558. doi: 10.1001/archpediatrics.2009.61

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(5), 489–509. doi: 10.1017/S0140525X08004998

Christiansen, M. H., Conway, C. M., & Onnis, L. (2012). Similar neural correlates for language and sequential learning: Evidence from event-related brain potentials. *Language and Cognitive Processes*, *27*(2), 231–256. doi: 10.1080/01690965.2011.606666

Conica, M., Nixon, E., & Quigley, J. (2020). Fathers' but not mothers' repetition of children's utterances at age two is associated with child vocabulary at age four. *Journal of Experimental Child Psychology*, *191*, 104738. doi: 10.1016/j.jecp.2019.104738

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). doi: 10.18653/v1/2020.acl-main.747

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single $&!#∗vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2126–2136). doi: 10.18653/v1/P18-1198

Conneau, A., Wu, S., Li, H., Zettlemoyer, L., & Stoyanov, V. (2020). Emerging cross-lingual structure

in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6022–6034). doi: 10.18653/v1/2020.acl-main.536

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, *47*(3), e13256. doi: P10.1111/cogs.13256

Cristia, A., Bulgarelli, F., & Bergelson, E. (2020). Accuracy of the Language Environment Analysis system segmentation and metrics: A systematic review. *Journal of Speech, Language, and Hearing Research*, *63*(4), 1093–1105. doi: 10.1044/2020_JSLHR-19-00017

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., ... Bergelson, E. (2021). A thorough evaluation of the Language Environment Analysis (LENA) system. *Behavior Research Methods*, *53*, 467–486. doi: 10.3758/s13428-020-01393-5

Culbertson, J. (2012). Typological universals as reflections of biased learning: Evidence from artificial language learning. *Language and Linguistics Compass*, *6*(5), 310–329. doi: 10.1002/lnc3.338

Culbertson, J. (2023). Artificial language learning. In J. Sprouse (Ed.), *Oxford Handbook of Experimental Syntax* (pp. 271–300). doi: 10.1093/oxfordhb/9780198797722.013.9

Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, *111*(16), 5842–5847. doi: 10.1073/pnas.1320525111

Culbertson, J., & Kirby, S. (2022). Syntactic harmony arises from a domain-general learning bias. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44). Retrieved from https://escholarship.org/uc/item/8ht6c0hm

Culbertson, J., Schouwstra, M., & Kirby, S. (2020). From the world to word order: Deriving biases in noun phrase order from statistical properties of the world. *Language*, *96*(3), 696–717. doi: 10.1353/lan.2020.0045

Culbertson, J., & Schuler, K. (2019). Artificial language learning in children. *Annual Review of Linguistics*, *5*, 353–373. doi: 10.1146/annurev-linguistics-011718-012329

Cutler, A., Hawkins, J. A., & Gilligan, G. (1985). The suffixing preference: A processing explanation. *Linguistics*, *23*, 723-758. doi: 10.1515/ling.1985.23.5.723

Cysouw, M. (2010). Dealing with diversity: Towards an explanation of NP-internal word order

frequencies. *Linguistic Typolgy*, *14*, 253–286. doi: 10.1515/lity.2010.010

Dautriche, I., Buccola, B., Berthet, M., Fagot, J., & Chemla, E. (2022). Evidence for compositionality in baboons (Papio papio) through the test case of negation. *Scientific Reports*, *12*(1), 19181. doi: 10.1038/s41598-022-21143-1

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavioral Research Methods*, *47*(1), 1–12. doi: 10.3758/s13428-014-0458-y

Devlin, J., Chang, M.-W., Lee, K., & Toutanova Kristina. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). doi: 10.18653/v1/N19-1423

Dirix, P., Augustinus, L., van Niekerk, D., & Van Eynde, F. (2017). Universal Dependencies for Afrikaans. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)* (pp. 38–47). Retrieved from https://www.aclweb.org/anthology/W17-0405

Dobashi, Y. (2003). *Phonological phrasing and syntactic derivation.* Cornell University.

Dobrovoljc, K., & Nivre, J. (2016). The Universal Dependencies Treebank of Spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1566–1573). Retrieved from https://www.aclweb.org/anthology/L16-1248

Doddapaneni, S., Ramesh, G., Khapra, M. M., Kunchukuttan, A., & Kumar, P. (2021). A primer on pretrained multilingual language models. *arXiv:2107.00676*. doi: https://doi.org/10.48550/arXiv.2107.00676

Dryer, M. S. (2018). On the order of demonstrative, numeral, adjective, and noun. *Language*, *94*(4), 798–833. doi: 10.1353/lan.2018.0054

Dufter, P., & Schütze, H. (2020). Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4423–4437). doi: 10.18653/v1/2020.emnlp-main.358

Edmiston, D. (2020). A systematic analysis of morphological content in BERT models for multiple languages. *arXiv:2004.03032*. doi: 10.48550/arXiv.2004.03032

Emmorey, K. (2020). The neurobiology of reading differs for deaf and hearing adults. , 347–359.

doi: 10.1093/oxfordhb/9780190054045.013.25

Emmorey, K., & Lee, B. (2021). The neurocognitive basis of skilled reading in prelingually and profoundly deaf adults. *Language and Linguistics Compass*, *15*(2), e12407. doi: 10.1111/lnc3.12407

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(5), 429–448. doi: 10.1017/S0140525X0999094X

Farran, L. K., Lee, C.-C., Yoo, H., & Oller, D. K. (2016). Cross-cultural register differences in infant-directed speech: An initial study. *PLOS ONE*, *11*(3), e0151518.

Fedzechkina, M., Newport, E. L., & Florian Jaeger, T. (2016). Miniature artificial language learning as a complement to typological data. In L. Ortega, A. E. Tyler, H. I. Park, & M. Uno (Eds.), *The Usage-Based Study of Language Learning and Multilingualism* (pp. 211–232). Georgetown University Press. doi: 10.1353/book45841

Ferjan Ramírez, N., Lytle, S. R., Fish, M., & Kuhl, P. K. (2018). Parent coaching at 6 and 10 months improves language outcomes at 14 months: A randomized controlled trial. *Developmental Science*, *22*(3), 1–14. doi: 10.1111/desc.12762

Ferjan Ramírez, N., Lytle, S. R., & Kuhl, P. K. (2020). Parent coaching increases conversational turns and advances infant language development. *Proceedings of the National Academy of Sciences*, *117*(7), 3484–3491. doi: 10.1073/pnas.1921653117

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant behavior and development*, *8*(2), 181–195. doi: 10.1016/S0163-6383(85)80005-9

Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development*, *10*(3), 279–293. doi: 10.1016/0163-6383(87)90017-8

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501. doi: 10.1017/S0305000900010679

Fitch, W. T., Hauser, M. D., & Chomsky, N. (2005). The evolution of the language faculty: Clarifications and implications. *Cognition*, *97*(2), 179–210. doi: 10.1016/j.cognition.2005.02.005

Folia, V., Uddén, J., De Vries, M., Forkstam, C., & Petersson, K. M. (2010). Artificial language

learning in adults and children. *Language learning*, *60*(s2), 188–220. doi: 10.1111/j.1467-9922.2010.00606.x

Futrell, R., Hickey, T., Lee, A., Lim, E., Luchkina, E., & Gibson, E. (2015). Cross-linguistic gestures reflect typological universals: A subject-initial, verb-final bias in speakers of diverse languages. *Cognition*, *136*, 215–221. doi: 10.1016/j.cognition.2014.11.022

Futrell, R., & Levy, R. P. (2019). Do RNNs learn human-like abstract word order preferences? In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019* (pp. 50–59). doi: 10.7275/jb34-9986

Genovese, G., Spinelli, M., Lauro, L. J. R., Aureli, T., Castelletti, G., & Fasolo, M. (2020). Infant-directed speech as a simplified but not simple register: a longitudinal study of lexical and syntactic features. *Journal of Child Language*, *47*(1), 22–44. doi: 10.1017/S0305000919000643

Gergely, A., Faragó, T., Galambos, Á., & Topál, J. (2017). Differential effects of speech situations on mothers' and fathers' infant-directed and dog-directed speech: An acoustic analysis. *Scientific Reports*, *7*(1), 1–10. doi: 10.1038/s41598-017-13883-2

Gerz, D., Vulić, I., Ponti, E., Naradowsky, J., Reichard, R., & Korhonen, A. (2018). Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the Association for Computational Linguistics*, *6*, 451–465. doi: 10.1162/tacl_a_00032

Gilkerson, J., & Richards, J. A. (2009). *The Power of Talk: Impact of Adult Talk, Conversational Turns, and TV During the Critical 0-4 Years of Child Development* (2nd ed.; Tech. Rep.). LENA Foundation.

Gilkerson, J., & Richards, J. A. (2020). *LENA: A guide to understanding the design and purpose of the LENA system* (Tech. Rep.). LENA Foundation. doi: TechnicalReportLTR-12

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Oller, D. K., ... Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, *26*(2), 248–265. doi: 10.1044/2016_AJSLP-15-0169

Goldin-Meadow, S., & Mayberry, R. I. (2001). How do profoundly deaf children learn to read? *Learning Disabilities Research & Practice*, *16*(4), 222–229. doi: 10.1111/0938-8982.00022

Goldin-Meadow, S., So, W. C., Özyürek, A., & Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences*, *105*(27), 9163–9168. doi: 10.1073/pnas.0710060105

Golinkoff, R. M., & Ames, G. J. (1979). A comparison of fathers' and mothers' speech with their young chidren. *Child Development*, *50*(1), 28–32. doi: 10.1111/j.1467-8624.1979.tb02975.x

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby)talk to me: The social context of infant-directed speech and its effects on early language acquisition. *Current Directions in Psychological Science*, *24*(5), 339–344. doi: 10.1177/0963721415595345

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4*(5), 178–186. doi: 10.1016/S1364-6613(00)01467-4

Greenberg, J. H. (1957). *Essays in Linguistics*. University of Chicago Press.

Greenberg, J. H. (1963). *Universals of Language*. MIT Press.

Grey, S. (2020). What can artificial languages reveal about morphosyntactic processing in bilinguals? *Bilingualism: Language and Cognition*, *23*(1), 81–86. doi: 10.1017/S1366728919000567

Grosjean, F. (1982). *Life with Two Languages: An Introduction to Bilingualism*. Harvard University Press.

Grosjean, F. (2008). *Studying Bilinguals*. Oxford University Press.

Grosjean, F. (2010). *Bilingual: Life and Reality*. Harvard University Press.

Gullifer, J. W., Kousaie, S., Gilbert, A. C., Grant, A., Giroud, N., Coulter, K., ... Titone, D. (2021). Bilingual language experience as a multidimensional spectrum: Associations with objective and subjective language proficiency. *Applied Psycholinguistics*, *42*(2), 245–278. doi: 10.1017/S0142716420000521

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1195–1205). doi: 10.18653/v1/N18-1108

Hajič, J., Smrž, O., Zemánek, P., Pajas, P., Šnaidauf, J., Beška, E., ... Hassanová, K. (2009). *Prague Arabic Dependency Treebank 1.0* (Tech. Rep.). Retrieved from http://ufal.mff.cuni.cz/padt

Hall, C. J. (1988). Integrating diachronic and processing principles in explaining the suffixing

preference. In J. A. Hawkins (Ed.), *Explaining Language Universals* (pp. 321–349). Blackwell.

Hart, B., & Risley, T. R. (1995). *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore, MD, US: Paul H. Brookes Publishing.

Hartman, K. M., Ratner, N. B., & Newman, R. S. (2017). Infant-directed speech (IDS) vowel clarity and child language outcomes. *Journal of Child Language*, *44*(5), 1140–1163. doi: 10.1017/S0305000916000520

Hattori, S. (1950). Clitics and bound forms. *Gengo Kenkyu (Journal of the Linguistic Society of Japan)*, *1950*(15), 89-89. doi: 10.11435/gengo1939.1950.1

Hauptman, M., Blanco-Elorrieta, E., & Pylkkänen, L. (2022). Inflection across categories: Tracking abstract morphological processing in language production with MEG. *Cerebral Cortex*, *32*(8), 1721–1736. doi: 10.1093/cercor/bhab309

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579. doi: 10.1126/science.298.5598.1569

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., . . . Ginter, F. (2014). Building the essential resources for Finnish: The Turku Dependency Treebank. *Language Resources and Evaluation*, *48*(3), 493–531. doi: 10.1007/s10579-013-9244-1

Hawkins, J. A., & Cutler, A. (1988). Psycholinguistic factors in morphological asymmetry. In J. A. Hawkins (Ed.), *Explaining Language Universals* (pp. 280–317). Blackwell.

Hawkins, J. A., & Gilligan, G. (1988). Prefixing and suffixing universals in relation to basic word order. *Lingua*, *74*(2-3), 219–259. doi: 10.1016/0024-3841(88)90060-5

Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics*, *39*, 1041-1070. doi: 10.1515/ling.2001.041

Hay, J. (2002). From speech perception to morphology: Affix ordering revisited. *Language*, *78*(3), 527–555. doi: 10.1353/lan.2002.0159

Hay, J., & Plag, I. (2004). What constrains possible suffix combinations? On the interaction of grammatical and processing restrictions in derivational morphology. *Natural Language & Linguistic Theory*, *22*(3), 565–596. doi: 10.1023/B:NALA.0000027679.63308.89

Henrich, J. (2020). *The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous*. Farrar, Straus and Giroux.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). Most people are not WEIRD. *Nature*, *466*, 29.

doi: 10.1038/466029a

Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. doi: 10.1017/S0140525X0999152X

Hewitt, J., & Liang, P. (2019). Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 2733–2743). doi: 10.18653/v1/D19-1275

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4129–4138). doi: 10.18653/v1/N19-1419

Hladik, E. G., & Edwards, H. T. (1984). A comparative analysis of mother-father speech in the naturalistic home environment. *Journal of Psycholinguistic Research*, *13*(5), 321–332. doi: 10.1007/BF01068149

Hockett, C. F. (1959). Animal "languages" and human language. *Human Biology*, *31*(1), 32–39. Retrieved from https://www.jstor.org/stable/41449227

Hoen, M., & Dominey, P. F. (2004). Evidence for a shared mechanism in linguistic and nonlinguistic sequence processing? ERP recordings of on-line function- and content-information integration. In M. Carreiras & C. Clifton Jr. (Eds.), *The On-line Study of Sentence Comprehension* (pp. 329–356). Psychology Press. doi: 10.4324/9780203509050

Hoff, E. (2003a). Causes and consequences of SES-related differences in parent-to-child speech. In M. H. Bornstein & R. H. Bradley (Eds.), *Socioeconomic Status, Parenting, and Child Development* (pp. 147–160). Lawrence Erlbaum Associates Publishers. doi: 10.4324/9781410607027-15

Hoff, E. (2003b). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74*(5), 1368–1378. doi: 10.1111/1467-8624.00612

Hoffmeister, R. J., & Caldwell-Harris, C. L. (2014). Acquiring English as a second language via print: The task for deaf children. *Cognition*, *132*(2), 229–242. doi: 10.1016/j.cognition.2014.03.014

Hollingshead, A. B. (1975). *Four factor index of social status.* Unpublished manuscript.

Hollingshead, A. B. (2011). Four factor index of social status. *Yale Journal of Sociology*, *8*, 21–51. Retrieved from https://sociology.yale.edu/sites/default/files/files/yjs_fall_2011.pdf

Hupkes, D., Veldhoen, S., & Zuidema, W. (2018). Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks rocess hierarchical structure. *Journal of Artificial Intelligence Research*, *61*, 907–926. doi: 10.1613/jair.1.11196

Hupp, J. M., Sloutsky, V. M., & Culicover, P. W. (2009). Evidence for a domain-general mechanism underlying the suffixation preference in language. *Language and Cognitive Processes*, *24*(6), 876–909. doi: 10.1080/01690960902719267

Ishola, & Zeman, D. (2020). Yorùbá Dependency Treebank (YTB). In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 5178–5186). Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.637

Jackendoff, R., & Pinker, S. (2005). The nature of the language faculty and its implications for evolution of language (Reply to Fitch, Hauser, and Chomsky). *Cognition*, *97*(2), 211–225. doi: 10.1016/j.cognition.2005.04.006

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic Influence in Language and Cognition*. Routledge.

Jawahar, G., Sagot, B., & Seddah, D. (2019). What dooes BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3651–3657). Association for Computational Linguistics. doi: 10.18653/v1/P19-1356

Jones, J., & Mosher, W. D. (2013). Fathers' involvement with their children: United States, 2006-2010. *National Health Statistics Reports*(71), 1–21. Retrieved from https://www.cdc.gov/nchs/data/nhsr/nhsr071.pdf

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). doi: 10.18653/v1/2020.acl-main.560

K, K., Wang, Z., Mayhew, S., & Roth, D. (2020). Cross-lingual ability of multilingual BERT: An empirical study. In *International Conference on Learning Representations.* Retrieved from https://openreview.net/forum?id=HJeT3yrtDr

Kageyama, T. (2020). Morphology in Japonic Languages. In *Oxford research encyclopedia of*

linguistics. doi: 10.1093/acrefore/9780199384655.013.538

Kalashnikova, M., & Burnham, D. (2018). Infant-directed speech from seven to nineteen months has similar acoustic properties but different functions. *Journal of Child Language*, *45*(5), 1035–1053. doi: 10.1017/S0305000917000629

Kanampiu, P. N., & Muriungi, P. K. (2019). Order of modifiers in Kî itharaka determiner phrase. *International Journal on Studies in English Language and Literature*, *7*, 10-21. doi: 10.20431/2347-3134.0706002

Kennison, S. M., & Byrd-Craven, J. (2015). Gender differences in beliefs about infant-directed speech: The role of family dynamics. *Child Development Research*, *2015*(871759), 1–6. doi: 10.1155/2015/871759

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language*, *42*(6), 703–735. doi: 10.1177/0142723721106640

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *arXiv:1412.6980*. doi: 10.48550/arXiv.1412.6980

Klein, S., & Tsarfaty, R. (2020, July). Getting the ##life out of living: How adequate are word-pieces for modelling complex morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 204–209). Association for Computational Linguistics. doi: 10.18653/v1/2020.sigmorphon-1.24

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science*, *10*(1), 110–120. doi: 10.1111/j.1467-7687.2007.00572.x

Kuhl, P. K. (2011). Social mechanisms in early language acquisition: Understanding integrated brain systems supporting language. In J. Decety & J. T. Cacioppo (Eds.), *The Oxford Handbook of Social Neuroscience* (pp. 650–667). doi: 10.1093/oxfordhb/9780195342161.013.0043

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., . . . Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*(5326), 684–686. doi: 10.1126/science.277.5326.684

Kuhl, P. K., Tsao, F. M., & Liu, H. M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, *100*(15), 9096–9101. doi: 10.1073/pnas.1532872100

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in

linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi: 10.18637/jss.v082.i13

Lago, S., Mosca, M., & Stutter Garcia, A. (2021). The role of crosslinguistic influence in multilingual processing: Lexicon versus syntax. *Language Learning*, *71*(S1), 163–192. doi: 10.1111/lang.12412

Lamb, M. E., & Tamis-LeMonda, C. S. (2004). The role of the father: An introduction. In M. E. Lamb (Ed.), *The role of the father in child development* (4th ed., pp. 1–31). John Wiley & Sons. doi: 10.1016/S0266-6138(88)80076-7

Leaper, C., Anderson, K. J., & Sanders, P. (1998). Moderators of gender effects on parents' talk to their children: A meta-analysis. *Developmental psychology*, *34*(1), 3–27. doi: 10.1037/0012-1649.34.1.3

Lehet, M., Arjmandi, M. K., Houston, D., & Dilley, L. (2020). Circumspection in using automated measures: Talker gender and addressee affect error rates for adult speech detection in the Language ENvironment Analysis (LENA) system. *Behavior Research Methods*. doi: 10.3758/s13428-020-01419-y

Levinson, S. C. (2012). The original sin of cognitive science. *Topics in Cognitive Science*, *4*(3), 396–403. doi: 10.1111/j.1756-8765.2012.01195.x

Liu, H. M., Kuhl, P. K., & Tsao, F. M. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science*, *6*(3), F1–F10. doi: 10.1111/1467-7687.00275

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., & Smith, N. A. (2019). Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 1073–1094). doi: 10.18653/v1/N19-1112

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*. doi: 10.48550/arXiv.1907.11692

Luk, G., & Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology*, *25*(5), 605–621. doi: 10.1080/20445911.2013.795574

Lupyan, G., & Winter, B. (2018). Language is more abstract than you think, or, why aren't languages more iconic? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752), 20170137. doi: doi.org/10.1098/rstb.2017.0137

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: A cognitive perspective. *arXiv:2301.06627*. doi: 10.48550/arXiv.2301.06627

Majid, A. (2023). Establishing psychological universals. *Nature Reviews Psychology*, *2*(4), 199–200. doi: 10.1038/s44159-023-00169-w

Majorano, M., Rainieri, C., & Corsano, P. (2013). Parents' child-directed communication and child language development: A longitudinal study with Italian toddlers. *Journal of Child Language*, *40*(4), 836–859. doi: 10.1017/S0305000912000323

Malin, J. L., Cabrera, N. J., & Rowe, M. L. (2014). Low-income minority mothers' and fathers' reading and children's interest: Longitudinal contributions to children's receptive vocabulary skills. *Early Childhood Research Quarterly*, *29*(4), 425–432. doi: 10.1016/j.ecresq.2014.04.010

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*. doi: 10.1044/1092-4388(2007/067)

Marian, V., & Shook, A. (2012). The cognitive benefits of being bilingual. *Cerebrum: The Dana Forum on Brain Science*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3583091/

Martin, A., & Culbertson, J. (2020). Revisiting the suffixing peference: Native-language affixation patterns influence perception of sequences. *Psychological Science*, *31*(9), 1107–1116. doi: 10.1177/0956797620931108

Martin, A., Holtz, A., Abels, K., Adger, D., & Culbertson, J. (2020). Experimental evidence for the influence of structure and meaning on linear order in the noun phrase. *Glossa*, *5*(1), 1–21. doi: 10.5334/gjgl.1085

Martin, A., Ratitamkul, T., Abels, K., Adger, D., & Culbertson, J. (2019). Cross-linguistic evidence for cognitive universals in the noun phrase. *Linguistics Vanguard*, *5*(1), 20180072. doi: 10.1515/lingvan-2018-0072

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In *Proceedings*

of the 2018 conference on empirical methods in natural language processing (pp. 1192–1202). doi: 10.18653/v1/D18-1151

McCarthy, A. D., Vylomova, E., Wu, S., Malaviya, C., Wolf-Sonkin, L., Nicolai, G., ... others (2019). The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th workshop on computational research in phonetics, phonology, and morphology* (pp. 229–244). doi: 0.18653/v1/W19-4226

McDonald, R. T., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., ... others (2013). Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 92–97). Retrieved from https://www.aclweb.org/anthology/P13-2017

McMillan-Major, A., Bender, E. M., & Friedman, B. (2023). Data statements: From technical concept to community practice. *ACM Journal on Responsible Computing*. doi: 10.1145/3594737

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, *90*(5), 862–877. doi: 10.1037/0022-3514.90.5.862

Morford, J. P., Wilkinson, E., Villwock, A., Piñar, P., & Kroll, J. F. (2011). When deaf signers read English: Do written words activate their sign translations? *Cognition*, *118*(2), 286–292. doi: 10.1016/j.cognition.2010.11.006

Morgan-Short, K. (2020). Insights into the neural mechanisms of becoming bilingual: A brief synthesis of second language research with artificial linguistic systems. *Bilingualism: Language and Cognition*, *23*(1), 87–91. doi: 10.1017/S1366728919000701

Musselman, C. (2000). How do children who can't hear learn to read an alphabetic script? a review of the literature on reading and deafness. *Journal of Deaf Studies and deaf education*, *5*(1), 9–31. doi: 10.1093/deafed/5.1.9

Newman, R. S., Rowe, M. L., & Bernstein Ratner, N. (2016). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of Child Language*, *43*(5), 1158–1173. doi: 10.1017/S0305000915000446

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., ... Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings*

of the tenth international conference on language resources and evaluation (lrec'16) (pp. 1659–1666). European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/L16-1262

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., ... Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 4034–4043). Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.497

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., ... Warren, S. F. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences*, *107*(30), 13354–13359. doi: 10.1073/pnas.1003882107

OpenAI. (2022). *Introducing ChatGPT.* Retrieved from https://openai.com/blog/chatgpt

OpenAI. (2023). *Gpt-4 technical report.* doi: 10.48550/arXiv.2303.08774

Orena, A. J., Byers-Heinlein, K., & Polka, L. (2019). Reliability of the language environment analysis recording system in analyzing french–english bilingual speech. *Journal of Speech, Language, and Hearing Research*, *62*(7), 2491–2500. doi: 10.1044/2019_JSLHR-L-18-0342

Ortega, L. (2020). The study of heritage language development from a bilingualism and social justice perspective. *Language Learning*, *70*, 15–53. doi: 10.1111/lang.12347

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language*, *31*(6), 785–806. doi: 10.1016/0749-596X(92)90039-Z

Osterhout, L., Kim, A., & Kuperberg, G. R. (2012). The neurobiology of sentence comprehension. *The Cambridge Handbook of Psycholinguistics*, 365–389. doi: 10.1017/CBO9781139029377.019

Pancsofar, N. (2020). Fathers' language input and early child language development. In H. E. Fitzgerald, K. V. von Klitzing, N. J. Cabrera, J. Scarano de Mendonça, & T. Skjøthaug (Eds.), *Handbook of fathers and child development* (pp. 393–409). Springer Nature Switzerland. doi: 10.1007/978-3-030-51027-5_23

Pancsofar, N., & Vernon-Feagans, L. (2006). Mother and father language input to young children: Contributions to later language development. *Journal of Applied Developmental Psychology*, *27*(6), 571–587. doi: 10.1016/j.appdev.2006.08.003

Pancsofar, N., Vernon-Feagans, L., & The Family Life Project Investigators. (2010). Fathers' early contributions to children's language development in families from low-income rural communities. *Early Childhood Research Quarterly*, *25*(4), 450–463. doi: 10.1016/j.ecresq.2010.02.001

Park, H. H., Zhang, K. J., Haley, C., Steimel, K., Liu, H., & Schwartz, L. (2021). Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, *9*, 261-276. doi: 10.1162/tacl_a_00365

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Retrieved from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pearl, L. (2022). Poverty of the stimulus without tears. *Language Learning and Development*, *18*(4), 415–454. doi: 10.1080/15475441.2021.1981908

Perniss, P., Thompson, R. L., & Vigliocco, G. (2010). Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in Psychology*, *1*, 227. doi: 10.3389/fpsyg.2010.00227

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227–2237). doi: 10.18653/v1/N18-1202

Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. *lingbuzz/007180*. Retrieved from https://lingbuzz.net/lingbuzz/007180

Pimentel, T., Saphra, N., Williams, A., & Cotterell, R. (2020). Pareto probing: Trading off accuracy for complexity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3138–3153). Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.254

Pimentel, T., Valvoda, J., Hall Maudslay, R., Zmigrod, R., Williams, A., & Cotterell, R. (2020). Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4609–4622). Association for

Computational Linguistics. doi: 10.18653/v1/2020.acl-main.420

Pinker, S., & Jackendoff, R. (2005). The faculty of language: What's special about it? *Cognition*, *95*(2), 201–236. doi: 10.1016/j.cognition.2004.08.004

Pires, T., Schlinger, E., & Garrette, D. (2019, July). How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4996–5001). Association for Computational Linguistics. doi: 10.18653/v1/P19-1493

Pleck, J. H. (2010). Paternal involvement: Revised conceptualization and theoretical linkages with child outcomes. In M. E. Lamb (Ed.), *The Role of the Father in Child Development* (5th ed., pp. 58–93). John Wiley & Sons. doi: 10.18653/v1/P19-1493

Pretzer, G. M., Lopez, L. D., Walle, E. A., & Warlaumont, A. S. (2019). Infant-adult vocal interaction dynamics depend on infant vocal type, child-directedness of adult speech, and timeframe. *Infant Behavior and Development*, *57*, 101325. doi: 10.1016/j.infbeh.2019.04.007

Pye, C., Pfeiler, B., De León, L., Brown, P., & Mateo, P. (2007). Roots or edges? Explaining variation in children's early verb forms across five Mayan languages. In *Learning Indigenous Languages: Child Language Acquisition in Mesoamerica* (pp. 15–46). Mouton de Gruyter. doi: 10.1515/9783110923148.15

Pyysalo, S., Kanerva, J., Missilä, A., Laippala, V., & Ginter, F. (2015). Universal Dependencies for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)* (pp. 163–172). Retrieved from https://www.aclweb.org/anthology/W15-1821

Quigley, J., & Nixon, E. (2020). Infant language predicts fathers' vocabulary in infant-directed speech. *Journal of Child Language*, *47*, 146–158. doi: 10.1017/S0305000919000205

Quigley, J., Nixon, E., & Lawson, S. (2019). Exploring the association of infant receptive language and pitch variability in fathers' infant-directed speech. *Journal of Child Language*, *46*(4), 800–811. doi: 10.1017/S0305000919000175

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from https://www.r-project.org/

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. Retrieved from https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners* (Tech. Rep.). OpenAI. Retrieved from https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: Speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental Science*, *17*(6), 880–891. doi: 10.1111/desc.12172

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2017a). The impact of early social interactions on later language development in Spanish–English bilingual infants. *Child Development*, *88*(4), 1216–1234. doi: 10.1111/cdev.12648

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2017b). Look who's talking NOW! Parentese speech, social context, and language development across time. *Frontiers in Psychology*, *8*(1008), 1–12. doi: 10.3389/fpsyg.2017.01008

Ramírez-Esparza, N., Mehl, M. R., Álvarez-Bermúdez, J., & Pennebaker, J. W. (2009). Are Mexicans more or less sociable than Americans? Insights from a naturalistic observation study. *Journal of Research in Personality*, *43*(1), 1–7. doi: 10.1016/j.jrp.2008.09.002

Ranathunga, S., & de Silva, N. (2022). Some languages are more equal than others: Probing deeper into the linguistic disparity in the nlp world. *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 823–848. Retrieved from https://aclanthology.org/2022.aacl-main.62

Ravishankar, V. (2017). A Universal Dependencies Treebank for Marathi. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories* (pp. 190–200). Retrieved from https://www.aclweb.org/anthology/W17-7623

Reynolds, E., Vernon-Feagans, L., Bratsch-Hines, M., Baker, C. E., & the Family Life Project Key Investigators. (2019). Mothers' and fathers' language input from 6 to 36 months in rural two-parent-families: Relations to children's kindergarten achievement. *Early Childhood Research Quarterly*, *47*, 385–395. doi: 10.1016/j.ecresq.2018.09.002

Rice, K. (2011). Principles of affix ordering: An overview. *Word Structure*, *4*(2), 169–200. doi: 10.3366/word.2011.0009

Ritwika, V., Pretzer, G. M., Mendoza, S., Shedd, C., Kello, C. T., Gopinathan, A., & Warlaumont, A. S. (2020). Exploratory dynamics of vocal foraging during infant-caregiver communication. *Scientific Reports*, *10*(1), 10469. doi: 10.1038/s41598-020-66778-0

Romaine, S. (2013). The bilingual and multilingual community. In T. K. Bhatia & W. C. Ritchie (Eds.), *The Handbook of Bilingualism and Multilingualism* (Second ed., pp. 445–465). Blackwell Publishing.

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language*, *35*(1), 185–205. doi: 10.1017/S0305000907008343

Rowe, M. L. (2018). Understanding socioeconomic differences in parents' speech to children. *Child Development Perspectives*, *12*(2), 122–127. doi: 10.1111/cdep.12271

Rowe, M. L., Leech, K. A., & Cabrera, N. J. (2017). Going beyond input quantity: Wh-questions matter for toddlers' language and cognitive development. *Cognitive Science*, *41*, 162–179. doi: 10.1111/cogs.12349

Sadde, S., Seker, A., & Tsarfaty, R. (2018). The Hebrew Universal Dependency Treebank: Past, present and future. In *Proceedings of the 2nd Workshop on Universal Dependencies* (pp. 133–143). doi: 10.18653/v1/w18-6016

Şahin, G. G., Vania, C., Kuznetsov, I., & Gurevych, I. (2020). LINSPECTOR: Multilingual probing tasks for word representations. *Computational Linguistics*, *46*(2), 335-385. doi: 10.1162/coli_a_00376

Saint-Georges, C., Chetouani, M., Cassel, R., Apicella, F., Mahdhaoui, A., Muratori, F., . . . Cohen, D. (2013). Motherese in interaction: At the cross-road of emotion and cognition? (A systematic review). *PLOS ONE*, *8*(10), e78103. doi: 10.1371/journal.pone.0078103

Saldana, C., Oseki, Y., & Culbertson, J. (2021). Cross-linguistic patterns of morpheme order reflect cognitive biases: An experimental study of case and number morphology. *Journal of Memory and Language*, *118*, 104204. doi: 10.1016/j.jml.2020.104204

Saratsli, D., Bartell, S., & Papafragou, A. (2020). Cross-linguistic frequency and the learnability of semantics: Artificial language learning studies of evidentiality. *Cognition*, *197*, 104194. doi: 10.1016/j.cognition.2020.104194

Sassenhagen, J., & Fiebach, C. J. (2019). Finding the P3 in the P600: Decoding shared neural

mechanisms of responses to syntactic violations and oddball targets. *NeuroImage*, *200*, 425–436. doi: 10.1016/j.neuroimage.2019.06.048

Sato, A., Schouwstra, M., Flaherty, M., & Kirby, S. (2020). Do all aspects of learning benefit from iconicity? Evidence from motion capture. *Language and Cognition*, *12*(1), 36–55. doi: 10.1017/langcog.2019.37

Schouwstra, M., & de Swart, H. (2014). The semantic origins of word order. *Cognition*, *131*(3), 431–436. doi: 10.1016/j.cognition.2014.03.004

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green ai. *Communications of the ACM*, *63*(12), 54–63. doi: 10.1145/3381831

Shapiro, N. T., Hippe, D. S., & Ferjan Ramírez, N. (2021). How chatty are daddies? An exploratory study of infants' language environments. *Journal of Speech, Language, and Hearing Research*, *64*(8), 3242–3252. Retrieved from https://pubs.asha.org/doi/abs/10.1044/2021_JSLHR-20-00727 doi: 10.1044/2021_JSLHR-20-00727

Shapiro, N. T., Paullada, A., & Steinert-Threlkeld, S. (2021). A multilabel approach to morphosyntactic probing. In *Findings of the association for computational linguistics: Emnlp 2021* (pp. 4486–4524). Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.findings-emnlp.382 doi: 10.18653/v1/2021.findings-emnlp.382

Shapiro, N. T., & Steinert-Threlkeld, S. (2023). Iconic artificial language learning: A conceptual replication with English speakers. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45). Retrieved from https://escholarship.org/uc/item/7b66s7c6

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy*, *14*(6), 654–666. doi: 10.1080/15250000903263973

Smith, N. A. (2020). Contextual word representations: Putting words into computers. *Communications of the ACM*, *63*(6), 66–74. doi: 10.1145/3347145

Smrž, O., Bielický, V., Kouřilová, I., Kráčmar, J., Hajič, J., & Zemánek, P. (2008). Prague Arabic Dependency Treebank: A word on the million words. In *Proceedings of the workshop on arabic and local languages* (pp. 16–23). Retrieved from https://ufal.mff.cuni.cz/~smrz/LREC2008/padt-lrec.pdf

Smrž, O., Šnaidauf, J., & Zemánek, P. (2002). Prague Dependency Treebank for Arabic: Multi-level

annotation of Arabic corpus. In *Proceedings of the International Symposium on the Processing of Arabic* (pp. 147–155). Retrieved from https://ufal.mff.cuni.cz/padt/PADT_1.0/docs/papers/2002-flm-padt.pdf

Snow, C. E. (1977). Mothers' speech research: From input to interaction. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to Children* (pp. 31–49). Cambridge University Press.

Snow, C. E. (1999). Social perspectives on the emergence of language. In *The Emergence of Language* (pp. 257–276). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers. doi: 10.4324/9781410602367-9

Søgaard, A. (2022). Should we ban English NLP for a year? In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 5254–5260). doi: 10.18653/v1/2022.emnlp-main.351

Song, J. Y., Demuth, K., & Morgan, J. (2010). Effects of the acoustic properties of infant-directed speech on infant word recognition. *Journal of the Acoustical Society of America*, *128*(1), 389–400. doi: 10.1121/1.3419786

St. Clair, M. C., Monaghan, P., & Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science*, *33*(7), 1317–1329. doi: 10.1111/j.1551-6709.2009.01065.x

Stańczak, K., Ponti, E., Hennigen, L. T., Cotterell, R., & Augenstein, I. (2022). Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1589–1598. doi: 10.18653/v1/2022.naacl-main.114

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645–3650). doi: 10.18653/v1/P19-1355

Sulubacak, U., Gökirmak, M., Tyers, F. M., Çöltekin, Ç., Nivre, J., & Eryiğit, G. (2016). Universal Dependencies for Turkish. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3444–3454). Retrieved from https://www.aclweb.org/anthology/C16-1325

Tamis-LeMonda, C. S., Baumwell, L., & Cabrera, N. J. (2012). Fathers' role in children's

language development. In N. J. Cabrera & C. S. Tamis-LeMonda (Eds.), *Handbook of father involvement: Multidisciplinary perspectives* (2nd ed., pp. 135–150). Taylor and Francis Group. doi: 10.4324/9780203101414.ch8

Tamis-LeMonda, C. S., Baumwell, L., & Cristofaro, T. (2012). Parent-child conversations during play. *First Language*, *32*(4), 413–438. doi: 10.1177/0142723711419321

Tartter, V. C. (1980). Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & Psychophysics*, *27*(1), 24–27. doi: 10.3758/BF03199901

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4593–4601). doi: 10.18653/v1/P19-1452

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*(1), 53–71. doi: 10.1207/s15327078in0701_5

Trask, R. L. (1999). *Key Concepts in Language and Linguistics*. Routledge.

Tsarfaty, R. (2013). A unified morpho-syntactic scheme of Stanford Dependencies. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 578–584). Retrieved from https://www.aclweb.org/anthology/P13-2103

Tsujimura, N. (2013). *An Introduction to Japanese Linguistics*. John Wiley & Sons.

Türk, U., Atmaca, F., Betül Özateş, c., Öztürk Başaran, B., Güngör, T., & Özgür, A. (2019). Improving the annotations in the Turkish Universal Dependency Treebank. In *Proceedings of the third workshop on universal dependencies (udw, syntaxfest 2019)* (pp. 108–115). doi: 10.18653/v1/w19-8013

Tyers, F. M., Washington, J., Çöltekin, Ç., & Makazhanov, A. (2017). An assessment of Universal Dependency annotation guidelines for Turkic languages. In *Proceedings of the 5th International Conference on Turkic Language Processing* (pp. 276–297). Retrieved from https://works.swarthmore.edu/fac-linguistics/226

Vance, T. J. (1993). Are Japanese particles clitics? *The Journal of the Association of Teachers of Japanese*, *27*(1), 3–33. doi: 10.2307/489122

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Sys-*

tems 30. Retrieved from https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vygotsky, L. S. (1978). Interaction between learning and development. In M. Gauvain & M. Cole (Eds.), *Readings on the Development of Children* (pp. 79–91). Harvard University Press.

Wang, Y., Williams, R., Dilley, L., & Houston, D. M. (2020). A meta-analysis of the predictability of lena™ automated measures for child language development. *Developmental Review*, *57*, 100921.

Warren-Leubecker, A., & Bohannon III, J. N. (1984). Intonation patterns in child-directed speech: Mother-father differences. *Child Development*, *55*, 1379–1385. doi: 10.2307/1130007

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language* (pp. 17–60). CRC Press.

Weissler, R. E., Drake, S., Kampf, K., Diantoro, C., Foster, K., Kirkpatrick, A., . . . Baese-Berk, M. M. (2023). Examining linguistic and experimenter biases through "non-native" versus "native" speech. *Applied Psycholinguistics*, *44*(4), 460–474. doi: 10.1017/S0142716423000115

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., . . . Rush, A. M. (2019). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). doi: 10.18653/v1/2020.emnlp-demos.6

Woo, B. (2019). *&^{0}: The Syntax and Semantics of 'Slash' and 'And/or'* (Doctoral dissertation, University of Washington). Retrieved from http://hdl.handle.net/1773/44341

Wu, S., & Dredze, M. (2019). Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 833–844). Retrieved from https://www.aclweb.org/anthology/D19-1077 doi: 10.18653/v1/D19-1077

Xu, D., Richards, J. A., & Gilkerson, J. (2014). Automated analysis of child phoetric production using naturalistic recordings. *Journal of Speech, Language, and Hearing Research*, *57*(5), 1638–1650. doi: 10.1044/2014_JSLHR-S-13-0037

Yedetore, A., Linzen, T., Frank, R., & McCoy, R. T. (2023). How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech. In *Proceedings*

of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 9370–9393). Toronto, Canada: Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.521

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., . . . Li, J. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 1–19). Retrieved from https://www.aclweb.org/anthology/K17-3001 doi: 10.18653/v1/K17-3001

Zhang, K. W., & Bowman, S. R. (2018). Language modeling teaches you more syntax than translation does: Lessons learned through auxiliary task analysis. *arXiv:1809.10040*. doi: 10.48550/arXiv.1809.10040

Zimmerman, F. J., Gilkerson, J., Richards, J. A., Christakis, D. A., Xu, D., Gray, S., & Yapanel, U. (2009). Teaching by listening: The importance of adult-child conversations to language development. *Pediatrics*, *124*(1), 342–349. doi: 10.1542/peds.2008-2267

Zwicky, A. M. (1985). Clitics and particles. *Language*, *61*(2), 283–305. doi: 10.2307/414146

Zwicky, A. M., & Pullum, G. K. (1983). Cliticization vs. inflection: English n't. *Language*, *59*(3), 502–513. doi: 10.2307/413900

A Feature labels

Tables A.1 and A.2 list the 166 feature labels we extracted in total across our experiments in Chapter 4. The monolingual probes were trained to extract every morphosyntactically relevant label that was available for a given language in its UD corpus. The multilingual probes focused on a subset of these labels.

Table A.1: The monolingual probes extracted different sets of features, while the multilingual probes extracted a semi-aggregated subset of these features (in bold under "Feature Labels").

Feature Labels	Afrikaans	Croatian	Finnish	Hebrew	Korean	Spanish	Turkish
ADJ	✓	✓	✓	✓	✓	✓	✓
ADP	✓	✓	✓	✓		✓	✓
ADV	✓	✓	✓	✓	✓	✓	✓
AUX	✓	✓	✓	✓	✓	✓	✓
CCONJ	✓	✓	✓	✓	✓	✓	✓
DET	✓	✓		✓	✓	✓	✓
NOUN	✓	✓	✓	✓	✓	✓	✓
NUM	✓	✓	✓	✓	✓	✓	✓
PART	✓	✓			✓	✓	
PRON	✓	✓	✓	✓	✓	✓	✓
PROPN	✓	✓	✓	✓	✓	✓	✓
SCONJ	✓	✓	✓	✓		✓	
VERB	✓	✓	✓	✓	✓	✓	✓
AdjType=Attr	✓						
AdjType=Pred	✓						
AdpType=Post			✓				
AdpType=Prep	✓		✓			✓	

Continuation of Table A.1:

Feature Labels	Afrikaans	Croatian	Finnish	Hebrew	Korean	Spanish	Turkish
AdpType=Preppron						✓	
AdvType=Tim						✓	
Animacy=Anim		✓					
Animacy=Inan		✓					
Aspect=Hab							✓
Aspect=Perf							✓
Aspect=Prog							✓
Aspect=Prosp							✓
Aspect=Rapid							✓
Case=Abe			✓				
Case=Abl			✓				✓
Case=Acc	✓	✓	✓	✓	✓	✓	✓
Case=Ade			✓				
Case=Advb					✓		
Case=All			✓				
Case=Com			✓			✓	
Case=Comp					✓		
Case=Dat		✓				✓	✓
Case=Ela			✓				
Case=Equ							✓
Case=Ess			✓				
Case=Gen		✓	✓	✓	✓		✓
Case=Ill			✓				
Case=Ine			✓				
Case=Ins		✓	✓				✓
Case=Loc		✓					✓
Case=Nom	✓	✓	✓		✓	✓	✓
Case=Par			✓				
Case=Tem				✓			
Case=Tra			✓				
Case=Voc		✓					

Continuation of Table A.1:

Feature Labels	Afrikaans	Croatian	Finnish	Hebrew	Korean	Spanish	Turkish
Clitic=Han			✓				
Clitic=Ka			✓				
Clitic=Kaan			✓				
Clitic=Kin			✓				
Clitic=Ko			✓				
Clitic=Pa			✓				
Clitic=S			✓				
Connegative=Yes			✓				
Definite=Cons				✓			
Definite=Def	✓	✓		✓		✓	
Definite=Ind	✓	✓				✓	
Degree=Abs						✓	
Degree=Cmp	✓	✓	✓			✓	
Degree=Dim	✓						
Degree=Pos	✓	✓	✓				
Degree=Sup	✓	✓	✓			✓	
Derivation=Inen			✓				
Derivation=Ja			✓				
Derivation=Lainen			✓				
Derivation=Llinen			✓				
Derivation=Minen			✓				
Derivation=Sti			✓				
Derivation=Tar			✓				
Derivation=Ton			✓				
Derivation=Ttain			✓				
Derivation=U			✓				
Derivation=Vs			✓				
Echo=Rdp							✓
Evident=Nfh							✓
Form=Adn					✓		
Form=Aux					✓		

Continuation of Table A.1:

Feature Labels	Afrikaans	Croatian	Finnish	Hebrew	Korean	Spanish	Turkish
Form=Compl					✓		
Gender=Fem		✓		✓		✓	
Gender=Masc		✓		✓		✓	
Gender=Neut		✓					
Gender[psor]=Fem		✓					
Gender[psor]=Masc		✓					
Gender[psor]=Neut		✓					
HebBinyan=HIFIL				✓			
HebBinyan=HITPAEL				✓			
HebBinyan=HUFAL				✓			
HebBinyan=NIFAL				✓			
HebBinyan=PAAL				✓			
HebBinyan=PIEL				✓			
HebBinyan=PUAL				✓			
HebExistential=True				✓			
InfForm=1			✓				
InfForm=2			✓				
InfForm=3			✓				
Mood=Cnd		✓	✓			✓	✓
Mood=Des							✓
Mood=Gen							✓
Mood=Imp		✓	✓	✓	✓	✓	✓
Mood=Ind		✓	✓		✓	✓	✓
Mood=Nec							✓
Mood=Opt							✓
Mood=Pot			✓				✓
Mood=Sub						✓	
NumType=Card		✓	✓		✓	✓	✓
NumType=Dist							✓
NumType=Frac						✓	
NumType=Mult		✓					

Continuation of Table A.1:

Feature Labels	Afrikaans	Croatian	Finnish	Hebrew	Korean	Spanish	Turkish
NumType=Ord		✓	✓			✓	✓
Number=Dual							
Number=Plur	✓	✓	✓	✓	✓	✓	✓
Number=Sing	✓	✓	✓	✓		✓	✓
Number[psor]=Plur		✓	✓			✓	✓
Number[psor]=Sing		✓	✓			✓	✓
PartForm=Agt			✓				
PartForm=Neg							
PartForm=Past			✓				
PartForm=Pres			✓				
PartType=Gen	✓						
PartType=Inf	✓						
PartType=Neg	✓						
Person=0			✓				
Person=1	✓	✓	✓	✓	✓	✓	✓
Person=2	✓	✓	✓	✓	✓	✓	✓
Person=3	✓	✓	✓	✓	✓	✓	✓
Person[psor]=1			✓				✓
Person[psor]=2			✓				✓
Person[psor]=3			✓				✓
Polarity=Neg		✓	✓	✓	✓	✓	✓
Polarity=Pos				✓			✓
Polite=Form					✓	✓	✓
Polite=Infm							✓
Poss=Yes	✓	✓				✓	
Prefix=Yes				✓			
PrepCase=Npr						✓	
PrepCase=Pre						✓	
PronType=Art	✓			✓		✓	
PronType=Dem	✓	✓	✓	✓		✓	✓
PronType=Emp				✓			

Continuation of Table A.1:

Feature Labels	Afrikaans	Croatian	Finnish	Hebrew	Korean	Spanish	Turkish
PronType=Ind	✓	✓	✓	✓		✓	✓
PronType=Int	✓	✓	✓	✓	✓	✓	
PronType=Neg		✓				✓	
PronType=Prs	✓	✓	✓	✓		✓	✓
PronType=Rcp			✓				
PronType=Rel	✓	✓	✓			✓	
PronType=Tot		✓				✓	
Reflex=Yes	✓	✓	✓	✓		✓	✓
Subcat=Intr	✓						
Subcat=Prep	✓						
Subcat=Tran	✓						
Tense=Fut				✓	✓	✓	✓
Tense=Imp		✓				✓	
Tense=Past	✓	✓	✓	✓	✓	✓	✓
Tense=Pqp							✓
Tense=Pres	✓	✓	✓			✓	✓
VerbForm=Conv		✓					✓
VerbForm=Fin	✓	✓	✓		✓	✓	
VerbForm=Ger					✓	✓	
VerbForm=Inf	✓	✓	✓	✓		✓	
VerbForm=Part	✓	✓	✓	✓		✓	✓
VerbForm=Vnoun							✓
VerbType=Aux	✓						
VerbType=Cop	✓			✓			
VerbType=Mod	✓			✓			
VerbType=Pas	✓						
Voice=Act		✓	✓	✓			
Voice=Cau					✓		✓
Voice=Mid				✓			
Voice=Pass		✓	✓	✓	✓		✓

Table A.2: The monolingual and multilingual probes were evaluated on seven "held-out" languages.

Feature Labels	Arabic	Chinese	Marathi	Slovenian	Tagalog	Yorùbá
ADJ	✓	✓	✓	✓	✓	✓
ADP	✓	✓	✓	✓	✓	✓
ADV	✓	✓	✓	✓	✓	✓
AUX	✓	✓	✓	✓	✓	✓
CCONJ	✓	✓	✓	✓		✓
DET	✓	✓	✓	✓	✓	✓
NOUN	✓	✓	✓	✓	✓	✓
NUM	✓	✓	✓	✓		✓
PART	✓	✓	✓	✓	✓	✓
PRON	✓	✓	✓	✓	✓	✓
PROPN	✓	✓	✓	✓	✓	✓
SCONJ	✓	✓	✓	✓	✓	✓
VERB	✓	✓	✓	✓	✓	✓
AdjType=Attr						
AdjType=Pred						
AdpType=Post						
AdpType=Prep						
AdpType=Preppron						
AdvType=Tim						
Animacy=Anim						
Animacy=Inan						
Aspect=Hab						
Aspect=Perf						
Aspect=Prog						
Aspect=Prosp						
Aspect=Rapid						
Case=Abe						
Case=Abl						
Case=Acc	✓		✓	✓		✓
Case=Ade						
Case=Advb						

Continuation of Table A.2:

Feature Labels	Arabic	Chinese	Marathi	Slovenian	Tagalog	Yorùbá
Case=All						
Case=Com						
Case=Comp						
Case=Dat			✓	✓	✓	
Case=Ela						
Case=Equ						
Case=Ess						
Case=Gen	✓	✓		✓		✓
Case=Ill						
Case=Ine						
Case=Ins			✓	✓		
Case=Loc			✓	✓	✓	
Case=Nom	✓		✓	✓		✓
Case=Par						
Case=Tem						
Case=Tra						
Case=Voc			✓			
Clitic=Han						
Clitic=Ka						
Clitic=Kaan						
Clitic=Kin						
Clitic=Ko						
Clitic=Pa						
Clitic=S						
Connegative=Yes						
Definite=Cons						
Definite=Def						
Definite=Ind						
Degree=Abs						
Degree=Cmp						
Degree=Dim						

Continuation of Table A.2:

Feature Labels	Arabic	Chinese	Marathi	Slovenian	Tagalog	Yorùbá
Degree=Pos						
Degree=Sup						
Derivation=Inen						
Derivation=Ja						
Derivation=Lainen						
Derivation=Llinen						
Derivation=Minen						
Derivation=Sti						
Derivation=Tar						
Derivation=Ton						
Derivation=Ttain						
Derivation=U						
Derivation=Vs						
Echo=Rdp						
Evident=Nfh						
Form=Adn						
Form=Aux						
Form=Compl						
Gender=Fem	✓		✓	✓	✓	
Gender=Masc	✓		✓	✓	✓	
Gender=Neut			✓	✓		
Gender[psor]=Fem						
Gender[psor]=Masc						
Gender[psor]=Neut						
HebBinyan=HIFIL						
HebBinyan=HITPAEL						
HebBinyan=HUFAL						
HebBinyan=NIFAL						
HebBinyan=PAAL						
HebBinyan=PIEL						
HebBinyan=PUAL						

Continuation of Table A.2:

Feature Labels	Arabic	Chinese	Marathi	Slovenian	Tagalog	Yorùbá
HebExistential=True						
InfForm=1						
InfForm=2						
InfForm=3						
Mood=Cnd				✓		
Mood=Des						
Mood=Gen						
Mood=Imp			✓	✓		
Mood=Ind	✓		✓	✓	✓	
Mood=Nec						
Mood=Opt						
Mood=Pot						
Mood=Sub						
NumType=Card						
NumType=Dist						
NumType=Frac						
NumType=Mult						
NumType=Ord						
Number=Dual						
Number=Plur	✓	✓	✓	✓	✓	✓
Number=Sing	✓		✓	✓	✓	✓
Number[psor]=Plur						
Number[psor]=Sing						
PartForm=Agt						
PartForm=Neg						
PartForm=Past						
PartForm=Pres						
PartType=Gen						
PartType=Inf						
PartType=Neg						
Person=0						

Continuation of Table A.2:

Feature Labels	Arabic	Chinese	Marathi	Slovenian	Tagalog	Yorùbá
Person=1	✓	✓	✓	✓	✓	✓
Person=2	✓	✓	✓	✓		✓
Person=3	✓	✓	✓	✓	✓	✓
Person[psor]=1						
Person[psor]=2						
Person[psor]=3						
Polarity=Neg	✓	✓	✓	✓	✓	
Polarity=Pos						
Polite=Form						
Polite=Infm						
Poss=Yes						
Prefix=Yes						
PrepCase=Npr						
PrepCase=Pre						
PronType=Art						
PronType=Dem	✓		✓	✓	✓	✓
PronType=Emp						
PronType=Ind				✓		✓
PronType=Int			✓	✓		✓
PronType=Neg				✓		
PronType=Prs	✓		✓	✓	✓	✓
PronType=Rcp						
PronType=Rel	✓		✓	✓		✓
PronType=Tot				✓		
Reflex=Yes						
Subcat=Intr						
Subcat=Prep						
Subcat=Tran						
Tense=Fut			✓	✓		
Tense=Imp						
Tense=Past			✓			

Continuation of Table A.2:

Feature Labels	Arabic	Chinese	Marathi	Slovenian	Tagalog	Yorùbá
Tense=Pqp						
Tense=Pres			✓	✓		
VerbForm=Conv						
VerbForm=Fin						
VerbForm=Ger						
VerbForm=Inf						
VerbForm=Part						
VerbForm=Vnoun						
VerbType=Aux						
VerbType=Cop						
VerbType=Mod						
VerbType=Pas						
Voice=Act	✓					
Voice=Cau						
Voice=Mid						
Voice=Pass	✓	✓				

B Monolingual and multilingual performance

In the following pages, Figures B.1 though B.7 report the global and feature-level F_1 results for the monolingual and multilingual probes from Chapter 4. In the monolingual experiments, we trained separate probes for Afrikaans, Croatian, Finnish, Hebrew, Korean, Spanish, and Turkish. In a set of multilingual experiments, we then trained probes on a shuffled combination of the training data from the monolingual probes. However, we excluded the Korean dataset from these experiments, due to the lack of documentation on its construction.

Figure B.1: Afrikaans F_1 results

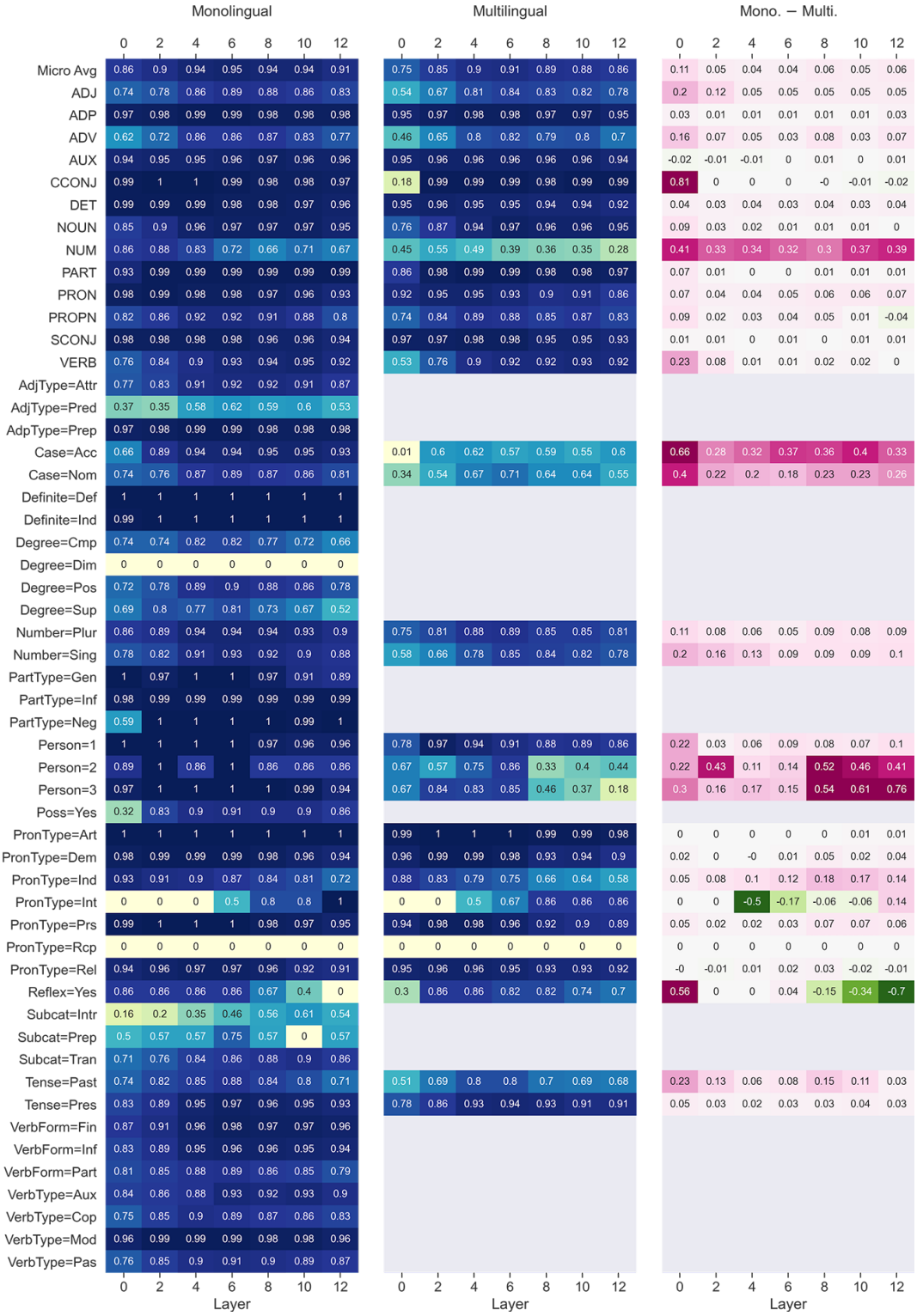
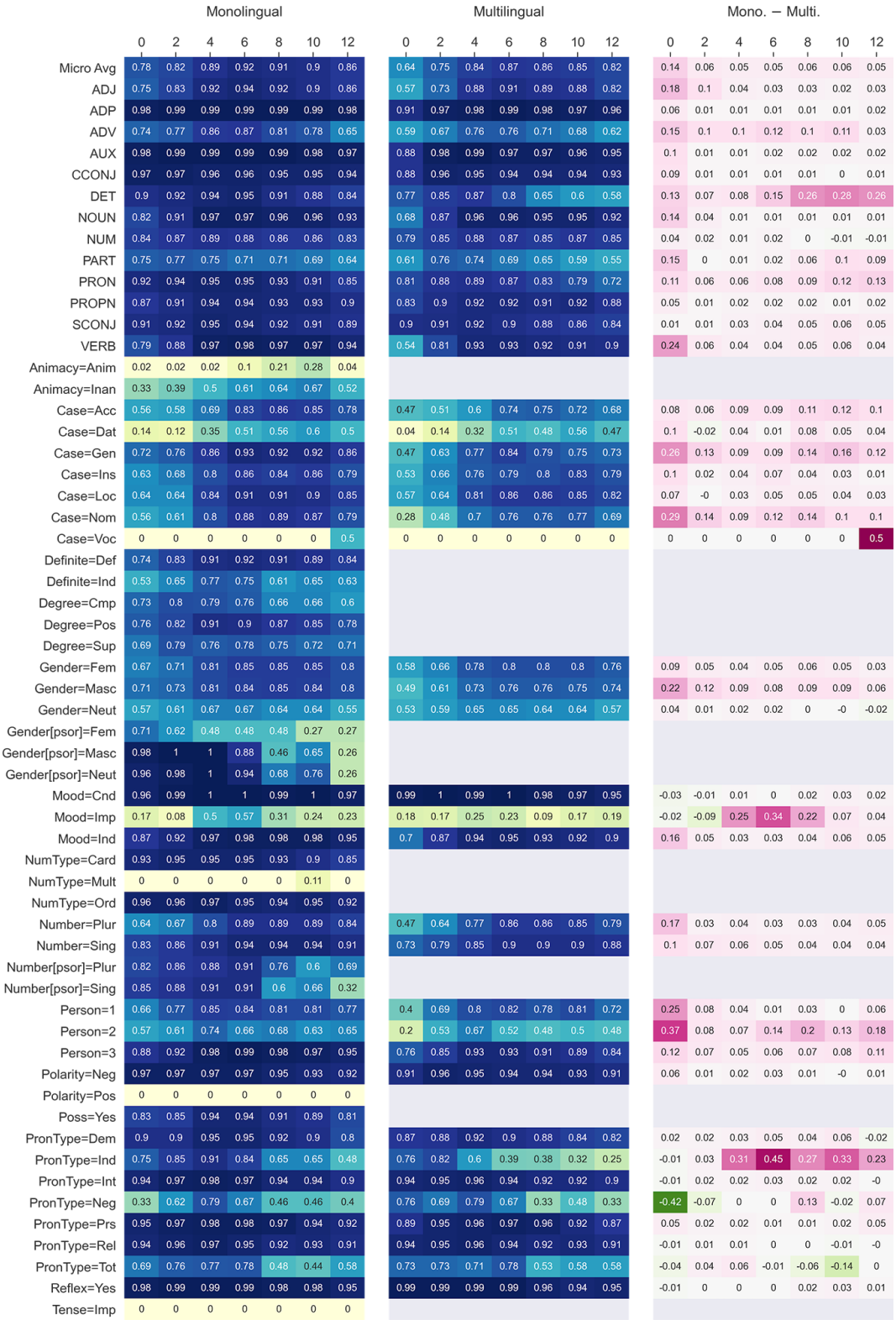


Figure B.2: Croatian F_1 results



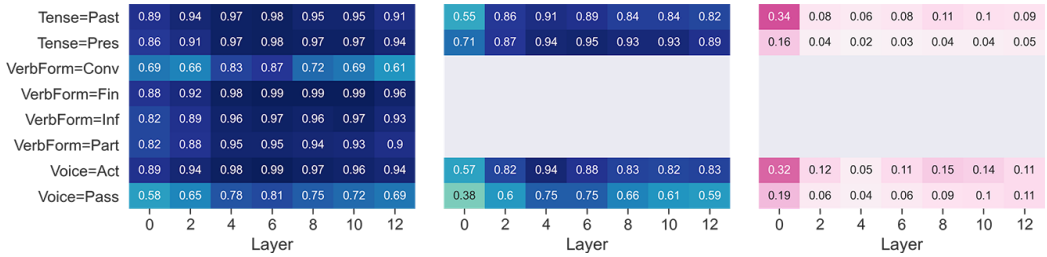
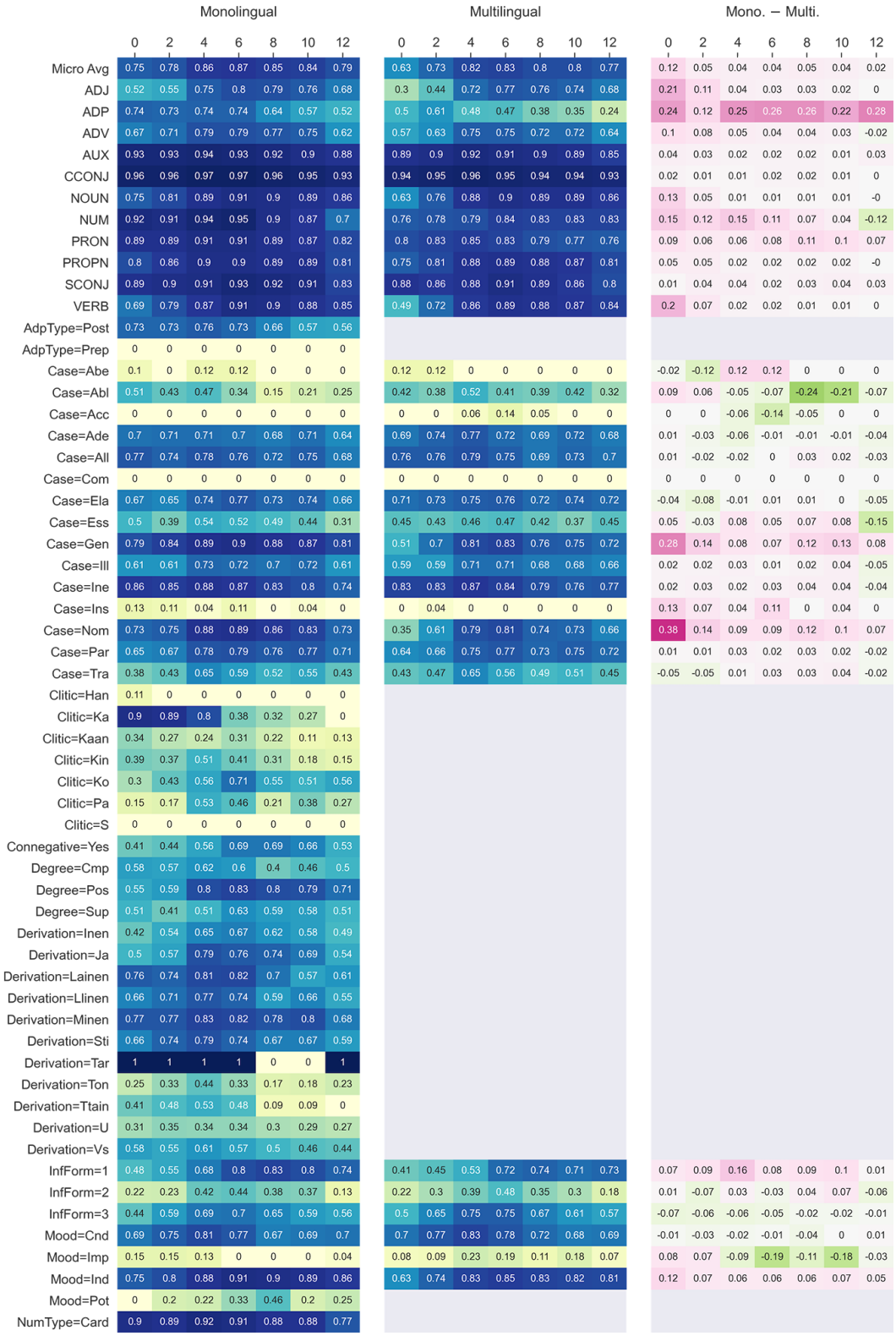

Tense=Past
Tense=Pres
VerbForm=Conv
VerbForm=Fin
VerbForm=Inf
VerbForm=Part
Voice=Act
Voice=Pass
0.89 0.94 0.97 0.98 0.95 0.95 0.91
0.86 0.91 0.97 0.98 0.97 0.97 0.94
0.69 0.66 0.83 0.87 0.72 0.69 0.61
0.88 0.92 0.98 0.99 0.99 0.99 0.96
0.82 0.89 0.96 0.97 0.96 0.97 0.93
0.82 0.88 0.95 0.95 0.94 0.93 0.9
0.89 0.94 0.98 0.99 0.97 0.96 0.94
0.58 0.65 0.78 0.81 0.75 0.72 0.69
0.55 0.86 0.91 0.89 0.84 0.84 0.82
0.71 0.87 0.94 0.95 0.93 0.93 0.89
0.57 0.82 0.94 0.88 0.83 0.82 0.83
0.38 0.6 0.75 0.75 0.66 0.61 0.59
0.34 0.08 0.06 0.08 0.11 0.1 0.09
0.16 0.04 0.02 0.03 0.04 0.04 0.05
0.32 0.12 0.05 0.11 0.15 0.14 0.11
0.19 0.06 0.04 0.06 0.09 0.1 0.11
0 2 4 6 8 10 12
Layer
0 2 4 6 8 10 12
Layer
0 2 4 6 8 10 12
Layer

Figure B.3: Finnish F_1 results



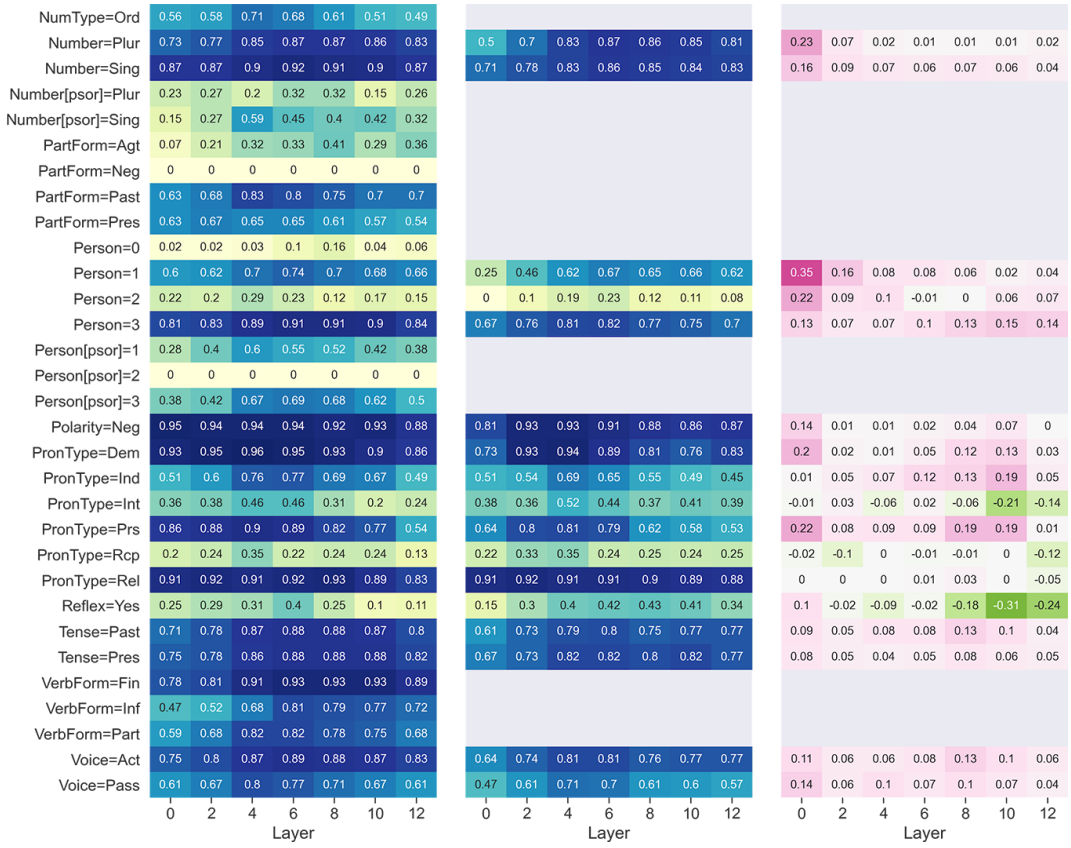

NumType=Ord 0.56 0.58 0.71 0.68 0.61 0.51 0.49
Number=Plur 0.73 0.77 0.85 0.87 0.87 0.86 0.83 0.5 0.7 0.83 0.87 0.86 0.85 0.81 0.23 0.07 0.02 0.01 0.01 0.01 0.02
Number=Sing 0.87 0.87 0.9 0.92 0.91 0.9 0.87 0.71 0.78 0.83 0.86 0.85 0.84 0.83 0.16 0.09 0.07 0.06 0.07 0.06 0.04
Number[psor]=Plur 0.23 0.27 0.2 0.32 0.32 0.15 0.26
Number[psor]=Sing 0.15 0.27 0.59 0.45 0.4 0.42 0.32
PartForm=Agt 0.07 0.21 0.32 0.33 0.41 0.29 0.36
PartForm=Neg 0 0 0 0 0 0 0
PartForm=Past 0.63 0.68 0.83 0.8 0.75 0.7 0.7
PartForm=Pres 0.63 0.67 0.65 0.65 0.61 0.57 0.54
Person=0 0.02 0.02 0.03 0.1 0.16 0.04 0.06
Person=1 0.6 0.62 0.7 0.74 0.7 0.68 0.66 0.25 0.46 0.62 0.67 0.65 0.66 0.62 0.35 0.16 0.08 0.08 0.06 0.02 0.04
Person=2 0.22 0.2 0.29 0.23 0.12 0.17 0.15 0 0.1 0.19 0.23 0.12 0.11 0.08 0.22 0.09 0.1 -0.01 0 0.06 0.07
Person=3 0.81 0.83 0.89 0.91 0.91 0.9 0.84 0.67 0.76 0.81 0.82 0.77 0.75 0.7 0.13 0.07 0.07 0.1 0.13 0.15 0.14
Person[psor]=1 0.28 0.4 0.6 0.55 0.52 0.42 0.38
Person[psor]=2 0 0 0 0 0 0 0
Person[psor]=3 0.38 0.42 0.67 0.69 0.68 0.62 0.5
Polarity=Neg 0.95 0.94 0.94 0.94 0.92 0.93 0.88 0.81 0.93 0.93 0.91 0.88 0.86 0.87 0.14 0.01 0.01 0.02 0.04 0.07 0
PronType=Dem 0.93 0.95 0.96 0.95 0.93 0.9 0.86 0.73 0.93 0.94 0.89 0.81 0.76 0.83 0.2 0.02 0.01 0.05 0.12 0.13 0.03
PronType=Ind 0.51 0.6 0.76 0.77 0.69 0.67 0.49 0.51 0.54 0.69 0.65 0.55 0.49 0.45 0.01 0.05 0.07 0.12 0.13 0.19 0.05
PronType=Int 0.36 0.38 0.46 0.46 0.31 0.2 0.24 0.38 0.36 0.52 0.44 0.37 0.41 0.39 -0.01 0.03 -0.06 0.02 -0.06 -0.21 -0.14
PronType=Prs 0.86 0.88 0.9 0.89 0.82 0.77 0.54 0.64 0.8 0.81 0.79 0.62 0.58 0.53 0.22 0.08 0.09 0.09 0.19 0.19 0.01
PronType=Rcp 0.2 0.24 0.35 0.22 0.24 0.24 0.13 0.22 0.33 0.35 0.24 0.25 0.24 0.25 -0.02 -0.1 0 -0.01 -0.01 0 -0.12
PronType=Rel 0.91 0.92 0.91 0.92 0.93 0.89 0.83 0.91 0.92 0.91 0.91 0.9 0.89 0.88 0 0 0 0.01 0.03 0 -0.05
Reflex=Yes 0.25 0.29 0.31 0.4 0.25 0.1 0.11 0.15 0.3 0.4 0.42 0.43 0.41 0.34 0.1 -0.02 -0.09 -0.02 -0.18 -0.31 -0.24
Tense=Past 0.71 0.78 0.87 0.88 0.88 0.87 0.8 0.61 0.73 0.79 0.8 0.75 0.77 0.77 0.09 0.05 0.08 0.08 0.13 0.1 0.04
Tense=Pres 0.75 0.78 0.86 0.88 0.88 0.88 0.82 0.67 0.73 0.82 0.82 0.8 0.82 0.77 0.08 0.05 0.04 0.05 0.08 0.06 0.05
VerbForm=Fin 0.78 0.81 0.91 0.93 0.93 0.93 0.89
VerbForm=Inf 0.47 0.52 0.68 0.81 0.79 0.77 0.72
VerbForm=Part 0.59 0.68 0.82 0.82 0.78 0.75 0.68
Voice=Act 0.75 0.8 0.87 0.89 0.88 0.87 0.83 0.64 0.74 0.81 0.81 0.76 0.77 0.77 0.11 0.06 0.06 0.08 0.13 0.1 0.06
Voice=Pass 0.61 0.67 0.8 0.77 0.71 0.67 0.61 0.47 0.61 0.71 0.7 0.61 0.6 0.57 0.14 0.06 0.1 0.07 0.1 0.07 0.04
0 2 4 6 8 10 12
Layer
0 2 4 6 8 10 12
Layer
0 2 4 6 8 10 12
Layer

Figure B.4: Hebrew F_1 results

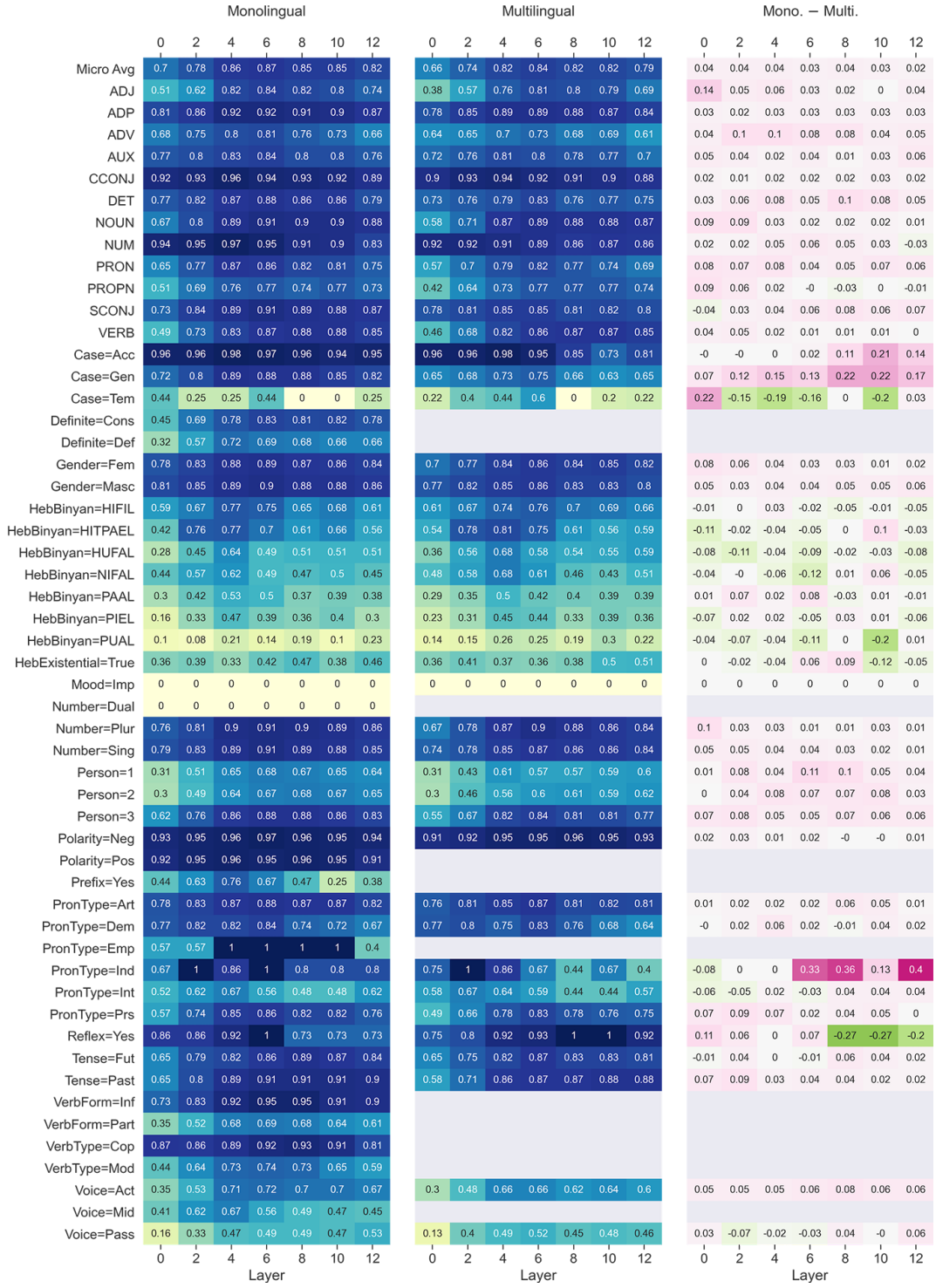


Figure B.5: Korean F_1 results

Due to the lack of documentation on Korean PUD corpus, no multilingual probes were trained on Korean.

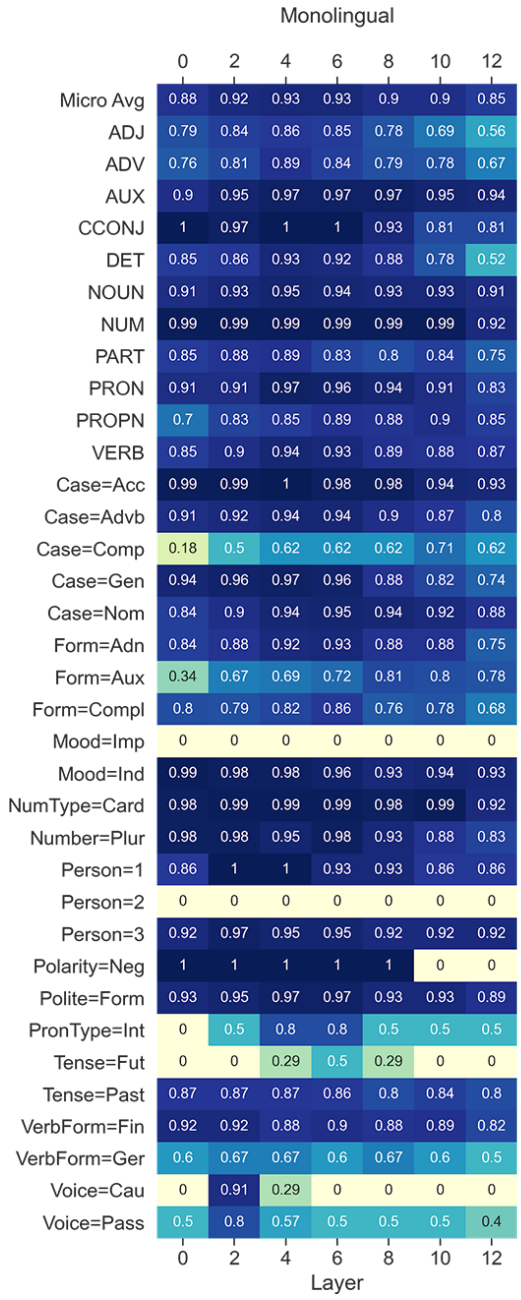
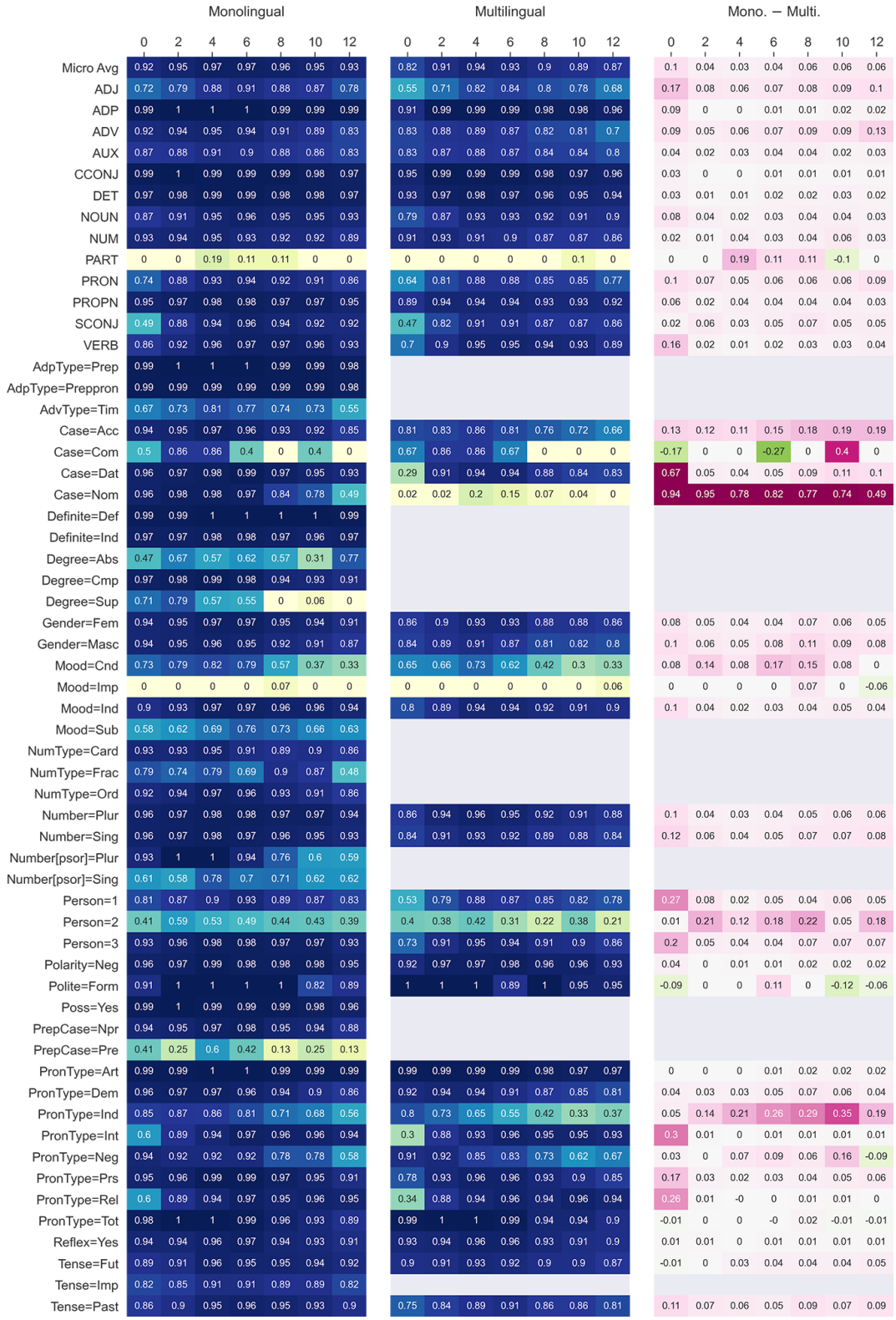


Figure B.6: Spanish F_1 results



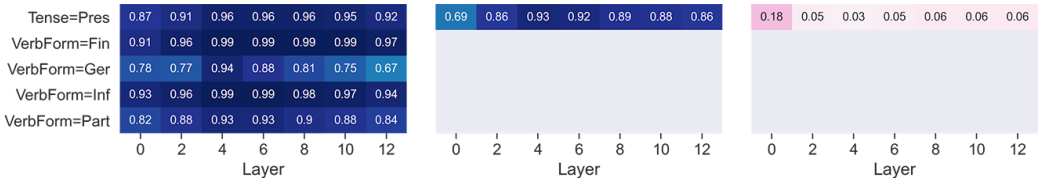
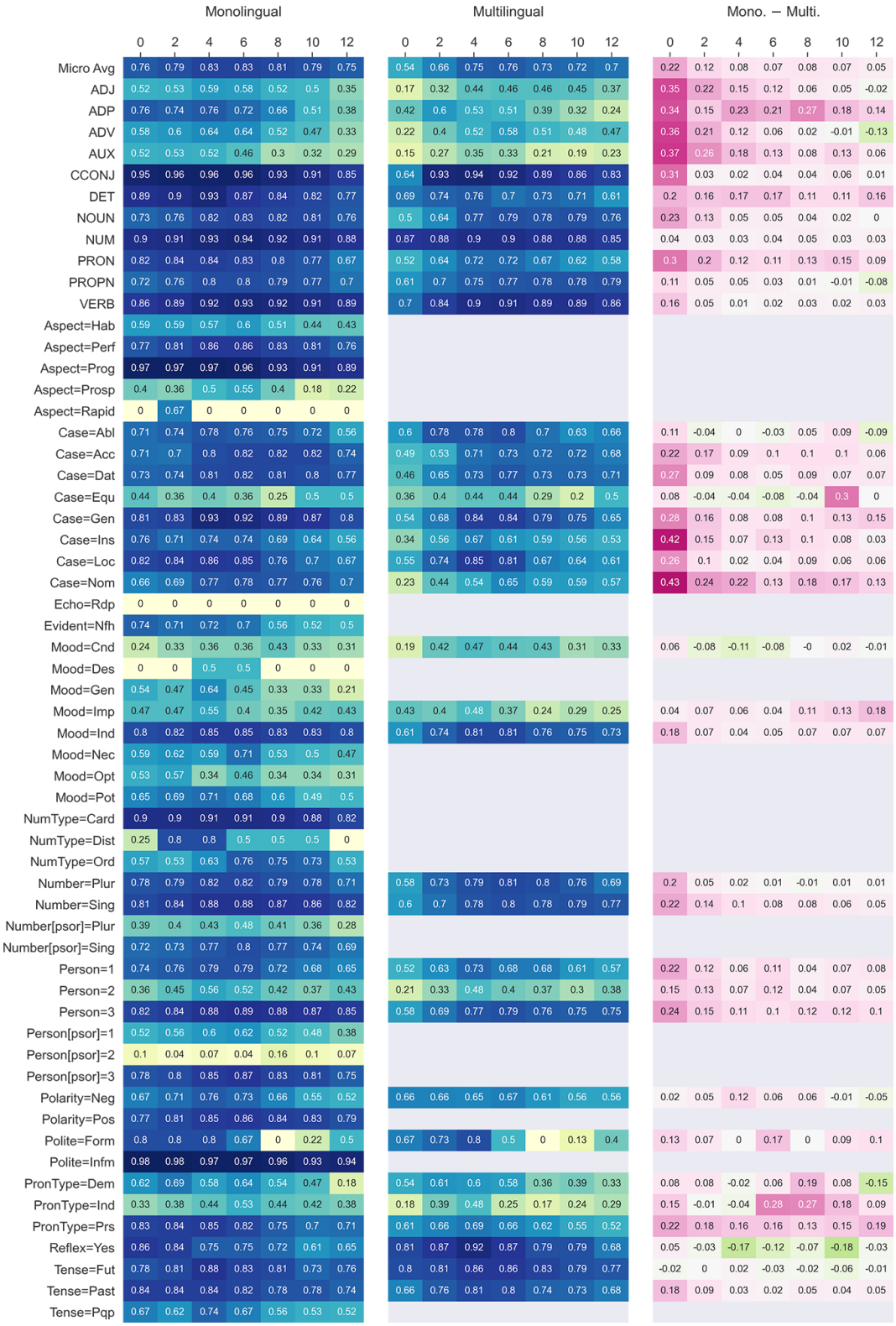
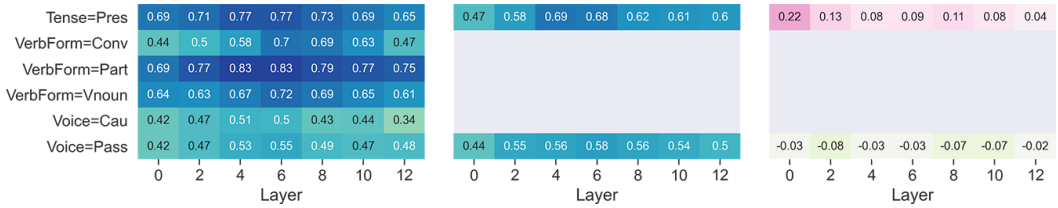

Tense=Pres
VerbForm=Fin
VerbForm=Ger
VerbForm=Inf
VerbForm=Part
0.87 0.91 0.96 0.96 0.96 0.95 0.92
0.91 0.96 0.99 0.99 0.99 0.99 0.97
0.78 0.77 0.94 0.88 0.81 0.75 0.67
0.93 0.96 0.99 0.99 0.98 0.97 0.94
0.82 0.88 0.93 0.93 0.9 0.88 0.84
0 2 4 6 8 10 12
Layer
0.69 0.86 0.93 0.92 0.89 0.88 0.86
0 2 4 6 8 10 12
Layer
0.18 0.05 0.03 0.05 0.06 0.06 0.06
0 2 4 6 8 10 12
Layer

Figure B.7: Turkish F_1 results

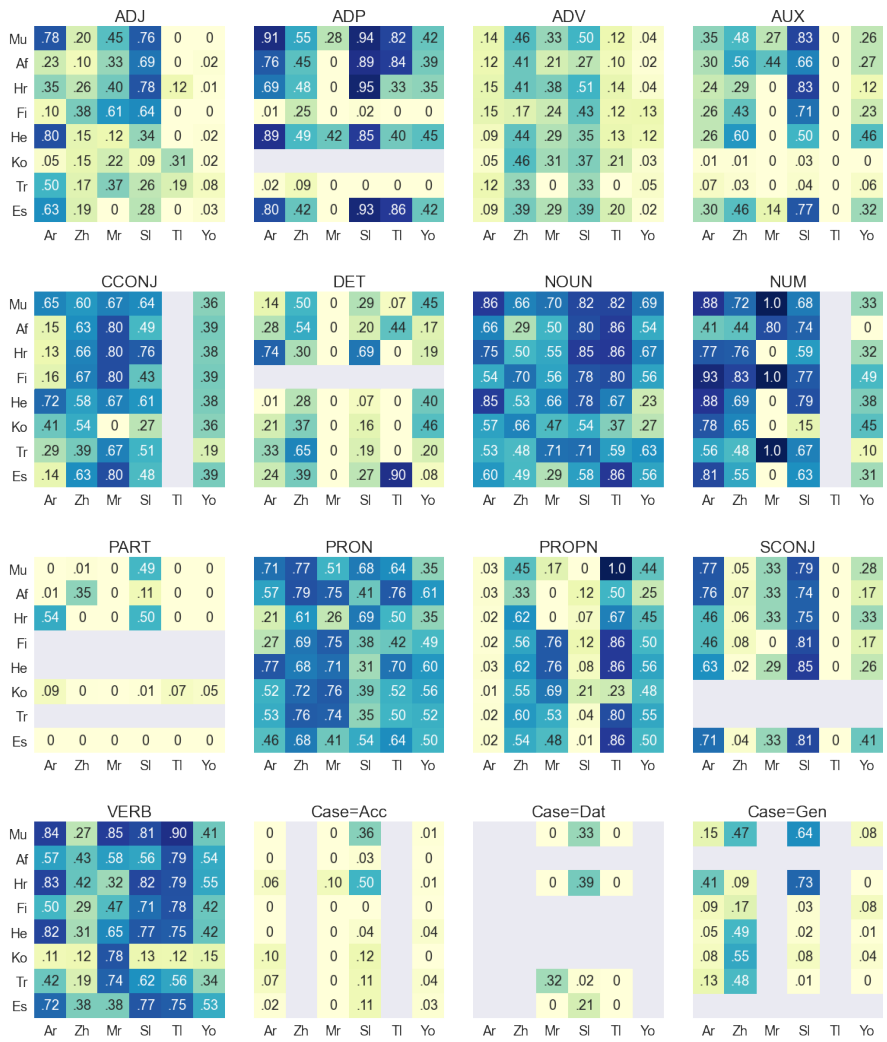


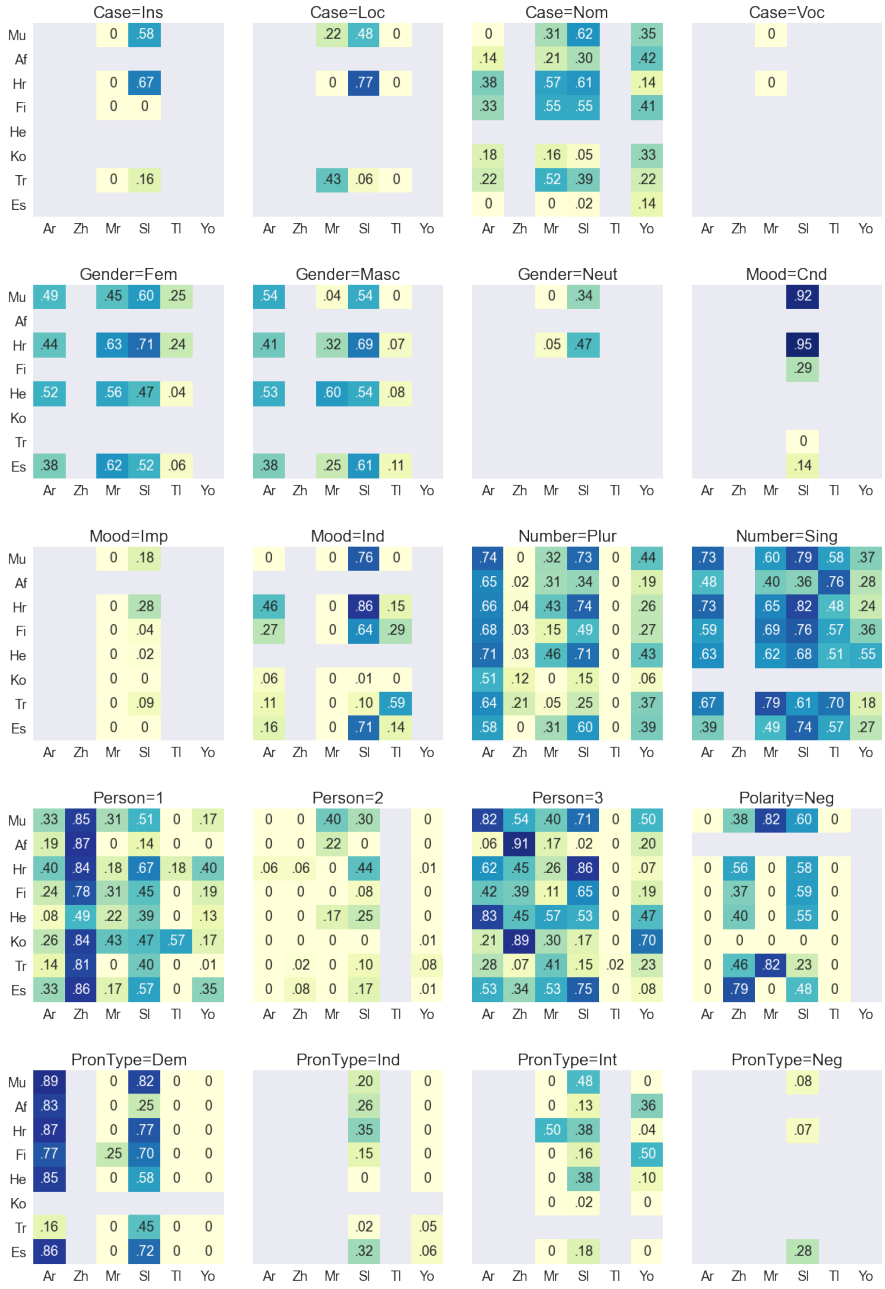

Tense=Pres
VerbForm=Conv
VerbForm=Part
VerbForm=Vnoun
Voice=Cau
Voice=Pass
0.69 0.71 0.77 0.77 0.73 0.69 0.65
0.44 0.5 0.58 0.7 0.69 0.63 0.47
0.69 0.77 0.83 0.83 0.79 0.77 0.75
0.64 0.63 0.67 0.72 0.69 0.65 0.61
0.42 0.47 0.51 0.5 0.43 0.44 0.34
0.42 0.47 0.53 0.55 0.49 0.47 0.48
0 2 4 6 8 10 12
Layer
0.47 0.58 0.69 0.68 0.62 0.61 0.6
0.44 0.55 0.56 0.58 0.56 0.54 0.5
0 2 4 6 8 10 12
Layer
0.22 0.13 0.08 0.09 0.11 0.08 0.04
-0.03 -0.08 -0.03 -0.03 -0.07 -0.07 -0.02
0 2 4 6 8 10 12
Layer

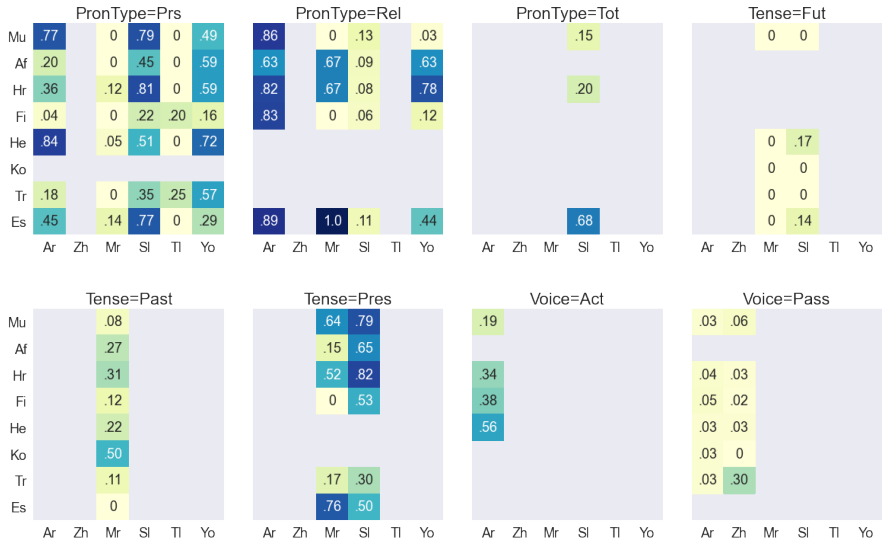
C Crosslingual performance

Figure C.1 shows the feature-level F_1 results from evaluating the mBERT-6 probes on the held-out languages in Chapter 4.

Figure C.1: Crosslingual F_1 results




Case=Ins
Case=Loc
Case=Nom
Case=Voc
Gender=Fem
Gender=Masc
Gender=Neut
Mood=Cnd
Mood=Imp
Mood=Ind
Number=Plur
Number=Sing
Person=1
Person=2
Person=3
Polarity=Neg
PronType=Dem
PronType=Ind
PronType=Int
PronType=Neg
Mu
Af
Hr
Fi
He
Ko
Tr
Es
Ar Zh Mr Sl Tl Yo


PronType=Prs
Mu .77 0 .79 0 .49
Af .20 0 .45 0 .59
Hr .36 .12 .81 0 .59
Fi .04 0 .22 .20 .16
He .84 .05 .51 0 .72
Ko
Tr .18 0 .35 .25 .57
Es .45 .14 .77 0 .29
Ar Zh Mr Sl Tl Yo
PronType=Rel
.86 0 .13 .03
.63 .67 .09 .63
.82 .67 .08 .78
.83 0 .06 .12
.89 1.0 .11 .44
Ar Zh Mr Sl Tl Yo
PronType=Tot
.15
.20
.68
Ar Zh Mr Sl Tl Yo
Tense=Fut
0 0
0 .17
0 0
0 0
0 .14
Ar Zh Mr Sl Tl Yo
Tense=Past
.08
.27
.31
.12
.22
.50
.11
0
Ar Zh Mr Sl Tl Yo
Tense=Pres
.64 .79
.15 .65
.52 .82
0 .53
.17 .30
.76 .50
Ar Zh Mr Sl Tl Yo
Voice=Act
.19
.34
.38
.56
Ar Zh Mr Sl Tl Yo
Voice=Pass
.03 .06
.04 .03
.05 .02
.03 .03
.03 0
.03 .30
Ar Zh Mr Sl Tl Yo