Language Documentation & Conservation Special Publication 25 179
Doing corpus-based typology with spoken language data: State of the art
edited by Geoffrey Haig, Stefan Schnell, and Frank Seifart, pp. 179 – 232
https://nflrc.hawaii.edu/ldc/sp25
https://hdl.handle.net/10125/74661

Language vs. individuals in cross-linguistic corpus typology

Danielle Barth,[1,2] Nicholas Evans,[1,2] I Wayan Arka,[1,2] Henrik Bergqvist,[3] Diana Forker,[4] Sonja Gipper,[5] Gabrielle Hodge,[6] Eri Kashima,[7] Yuki Kasuga,[8] Carine Kawakami,[9] Yukinori Kimoto,[10] Dominique Knuchel,[11] Norikazu Kogura,[9] Keita Kurabe,[9] John Mansfield,[12] Heiko Narrog,[13] Desak P. Eka Pratiwi,[14] Saskia van Putten,[15] Chikako Senge,[16] Olena Tykhostup [4]

[1] *Australian National University,* [2] *Centre of Excellence for the Dynamics of Language,* [3] *Stockholms Universitet,* [4] *Friedrich-Schiller-Universität Jena,* [5] *Universität zu Köln,* [6] *University College London,* [7] *University of Helsinki,* [8] *independent researcher,* [9] *Tokyo University of Foreign Studies,* [10] *University of Hyogo,* [11] *Universität Bern,* [12] *University of Melbourne,* [13] *Tohoku University,* [14] *Mahasaraswati Denpasar University,* [15] *Radboud Universitet,* [16] *University of Turku*

Abstract

There is a long tradition in linguistics of seeing each language as a powerful factor setting out predetermining grooves in how people express themselves. But how strong is this effect? We know that despite the forces of linguistic habit people nonetheless enjoy some freedom in formulating their thoughts. Can we measure the relative contributions of language structures and individual variation to how people formulate statements about the world? Do accounts of typological differences need to take individual vari-

ation into account, and is such variation more prevalent in some kinds of linguistic domains than others? In this paper, we deploy a parallax corpus across thirteen languages from around the world and explore four case studies of linguistic choice, two grammatical and two semantic. We assess whether differences are accounted adequately just by individual participant variation, just by language information, or whether taking into account both helps account for the patterns we see. We do this through comparisons of statistical models. Our results make it clear that participants using the same language do not always behave similarly and this is especially true of our semantic variables. We take this to be a strong caution that the behaviour of individual participants should be considered when making typological generalisations, but also as an exciting outcome that corpus typology as a field can help us account for intra- and inter-language variation.

Keywords: social cognition, corpus-based typology, *Family Problems* picture task, Sapir-Whorf hypothesis, model comparison

1 Introduction

There is a long tradition in linguistics of seeing each language – in particular its grammar and its semantic categories – as a powerful factor setting out predetermining grooves in how people express themselves. For Whorf (1956: 221) "users of markedly different grammars are pointed by their grammars toward different types of observations and different evaluations of externally similar acts of observations", for Levinson (2003: 325) "learning a language seems to play an important role in restricting cognition. A language 'canalizes' the mental landscape", and dozens of comparable quotes could be given, including a chapter title by one of the authors of this article: *Trellises of the mind: How language trains thought* (Evans 2010: 159).[1]

1 First we would like to thank the speakers and signers of all the languages represented in this study, both for their participation in the Family Problems Task and, in many cases, for teaching us their languages over many years. This work originated as a project funded by the Australian Research Council (*Language and Social Cognition: The Design Resources*

But – even if it is at least partially true – how strong is this effect? We know that humans can think, and talk, outside the box, and move beyond the grooves of their language's habitual structures and categories. In other words, despite the forces of linguistic habit, and pre-evolved ease of encoding, people nonetheless enjoy some freedom in formulating their thoughts. We ask:

a. Can we measure the relative contributions of language structures and individual variation to how people formulate statements about the world?

b. Do accounts of typological differences need to take individual variation into account, and is such variation more prevalent in some kinds of linguistic domains than others?

We see the role of individual language knowledge, the utterances they produce, and language-as-social-code, as a multidirectional set of stochastic causal interactions, unfolding through time at a number of different timescales (cf. Enfield 2014). Language, as well as culture, influences the express-

---

of Grammatical Diversity, DP0878126), when the actual task was developed and many recordings made. Continuation into the present phase of analysis, in particular the employment of Barth as a postdoc and the funding of the ongoing workshops bringing together language-specific investigators for "annotation jams", was made possible by an Anneliese-Maier Forschungspreis awarded to Evans by the Alexander von Humboldt Foundation and the German Federal Ministry of Education and Research, plus support from the ARC Research Centre for the Dynamics of Language (CoEDL), funded by the Australian Research Council (CE140100041) and the School of Culture History and Language, ANU College of Asia and the Pacific. Sonja Gipper received funding from the Deutsche Forschungsgemeinschaft (DFG project number 275274422, reference number GI 1110/1-1). We thank the above-named institutions for their generous support of our research. We also thank universities who have hosted SCOPIC project meetings: Otto-Friedrich-Universität Bamberg, Universität Leipzig, Universität zu Köln, Stockholms Universitet and The Research Institute for Languages and Cultures of Asia and Africa at Tokyo University of Foreign Studies. We also thank Naijing Liu for her meticulous job in formatting the manuscript, Nick Thieberger for good advice along the way, the participants at the DGfS workshop *Corpus-based typology: Spoken language from a cross-linguistic perspective* in Hamburg in March 2020 for their questions and comments, Geoffrey Haig and Stefan Schnell for organising the workshop, and Frank Seifart and an anonymous reviewer for their usefully critical comments on an earlier version of this manuscript.

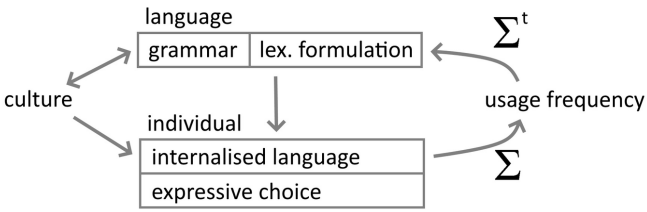

Figure 1 Language and individuals affect usage frequency and each other. Figure produced by Wolfgang Barth.

ive choices of individuals, which we can measure in terms of usage frequencies. Over time, these feed back to change language in turn (cf. Figure 1).

We pursue our questions in this article by drawing on a "parallax" corpus of thirteen languages from around the world (the SCOPIC corpus; see below) which offers people freedom of naturalistic formulation around a series of common problem-solving tasks.

The primary strength of corpus linguistics is its ability to identify variable patterns. When we look at usage across a community, and repeated usage from individuals within a community, we see that there are very few absolutes. Language users (a term we use to include speakers of spoken languages and signers of signed languages) can have individual preferences. We may well see tendencies across communities, but there is also variation at all levels of linguistic organisation, including grammar and vocabulary. The second strength of corpus linguistics is its ability to identify the factors that condition variability. When we examine patterns, we see that variation is rarely random: Characteristics of genre, context and language users all constrain variability and make outcomes more or less likely. This variation, in turn, enables corpus linguistics to show us how particular grammatical innovations arise from the flux of variation (Croft 2010).

Yet, as is clear from the rationale for this special issue, corpus linguistics in general, and the documentation of variable usage in particular, has been rather narrow in its choice of languages, dominated by studies of English, Mandarin, Arabic, Spanish and Japanese which together account for 85% of the corpora available (Anand et al. 2010). We should expect that at least the

same kind of variability is present in the language use of under-resourced and less studied languages as well. Corpus typology or "token-based typology" (Levshina 2019), as a new but growing field, builds on the idea of corpus linguistics, trying to capture and account for variability both between and within languages.

In this paper, we compare a parallax corpus across thirteen languages from around the world. In an earlier publication (Barth & Evans 2017: 1) we defined a parallax corpus as one containing "broadly comparable formulations resulting from a comparable task", to avoid the implications, associated with the term "parallel corpus", that there will be exact semantic equivalence across languages. The multilingual corpus built on the *Pear* stories (Chafe 1980) is also a parallax corpus on our definition, and formed the basis for the study by Croft (2010) mentioned above, but by getting speakers to comment on a pre-existing film it presents a ready-made plot, whereas the SCOPIC corpus growing out of the *Family Problems* picture task (San Roque et al. 2012) puts speakers in the driver's seat in terms of plot-formation, often generating lively debate among the participants about the narrative they are constructing and hence introducing more naturalistic conversation. More information about the task on which this corpus is based will be given in Section 2 below. We use this cross linguistic corpus – with closely-matched elements, but giving complete freedom of formulation to each individual – to assess the relative importance of language, individual, and genre factors in accounting for the total variability of the data.

Returning to the claims about linguistic relativity with which we introduced the paper, these need to be evaluated against specific domains, such as space (Levinson 2003). Within the chosen domain, a cross-linguistic typology of possible conceptualisations or expressive options is elaborated (e.g. the well-known division between egocentric [left / right] and allocentric [north / south; uphill / downhill] reckoning in terms of spatial orientation). Then, across a sample of languages, some standard procedure is used to see which linguistic devices people employ to talk about the domain (or, in some cases, to carry out various forms of non-verbal task like reproducing a spatial layout). Language becomes the independent variable, and particular responses the dependent one; models of causality build in a range of intermediate effects including culture (e.g. differential socialisation of attention) and cognition. See Lucy (1992) for a foundational discussion of the logic of this approach.

Within this neo-Whorfian tradition, most research has focused on aspects of the physical world, such as space (Levinson 2003; Nuñez et al. 2019) or colour (Winawer et al. 2007). Further, it has involved quite different data-gathering paradigms, such as eliciting short isolated descriptions or "director-matching" tasks, or non-verbal material such as reproduction of remembered spatial layouts or reaction times to "same-versus-different" questions about colour stimuli (Winawer et al. 2007). Our research differs from this both in its choice of semantic domain – social cognition – and in its orientation to gathering a more substantial corpus that includes conversation and narrative as well as more focussed descriptions. We chose the domain of social cognition for three main reasons, (i) it is the domain for which we already had a good parallax corpus, (ii) there is prima facie evidence (based on our own knowledge of a range of languages) that social cognition is particularly amenable to cross-cultural and cross-linguistic variation, (iii) it is increasingly clear that social cognition has played a central role in the evolution of both language and the human cognitive and cultural explosions (e.g. Tomasello 2014; Enfield & Levinson 2006), so that understanding the cross-linguistic variability in the coding of this domain is a prerequisite for linguistically realistic models of how human cognitive capacity, and culture, co-evolved.

We focus on two grammatical domains (propositional framing / complementation and direct / indirect speech) and two semantic–lexical variables (formulation of reference, and choice of cognitive or speech-act verbs in representing thought).[2] By hypothesis, we expect fewer differences between language users when we investigate grammatical variables (since they form part of a closed contrast set) and more differences when we investigate semantic variables. We also assess whether differences are best accounted for just by individual participant variation, or whether taking into account language differences also helps account for the patterns we see. We compare the four different domains by comparing logistic regression models that include information about language users, languages, or both. The models also allow

2 We chose these domains for investigation here because (i) they are easier to operationalise and annotate values for than some of the other domains we are looking at (e.g. stance, benefaction, private predicates), (ii) since not all SCOPIC researchers have annotated their language-specific corpora for all domains yet, these gave us the best cross-linguistic coverage.

us to assess whether or not we need to include participant variability in statistical tests of data from (corpus) typology and for which linguistic domains this is most important.

Our results make it clear that participants using the same language do not always behave similarly and this is especially true of our semantic variables. We take this to be a strong caution that the behaviour of individual participants should be considered when making typological generalisations, but also as an exciting outcome that corpus typology as a field can help us account for intra- and inter-language variation.

2 Data

2.1 The *Social Cognition Parallax Interview Corpus* (SCOPIC)

The *Family Problems* picture task is a specially-designed stimulus-elicitation methodology deployed in a language documentation context, resulting in contextualized data in conversations and narratives (San Roque et al. 2012). It is a broad-spectrum task aimed at eliciting rich and engaged data in a wide range of subdomains relevant to social cognition, including social relationships between participants (e.g. kinship), the feelings and thoughts of a range of characters in the story (as well as the task participants themselves), and other social consequences of the events depicted (e.g. benefit or detriment to different characters). In the task, two participants are presented with a series of 16 cards that they describe, then put into a sequential order. These are our stimuli. This triggers discussion about the motivations for the ordering chosen but also questions like *Who is this guy?* and commands like *Put this card here!* Then the participants tell the narrative they have jointly constructed, usually in both third-person and imagined first-person versions, to a third participant who joins the group for the second half of the task. This resets, to zero, the common ground that the first pair have built up with each other. Language users choose their own formulations for the same depicted situations, so there is no founder bias from a source language (in contrast to translation tasks).

The task has generated a multilingual corpus called the *Social Cognition Parallax Interview Corpus* (SCOPIC, Barth & Evans 2017). SCOPIC is an open-

ended, multilingual corpus annotated with functional categories relevant to social cognition, which is the semantic/functional domain for which the *Family Problems* picture task was designed to elicit rich data. Our method allows for comparability across very different languages, and it ensures that conversations (while first describing and arranging the stimuli cards) and narratives (both third person and first person) can also be compared – we will refer to these as "genres" here. Sessions typically run for around thirty minutes of speech per participant-pair.

2.2 Domains examined

In this showcase, we look at four different domains related to social cognition, two grammatical (a. and b.) and two semantic/lexical (c. and d.):

a. *Propositional framing (grammatical).* Commenting on a proposition, that is framing it within a higher (semantic predicate), best known through complement clauses in languages like English and German, but we wish to distinguish the function from the structure, hence our use of "propositional framing" rather than "complementation"; propositional framing may be accomplished by subordination or paratactic encoding.
b. *Reported speech, thought, and action constructions (grammatical).* Relaying purported quoted speech, thoughts, or actions; expressed in a direct or indirect formulation.
c. *Reported speech, thought, and action predicates (semantic).* What kinds of predicates are used to encode cognition, utterance, and quoted action; speech, thought, or emotional predicates. Predicates were coded based on specific usage, not dictionary definitions.
d. *Human reference lexical choice (semantic).* Words used to make reference to human entities; kinship formulations or other kinds of formulations.

Looking at these domains from another angle, the two semantically oriented ones focus on how people classify others (d.) and how they reveal (or not) mental inner worlds (c.), and the two grammatical ones, both relevant to the representation of speech, thought and action (the "outer/social" and "inner/psychological" aspects of social cognition), look at the grammar of how

these propositions are framed (a.) and whether the formulation of reported speech, thought, action, and emotion uses direct or indirect discourse (b.). For each of these domains, we assess how similar participants are: Do they pattern most similarly to other users of their language, as they would on a radically "language-deterministic" model, or are other factors such as individual variability or genre equally important?

2.3 Languages in the study

This study includes data from thirteen languages, twelve spoken and one signed (cf. Table 1). These languages are a subset of the total SCOPIC corpus, which has over 30 languages (Barth & Evans 2017). For each of the four domains we consider here, we restrict our study to languages which had (i) comprehensive coding for at least one respective variables, and (ii) at least three sessions of the task for which coding was available. We have restricted our analyses to languages that have three or more annotated sessions in a domain, as three seems to be the minimum number for which we can compare language user behaviour within a language. Just having three participants (within, say, a single session or two sessions) is also not enough, since language users in the same sessions may affect each other's grammatical choices, so we also look at session groups in addition to individual users as sources of variability. For various reasons we were not able to draw on the coding of all four domains across all thirteen languages in the current sample, but for four languages (Balinese, Dalabon, Kogi, and Matukar Panau) we do have data for all four domains, which we discuss further in Section 5. Table 2 provides a list of the languages included and how many participants' data were included for each domain. Some domains, such as lexical human reference have more sessions and more languages coded than other domains such as propositional framing, resulting in different numbers of participants.

language	ID	language family	geographic region	researcher(s)
Arta	ART	Austronesian	Philippines	Yukinori Kimoto
Auslan	AUS	British Sign Language	Australia	Gabrielle Hodge
Avatime	AVA	Niger-Congo	Ghana	Saskia van Putten
Balinese	BAN	Austronesian	Indonesia	I Wayan Arka, I Desak P. Eka Pratiwi
Dalabon	DAL	Gunwinyguan	Australia	Nicholas Evans
Japanese	JPN	Isolate	Japan	Heiko Narrog, Nicholas Evans, Eri Kashima, Yuki Kasuga, Carine Kawakami, Yukinori Kimoto, Norikazu Kogura, Keita Kurabe, Chikako Senge
Jinghpaw	JIN	Tibeto-Burman	China, India, Myanmar	Keita Kurabe
Kogi	KOG	Chibchan	Colombia	Dominique Knuchel, Henrik Bergqvist
Matukar Panau	MAT	Austronesian	Papua New Guinea	Danielle Barth
Murrinhpatha	MUR	Southern Daly	Australia	John Mansfield
Russian	RUS	Indo-European	Russia	Olena Tykhostup
Sanzhi Dargwa	SAN	Nakh-Daghestanian	Russia	Diana Forker
Yurakaré	YUR	Isolate	Bolivia	Sonja Gipper

Table 1 Language sample data. The ID codes are our own abbreviations for the languages, used in some of the figures in the following.

language	propositional framing	reported speech thought, and action (grammatical)	(semantic)	human reference
Arta	3	—	—	—
Auslan	—	10	10	10
Avatime	—	8	9	8
Balinese	27	20	37	38
Dalabon	4	4	4	4
Japanese	—	7	6	8
Jinghpaw	—	6	6	6
Kogi	6	4	6	6
Matukar Panau	9	16	25	21
Murrinhpatha	—	7	6	4
Russian	—	6	7	—
Sanzhi Dargwa	—	7	10	6
Yurakaré	—	13	15	15

Table 2 Languages and participant numbers included for each domain of interest. Differing numbers of participants across domains within a language reflects different amounts of annotation available and/or differences in meeting thresholds for each variable. That is, a participant had to produce multiple instances within a domain to be included in analyses (specifics follow in each subsection). Numbers here are after data exclusion, so sometimes language user numbers are lower than the number people who participated in the documented sessions.

3 Methods

Multivariate statistical modelling has a long tradition in linguistics, often being used in psycholinguistics (cf. Quené & van den Bergh 2008), sociolinguistics (cf. Tagliamonte & Baayen 2012) and corpus linguistics (cf. Bresnan et al. 2007), as these are fields where research questions often involve multiple factors (or predictors) that potentially affect an outcome. In this paper, we use one type of multivariate analysis, namely logistic regression models, to quantify patterns of linguistic use: What factors influence (or predict, in

the terminology used here) the use of one variant over another? We limit our question to alternation between variant A and variant B, and so use the technique of binary logistic regression. Any logistic regression model can include various predictors that we can test as having an effect on the outcome of A versus B, which is why it is a kind of multivariate statistics.

Mixed-effects models are a kind of generalised linear mixed-effects model (GLMM) and have so-called fixed-effect predictors as well as random-effects (thus making them "mixed"). Generally, mixed-effects models provide better estimates of fixed-effects coefficients than fixed-effects-only logistic regression (GLM) and help avoid spurious significance of fixed effects (Barth & Kapatsinski 2018; Baayen et al. 2008; Gries 2015; Winter 2020).

Fixed-effects are predictors that could apply to any sample of a population in predicting a particular outcome. We expect that the fixed-effects would show similar patterns, even if we added new data or ran our model on a different sample from the population. For example, no matter which sample of language users we include, we may expect that the specific stimuli (picture task cards) drive similar kinds of usages, albeit to different degrees. In the models, we report coefficients for each level of the fixed-effects predictor and these show how much change from the mean (perhaps best conceptualised as a baseline) is reflected by each level.

Random-effects reflect a random sample of the population we are interested in, for instance, individual language users from a population of all users of that language, or the sample of texts from all possible texts that could be produced. In the model, the random effects predictors are treated as random deflections from the population mean and are reported separately from the fixed-effects coefficients. Practically, random-effect intercepts allow us to include different baselines for each person, word, text, and so on, that we include in our data. We may expect, say, people's averages to be different from one another, but all affected in the same way by a fixed-effects variable. We then include a random effect for individual where there are intercept (baseline) levels for each person in the corpus. We then see if their baseline is changed (and how much) by each level of the fixed effect.

Language is an interesting predictor as it could be considered a fixed or random effect depending on what we want to determine. Does each language stand in for a random possible language out of the world's population of languages? Or are we interested in the effects of each language? In our case it is

the latter, so we include each language as a fixed effect, where each language is a separate level of that predictor. We test if some languages are significantly different from our mean. We choose one language for each model to form the baseline for the model (called a reference level) and our choice of language varies for each model: We select the language with the median use of the outcome variable, and this naturally varies for each domain. Using the language with the median usage of each category as the baseline makes it as hard as possible to get a significant difference between languages. This ensures that when we see significant differences, we can be confident they represent well-founded effects.

Our procedure, for each domain, is to run (i) a mixed-effects model with fixed effects for language and other predictors that relate to the context of use. The specific factors vary for each domain and are described in the sections below. The sole random effect for the first model is the stimuli card (one of 16, plus an unspecified/general option) to take into account the contribution from different pictures when we assess the effects of kinds of stimuli that have been grouped for fixed effects. We call this model *Mod-L* for 'language model'.

We then run (ii) the same model, but with two additional random effects for individual participants. This second model includes also random effects for the language user and session group. We include language users to see if language effects remain once we take into account the variability from individuals. We include groups (which corresponds to the session) to take into account the effects of the groupings of people, acknowledging that individuals may well align with their partner in the tasks in their linguistic choices. These are nested and cannot be teased apart given our corpus data. We call this model *Mod-LI* for 'language plus individual model'.

Finally, we run (iii) a model with all three random effects, but exclude language as fixed-effect to assess whether including language makes a significant contribution beyond what the individual users of a language contribute. We call this model *Mod-I* for 'individual model'.

We used R (R Core Team 2021) with the packages *lme4* (Bates et al. 2015) and *lmerTest* (Kuznetsova et al. 2017) to produce and assess models in terms of their Akaike information criterion (AIC, Akaike 1973). The AIC can be used to assess how well a model is predicting the outcome in terms of true positives and true negatives. We report the AICc (Sugiura 1978), which is a

corrected AIC to take into account small sample sizes and model complexity. More complex models are punished so that models cannot achieve lower AIC values (which are better than higher ones) simply by adding in extra factors; rather, those extra factors must make a substantial difference to be worth including. AICc values can be compared resulting in delta Δ, or the difference between the models' AICc. Models with a delta less than four are not substantially different, but those with delta greater than four have substantially less empirical support (Burnham & Anderson 2002). Further, we report Akaike weights w which show the probability that the best model is the most predictive one (Wagenmakers & Farrell 2004). We use the *qpcR* (Spiess 2018) R package to obtain the AICc, Δ, and w values.

Our model comparison allows us to test whether or not it is necessary to include participant information in our (and other) typological models. It also helps us to assess whether or not language effects are genuine or not once we take into account the variable contributions from individuals. Finally, the process allows us to ask whether language is still a useful grouping factor once individual variation is taken into account, a question that is important for typology and possible to test in corpus typology.

4 Results

We apply our model testing approach to four different functional domains, which we discuss here in the order (i) propositional framing (a grammatical feature, Section 4.1), (ii) quotation strategies (a grammatical feature, Section 4.2), (iii) predicates for representing reported quotations (a lexical/semantic feature, Section 4.3), and (iv) human reference (a lexical/semantic feature, Section 4.4).

4.1 Propositional framing (grammatical)

Propositional framing is about which constructions are possible for commenting on propositional content, by embedding one proposition as the argument of another, at some level of semantic representation. The classic survey by Noonan (1985) equivocates in its definition: The chapter starts with a syntactic definition – "by complementation we mean the syntactic situation that arises when a notional sentence of predication is an argument of a predic-

ate" (p. 42), but later (p. 64) states that "we have defined complementation as the grammatical state where a predication functions as an argument of a predicate", then immediately referring to this as "this (universal) semantic characterization". These ambiguities are partly due to the ambiguities of "argument" and "predicate" in linguistic metalanguage, as both syntactic and semantic terms, but because we are interested in cross-linguistic grammatical alternatives for realising these, we adopt the term "propositional framing" as the functional term and reserve "complementation" for one of the structural alternatives for expressing it – along with such alternatives as particles (like the Dalabon particle *yangdjehneng* '(someone unspecified) believed that', Barth & Evans 2021: Section 2.4), or, in some cases, unframed direct quotation which context shows to be someone's reported thought.

Complement clauses, then, are one well-known strategy for packaging a proposition into a higher-order predicate, which may be a locution *(I said* [*that I would do it*]), perception (*I saw* [*that he left*]) as in (1), desire (*I want* [*to leave now*]), knowledge (*I know* [*that is not true*]), or evaluative commentary (*It is good* [*that you will see her*]) as in (2). There are other subordinating strategies for the same purpose such as clause chains, converbs or participial constructions, among others (see Cristofaro 2003 for a typology, including a discussion of the gradient nature of "subordination" as a concept).[3]

(1) Balinese

tepuk-in	*be ee*	[*kene*	*deng deng*	*ngorte*	*raga*	*jak*	*kuren*	*raga*]
seen-APPL1	PRF	like.this	while	talk	I	with	husband	my

'I was seen like this, while I was having a conversation with my husband.'
(speaker Made Ratnadi, SocCog-ban02-badung2-task_4 3:25–3:30)

3 Morphological glossing follows the Leipzig Glossing Rules. Additional abbreviations: ADN – adnominal; BEN – benefactive; C – class; EMPH – emphatic; GERUND – gerund; MID – middle voice; MSG – masculine singular; PCUS – customary plural; POSR – possessor; RR – reflexive/reciprocal; SPK – speaker anchored.

(2) Matukar Panau

nen *aipain-da* *di-mado-ndo* [*hum-e* *main*] *uyan* *ti*
mother daughter-COM 3PL-sit-D:R hit.PL-R.I.PFV COMP good NEG

'The mother is sitting with her daughter and his hitting them is not good'
(speaker Mingkui Agid, SocCog-mjk10-ckd_ma_2 4:39–4:41)

Another strategy realising the same functions is to use paratactic structures. For instance, juxtaposition can be a syntactic strategy for linking propositional content and a higher-level layer of conceptualization as in (3).

(3) Matukar Panau

y-a-we-nga *main* *ha-n* *yawa-n* *i* *tuli-nggo*
3SG.S-go-R.I.PFV-FOC TOP CLF-3SG spouse-3SG 3 tell-R.I.IPFV

[*"manag manag matan ta ti pan-au-go,* *main* *tagtag?"*]
like.that like.that money INDF NEG give-1SG.R-R.I.IPFV PROX what

'She is telling her husband, "Why do you not give me money, why is that?"'
(speaker John Bogg, SocCog-mjk02-tk_jb_1 5:51–5:56)

As a sub-type of paratactic structures, some languages have independent strategies for conveying speech or thought content. The content proposition is simply presented without any framing verb or other element, and instead shifts in tense, deixis, person, and perhaps prosody show that the proposition is at a different discourse level and part of a higher-order category as in (4).

(4) Direct speech with no preceding or following framing predicate:

Matukar Panau

C: *haun-da* *tamat* *aim* *di-da* *di-tor-dop* *nub* *lumi-k*
again-COM man boy 3PL-COM 3PL-go.around-D.IRR beer drink-NMLZ

wai *di-bul-dop*
like 3PL-try-D.IRR

'The man and the boy walk around again, they [friends] are about to drink beer...'

M: [*"nga-mai-ye"*]
1SG.S-NEG-R.I.PFV
"'I don't want [to]!" (as main male character)'
(speakers Clara Kusos Darr and Mingkui Agid, SocCog-mjk10-ckd_ma_2 3:50–3:59)

Our data are coded for subordinating versus paratactic framing as well several broad function types, based on Noonan (1985), including reporting utterance, thought, probability assessment, knowledge, perception, evaluative commentary, desire, and fear. Some languages in our sample have a predominant strategy for the expression of these propositions and frames, like Matukar Panau which has very little subordination for any of these kinds of clause combination. Other languages in our sample have variable strategies that are contingent on specific functions. For instance, Balinese uses primarily paratactic structures for evaluative commentary as in (5). Complement clauses in Balinese are mostly limited to reporting utterances, thoughts, perception as in (1), knowledge, and probability assessment.

(5) Balinese

dadinne *tiang me..* *ma-kaeng-an*
so I MID-regret-NMLZ

[*inget* *teken* *panak* *inget* *teken* *kurenan*]
remember with son remember with wife

'So I regretted (it), (when) I remembered my son and my wife.'
(speaker Desak Putu Eka Pratiwi, SocCog-ban01-badung1-task_4 4:42–4:49)

Five languages in our corpus have three or more sessions with annotations for propositional framing structures, which we group as either subordinate or paratactic. We included 49 speakers from these languages who produced five or more propositional framing structures. Table 3 gives summary statistics for these data, including the overall language mean for each propositional framing strategy, as well as the minimum and maximum means from language users for each strategy. Arta, for instance, shows a very strong tend-

language	sessions ▼	users	tokens	propositional framing structure	language mean	min. user mean	max. user mean
Arta	3	3	14	subordinate	7.65%	4.59%	19.05%
			169	paratactic	92.35%	80.95%	95.41%
Balinese	11	27	181	subordinate	62.63%	0.00%	100.00%
			108	paratactic	37.37%	0.00%	100.00%
Dalabon	3	4	4	subordinate	3.48%	0.00%	7.27%
			111	paratactic	96.52%	92.73%	100.00%
Kogi	3	6	126	subordinate	89.36%	82.76%	100.00%
			15	paratactic	10.64%	0.00%	17.24%
Matukar Panau	3	9	48	subordinate	13.68%	0.00%	83.33%
			303	paratactic	86.32%	16.67%	100.00%

Table 3 Distribution of subordinate and paratactic framing structures across languages and language users. Shaded cells indicate over 60% usage for a structure in the language, or minimum and maximum user means.

ency towards paratactic framing: The mean for the language is 92.35%, the minimum mean from any one Arta speaker is 80.95% (meaning each speaker uses mostly paratactic structures), and the maximum mean from any one Arta speaker is 95.41%, which is high, but also indicates that all speakers use at least some subordination (minimum is 4.59%) for propositional framing. This strong language-plus-user pattern is reflected by the shading applied to the last three columns of the table for Arta's paratactic frequencies (also the case for Dalabon). Kogi shows the inverse pattern for the subordinate strategy, which is the most frequent strategy for the language and for every speaker of the language in our sample. Balinese and Matukar Panau have an overall higher tendency towards using paratactic structures, but for each language, there are speakers who use predominantly, or in the case of Balinese categorically, either subordinating or paratactic framing. In our table, this is visually reflected by shaded cells in the language user maximums columns for both of framing structures.

We model the likelihood of using subordinate versus paratactic framing structures using multivariate logistic regression models, as explained in Sec-

tion 3. Our fixed effects in each model are (i) language, (ii) genre, (iii) function of the propositional framing, and (iv) an interaction effect between genre and function. This interaction assesses whether or not the function affects the structural outcome differently in each genre. Because of data sparsity, we collapse the functional category to either speech and utterance framing or other.[4] We ran three models, *Mod-L* which lacks a random effect for individuals, *Mod-I* which lacks a fixed effect for language, and *Mod-LI* which includes a fixed effect for language and the nested random effects for individuals. The results are reported in Table 4, with significant effects highlighted in colour.

Our results in Table 4 show first that the model with both individuals and languages is the most likely to be the best one (for *Mod-LI*, $w \approx 1$) and that the AICc Δ between it and the other models are high, especially between it and *Mod-L* ($\Delta = 87.92$). We take this as support that including individuals and languages make for the best possible model.

Looking now at language as a fixed effect, we see that that Kogi and Balinese have significantly more subordinated constructions than the baseline level of the median language Matukar Panau in all both models where language is a predictor. In *Mod-L*, Dalabon shows significantly more paratactic usage than the baseline level, but this effect is not significant once variability from participants and session groups are taken into account. This is a strong signal that it is important to include individual variability, even in assessing the typology of grammatical structures.

For all three models genre is significant, with narrative genre resulting in more paratactic constructions. Function is also significant across all three models with the quotation functions (here speech and thought) resulting in more paratactic constructions. Further, these two predictors interact[5] in a subtractive way: Although there are more paratactic constructions with quotation functions, those functions are expressed with more subordinate constructions within narrative genres (cf. Table 5). This points to the need for

4 There are not enough tokens of each of our other functional categories to assess significant differences when taking our other variables into account. Exploratory analysis not reported here does not show strong differences between these categories. Other work (Kimoto et al., in preparation) will focus more on this area.

5 Interactions between function × language and genre × language were tested, but resulted in model non-convergence or extremely high standard error.

	Mod-L: language only		*Mod-LI:* both		*Mod-I:* individual only	
	$\Delta = 87.92$, $w \approx 0$		AICc = 706.21, $w \approx 1$		$\Delta = 16.88$, $w \approx 0$	
	$\beta \pm$ SE	p	$\beta \pm$ SE	p	$\beta \pm$ SE	p
(intercept)	−0.05 ± 0.27	0.85	−0.39 ± 0.78	0.62	−2.08 ± 0.58	< 0.01
Kogi	−4.48 ± 0.35	< 0.01	−4.76 ± 1.07	< 0.01		
Balinese	−2.46 ± 0.24	< 0.01	−2.57 ± 0.84	< 0.01		
Arta	0.32 ± 0.35	0.36	0.03 ± 1.08	0.98		
Dalabon	1.21 ± 0.56	0.03	1.87 ± 1.33	0.16		
narrative genre	1.35 ± 0.33	< 0.01	1.53 ± 0.40	< 0.01	1.51 ± 0.41	< 0.01
quotation function	2.77 ± 0.30	< 0.01	3.50 ± 0.38	< 0.01	3.60 ± 0.38	< 0.01
narrative genre × quotation function	−1.58 ± 0.40	< 0.01	−2.22 ± 0.47	< 0.01	−2.32 ± 0.48	< 0.01

Table 4 Model comparison for subordinate vs. paratactic constructions for propositional framing. The models include 1079 observations. Positive coefficients (β) are associated with more paratactic constructions for propositional framing. Significant results are shaded. The reference levels are Matukar Panau for language, descriptive for genre, and non-quotation (which includes commentative, desiderative, fear, perception, knowledge, and probability functions) for function. For *Mod-L*, the variance for stimuli cards ($n = 17$) is $s^2 = 0.06$ (standard deviation SD = 0.19). For *Mod-LI*, the variance for stimuli cards is $s^2 = 0.04$ (SD = 0.20), for language users ($n = 49$) $s^2 = 0.20$ (SD = 0.45), and for session groups ($n = 23$) $s^2 = 1.39$ (SD = 1.18). For *Mod-I*, the variance for stimuli cards is $s^2 = 0.06$ (SD = 0.24), for language users $s^2 = 0.26$ (SD = 0.51), and for session groups $s^2 = 5.09$ (SD = 2.26). Abbreviations: AIC – Akaike information criterion; AIC – corrected Akaike information criterion; β – regression coefficient; Δ – difference; SE – standard error; p – probability; w – Akaike weight. Note that the value of w is never 0 (or 1); we use the approximation symbol (≈) to indicate values very close to either extreme.

genre	function	subordinated	paratactic
descriptive	quotation	86	443
	other	106	38
narrative	quotation	134	173
	other	60	65

Table 5 Cross-tabulation of genre and function for propositional framing constructions.

further analysis of construction types and functions and is part of ongoing work (cf. Kimoto et al., in preparation).

Propositional framing is a domain with some language user effects. Generally, however, language, function, and context effects come out strong whether or not language user variability is taken into account, indicating fairly robust effect from language. However, *Mod-I* is still better than *Mod-L* in terms of the AICc, showing that individual behaviour is important to consider. Figure 2 shows the estimated means of language for our best model with confidence intervals, produced with the *ungeviz* package (Wilke 2021) and obtained from the *emmeans* package (Lenth 2021). These show the contribution of language to construction choice in our multivariate model. The estimate (β) mid-point of 0 (in both the model and the corresponding figure) indicates that the estimate for the language is very close the mean (average) across all data. The further away from 0 it falls, the greater the association with either subordination (negative values) or paratactic clause combinations (positive values) is.

4.2 Reported speech, thought, and action constructions (grammatical)

The second grammatical variable we examine is how the language users in our corpus represent content that is purported to be either spoken, thought, or enacted by the characters in the stimuli. Reported actions are those that a participant indicates were enacted by a quotee. Reported actions were largely, but not exclusively, limited to Auslan. Quoted action is an enactment

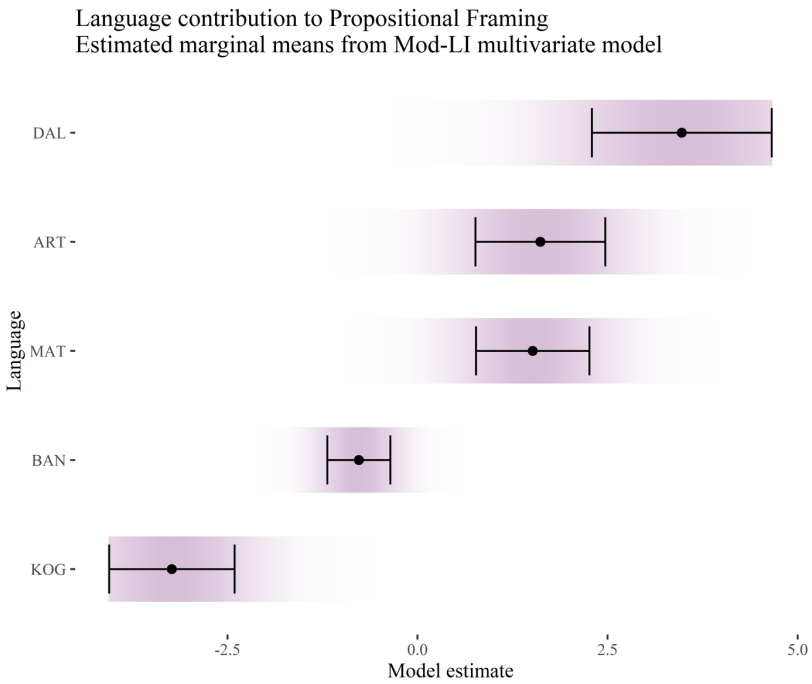

Figure 2 Model estimates with confidence intervals for the contribution of language to propositional framing construction choice. This and similar figures in the following were plotted using the R packages *ggplot2* (Wickham 2016) and *ggthemes* (Arnold 2021).

of an action or behaviour done by some referent in the narrative. Diagnostics include a construction of a surrogate blend enacting the action or behaviour.

The two primary means of quotation, and the ones we examine here, are direct and indirect constructions (but see Evans 2013 for a more elaborated typology). Take, for example, quoted speech: In canonical direct speech constructions, the deictic values of the purported speech act are reproduced in the quoted portion, resulting in an apparent shift within the sentence

between the time of reporting (time of expression) and the time of the purported speech act event, as well as between the person values, honorifics, or location-deictic expressions. In canonical indirect speech constructions the deictic values of the purported speech act (tense, person, honorifics, etc.) are calculated with respect to the primary speech event. There is less enactment of the quotee, with the main focus being on reporting quoted content. The difference between direct and indirect constructions is best understood through examples, which we give below. Thoughts, like speech, can also be reported through direct and indirect constructions, but actions are always part of direct constructions.

Example (6) from Matukar Panau shows a direct thought quotation (highlighted in colour), framed by the complex predicate (*ilon haiyando* 'worry'). The framing predicate is part of a clause chain that is in a perfective-realis aspect-mood and is in the third person. The reported thought, however, is in the imperfective-realis aspect-mood and the person is first person plural exclusive. These shifts are clear diagnostics of a direct construction in the language.

(6) Matukar Panau

ha-n	*pain*	*ilo-n*	*haiyan-do*	***"tag***	***tag***	***ngam-bul-ago?"***	
CLF-3S	woman	inside-3S	bad-D.R	**what**	**what**	**1P.EXCL-do-R.I.IPFV**	
di-bal-do	*di-kabiyai-e*						
3P-throw-D.R	3P-discuss-R.I.PFV						

'His wife worries, "What will we do?", they talk and meet.'

(speaker Kadagoi Rawad Forepiso, SocCog-mjk01-krf_spw_3 7:13–7:18)

Example (7) from Avatime shows an indirect speech quotation (in highlighted colour), framed by a quotative predicate (*εεpani* 'talking'). In the quoted speech, we see third person (noun class 1) singular pronoun *yε* referring to the women, which would be in first person if the quotation was direct. These matches between the time/person of expression and of quoted material are clear diagnostics of an indirect construction in the language, as is the Avatime-specific use of *kílε gì* to introduce the quoted material.

(7) Avatime

lɛ	*one*	*ɛɛpani*	*kílɛ*	*gì*
and	C1.SG-mother=DEF	C1.SG.PROG-talk	how	REL

ɔkaɛ	***asa***	***yɛ***	***kikpafu***
C1.SG-father=DEF	**C1.SG.PFV-hit**	**C1.SG**	**C4.SG-fist**

'And the woman is talking about how the man hit her.'

(SocCog-avn01-MM_AIA_all 32:32–32:37)

A view found in the literature (e.g. Coulmas 1986) is that direct speech simply reproduces what was said in the reported speech event, but there is an extra imposition of the quoter's perspective in indirect speech.[6] For instance, in German there is a choice between using the indicative or the subjunctive in indirect quoted speech, depending on how far the primary speaker endorses the truthfulness of the report offered by the quotee. In this view, indirect speech reveals more of the quoter and direct speech hides their perspective. However, the direct speech allows for the possibility of dramatising the represented speech through enactment of prosody or bodily gestures, and since these elements are chosen by the primary speaker, they may be another means of injecting the primary speaker's take on how a quote was expressed, so that direct constructions can give a different kind of window into the quoter's assessment of the quotee, adding an additional layer social cognition, beyond simply revealing the quoted content (cf. Clark 1996; Tannen 1989).

Twelve languages in our corpus have three or more sessions with annotations of reported speech, thought, and action structures. We included the 108 language users who produced six or more reported speech or thought constructions, either direct or indirect. Table 6 gives summary statistics of

6 Consider, for example, Coulmas (1986: 2): "The fundamental difference between the two lies in the speaker perspective or point of view of the reporter: In direct speech the reporter lends his voice to the original speaker and says (or writes) what he said, thus adopting his point of view, as it were [...] In indirect speech, on the other hand, the reporter comes to the fore. He relates a speech event as he would relate any other event: [F]rom his own point of view."

language	sessions ▼	users	tokens	reported speech structure	language mean	min. user mean	max. user mean
Auslan	5	10	197	direct	93.69%	68.75%	97.50%
			39	indirect	6.31%	2.50%	31.25%
Avatime	4	8	2	direct	3.48%	0.00%	20.00%
			73	indirect	96.52%	80.00%	100.00%
Balinese	10	20	171	direct	54.49%	0.00%	100.00%
			125	indirect	45.51%	0.00%	100.00%
Dalabon	3	4	114	direct	97.58%	92.31%	100.00%
			3	indirect	2.42%	0.00%	7.69%
Japanese	3	7	262	direct	93.49%	87.50%	100.00%
			16	indirect	6.51%	0.00%	12.50%
Jinghpaw	3	6	59	direct	82.72%	66.67%	88.89%
			12	indirect	17.28%	11.11%	33.33%
Kogi	3	4	76	direct	78.57%	56.25%	88.64%
			21	indirect	21.43%	11.36%	43.75%
Matukar Panau	8	16	373	direct	98.76%	84.62%	100.00%
			4	indirect	1.24%	0.00%	15.38%
Murrinhpatha	4	7	165	direct	99.40%	96.77%	100.00%
			1	indirect	0.60%	0.00%	3.23%
Russian	3	6	117	direct	34.15%	7.50%	46.00%
			241	indirect	65.85%	54.00%	92.50%
Sanzhi Dargwa	4	7	215	direct	81.72%	50.00%	90.91%
			45	indirect	18.28%	9.09%	50.00%
Yurakaré	7	13	375	direct	75.45%	27.27%	97.62%
			123	indirect	24.55%	2.38%	72.73%

Table 6 Distribution of direct or indirect constructions across languages and language users. Shaded cells indicate over 60% usage for a structure in the language, or minimum and maximum user means.

Figure 3 The stimulus card 'Thinking about gaol', which includes thought bubbles.

these data, showing that some languages (Auslan, Dalabon, Japanese, Jinghpaw, Matukar Panau, Murrinhpatha, Sanzhi Dargwa) have almost categorical usage of direct structures which is reflected by the shading in the table spanning across the language mean and minimum and maximum language user means. Avatime has near categorical indirect construction usage. For these languages, all language users in our sample follow the overall pattern of the language. For the languages Balinese and Yurakaré, there are participants whose maximum means are (close to) 100%, indicating that some users have (near) categorical usage of one or another construction type, but this could be in either direction. Other languages have strong tendencies towards one construction type, but a fair amount of deviation from the overall language pattern among individual language participant.

During the majority of each session, language users are looking at picture stimuli cards while they formulate expressions. Two of our picture cards have thought bubbles (cf. Figure 3) and five have speech bubbles (cf. Figure 4);

Figure 4 The stimulus card 'Receiving clothes', which includes a speech bubble.

the rest have neither (cf. Figure 5). As this is an element that may affect expression of quotation, we include this categorical distinction in our analyses (although cf. San Roque et al. 2012 for a discussion of languages that lack familiarity with this convention).

Our quantitative results below investigate the likelihood of languages being more or less associated with either a direct or indirect strategy for quotation. In each model, we explore differences between (i) languages, (ii) genre, and (iii) whether or not the picture card stimuli discussed at the time of expression includes a thought bubble, speech bubble, or none at all (these are our fixed effects). As discussed in Section 3, we modelled this likelihood in three ways: *Mod-L* lacks a random effect for individuals, *Mod-I* lacks a fixed effect for language, and *Mod-LI* includes a fixed effect for language and the nested random effects for individuals. The results are reported in Table 7, with significant effects highlighted.

Our results in Table 7 show first that the model with both individuals and

Figure 5 The stimulus card 'Refusing drink', which features neither thought nor speech bubbles.

languages is the most likely to be the best one (for *Mod-LI*, $w \approx 1$) and that the AICc Δ between it and the other models are very high, especially between it and the fixed-effects-only model ($\Delta = 163.3$). We take this as strong support that including individuals and languages make for the best model in our set.

We see that for our first fixed-effect, language, there is consistency across the two models that include it as a predictor: Murrinhpatha Matukar Panau, Dalabon, Japanese, and Auslan have significantly more direct constructions than middle-of-the-road Kogi, while Balinese, Russian, and Avatime have significantly more indirect constructions. For the second fixed effect of genre, we see a significant effect across all three genres, where narratives are more likely to have direct constructions. Our third fixed effect is not significant across any of the models.

Our main take-away from this case study is that language effects come out strong whether or not language user variability is taken into account,

	Mod-L: language only		*Mod-LI:* both		*Mod-I:* individual only	
	Δ = 163.30, $w \approx 0$		AICc = 2292.97, $w \approx 1$		Δ = 73.17, $w \approx 0$	
	$\beta \pm$ SE	*p*	$\beta \pm$ SE	*p*	$\beta \pm$ SE	*p*
(intercept)	−1.44 ± 0.30	< 0.01	−1.32 ± 0.68	0.05	−1.87 ± 0.42	< 0.01
Murrinhpatha	−3.92 ± 1.03	< 0.01	−4.50 ± 1.34	< 0.01		
Matukar Panau	−3.09 ± 0.52	< 0.01	−3.52 ± 0.89	< 0.01		
Dalabon	−2.50 ± 0.64	< 0.01	−3.07 ± 1.15	0.01		
Japanese	−1.55 ± 0.35	< 0.01	−2.18 ± 0.91	0.02		
Auslan	−1.41 ± 0.30	< 0.01	−1.63 ± 0.79	0.04		
Yurakaré	0.04 ± 0.27	0.88	−0.35 ± 0.75	0.65		
Jinghpaw	−0.09 ± 0.39	0.81	−0.34 ± 0.89	0.70		
Sanzhi Dargwa	−0.14 ± 0.30	0.64	−0.16 ± 0.83	0.85		
Balinese	1.16 ± 0.28	< 0.01	1.23 ± 0.73	0.09		
Russian	1.93 ± 0.28	< 0.01	2.12 ± 0.85	0.01		
Avatime	4.92 ± 0.57	< 0.01	5.21 ± 0.98	< 0.01		
narrative genre	−0.36 ± 0.11	< 0.01	−0.32 ± 0.13	0.02	−0.31 ± 0.13	0.02
card type: speech bubble	0.48 ± 0.25	0.05	0.49 ± 0.26	0.06	0.49 ± 0.25	0.05
card type: thought bubble	0.32 ± 0.34	0.34	0.30 ± 0.35	0.38	0.32 ± 0.35	0.36

Table 7 Model comparison for direct vs. indirect reported speech, thought, and action constructions. The models include 3367 observations. Positive coefficients (β) are associated with more indirect speech constructions. Significant results are shaded. The reference levels are Kogi for language, descriptive for genre, no bubble for stimulus card. For *Mod-L*, the variance for stimuli cards ($n = 17$) is $s^2 = 0.14$ (SD = 0.38). For *Mod-LI*, the variance for stimuli cards is $s^2 = 0.14$ (SD = 0.38), for language users ($n = 108$) $s^2 = 0.79$ (SD = 0.89), and for session groups ($n = 57$) $s^2 = 0.42$ (SD = 0.65). For *Mod-I*, the variance for stimuli cards is $s^2 = 0.14$ (SD = 0.38), for language users $s^2 = 0.78$ (SD = 0.89), and for session groups $s^2 = 7.39$ (SD = 2.72).

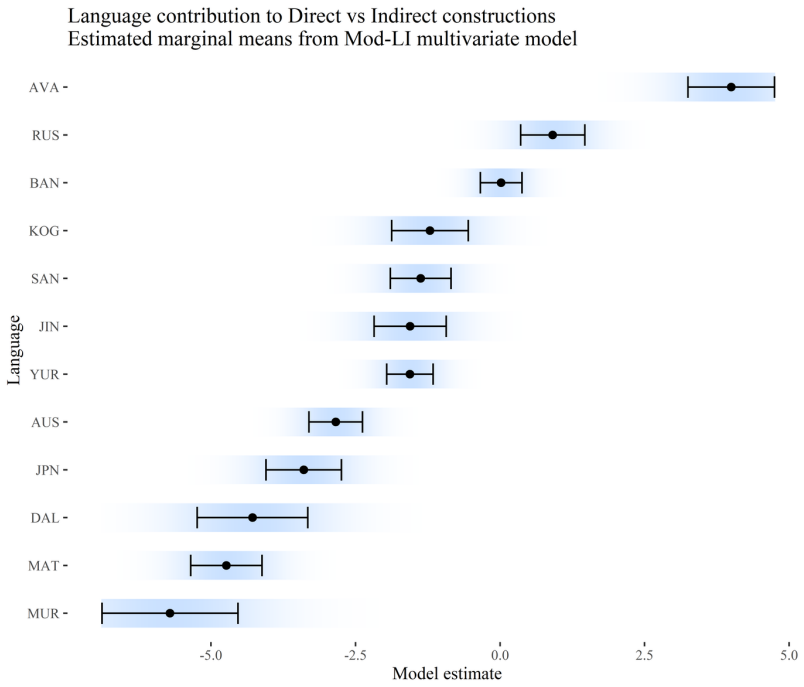

Figure 6 Model estimates with confidence intervals for the contribution of language to direct or indirect construction choice.

indicating fairly robust patterns within several individual languages. Figure 6 shows the estimated means of language for our best model with confidence intervals to assess the contribution of language to construction choice in our multivariate model.

4.3 Reported speech, thought, and action predicates (lexical/semantic)

We now turn to the lexical/semantic domain: The specific lexical items used to frame reported speech, thoughts, and quoted actions.[7] When language users express that characters are speaking, arguing, thinking, or reflecting, they build and convey the perspectives of the characters in their stories. In so doing, they may focus either on outer or on inner worlds. The outer world is the one of speaking, yelling, signalling, or actions that focuses on how characters overtly interact. The inner world is the one of thinking, imagining, feeling. or dreaming – here, language users focus on expressing information about an individual and what can only be known by that individual. All the languages in our sample have at least some distinct linguistic devices (predicates) for both of these functions, but in this case study we assess whether some users, or some languages, strongly favour one or the other strategy: Do language users avoid revealing inner worlds and focus on the visible social interaction of their characters or do they turn inward and bring to light private and hidden thoughts? In this case study, we include predicates that frame direct or indirect quoted material as well as predicates that simply express that speaking or thinking took place. As a secondary question, in our models below we can assess whether framing is associated to a higher degree with either quotation or cognition and emotion, or neither.

In the examples below we see a cognitive predicate from Russian that frames a thought (8), as well as two cognitive predicates without quotations from Kogi (9). Example (10) is from Dalabon, with a quotative predicate framing speech, and (11) is from Murrinhpatha, with a stand-alone quotative predicate. Also refer back to (6) for an emotional, inner-world predicate from Matukar Panau and (7) for a quotative predicate from Avatime.

7 Quoted action is an enactment of an action or behaviour done by some referent in the narrative. Diagnostics include a construction of a surrogate blend enacting the action or behaviour. Reported actions were largely, but not exclusively, limited to Auslan. For more see Ferrara & Johnston (2014) and Hodge & Johnston (2014).

(8) Cognitive predicate framing quoted content:

Russian

dumajet,	*"tak*	*i*	*zna-l,*	*što*	*èto*	*byl*
think-3SG.PRS	thus	and	know-PST.MSG	that	this	be-PST.MSG

nesoglasovannyj	*miting"*
unauthorised	rally

'He thinks, "I knew, it was actually an unauthorised rally!"'

(speaker Natalia, SocCog-rus01-nm_og_1 6:28–6:32)

(9) Stand-alone cognitive predicate:

Kogi

hẽ	*aiza*	***hangwá,***	*hieka*	*akbẽyatukka,*	*aiza*
DEM.ADN.SPK	only	**think**	something	3SG.IO-say-PROG-PRS	only

hangwá
think

'He is just thinking, he is telling him something, he is just thinking.'

(speaker MM, SocCog-kog01-LCZ_170803_1 1:50–1:53)

(10) Quotative predicate framing quoted content:

Dalabon

kirdikird	*ka-h-marnû-yi-ninjyi,*	*kirdikird-ngan*
woman	3SG.A>1SG.O-R-BEN-say-PCUS	wife-1SG.POSR

"Manjhkerninjh	*mah*	*da-h-kolhngu-n*	*wah?"*
why	EMPH	2SG.A>3SG.O-R-drink-PRS	grog

'The woman used to say to me, my wife: "Why are you drinking grog?"'

(speaker Manuel Pamkal,
SocCog-dal03-20200117ManuelFamilyProblems_3 1:38–1:45)

(11) Stand-alone quotative predicate:

Murrinhpatha

dem-nintha-ngkabirr-kanam

pierce.RR.3SG.NFUT-DU.M-scold-BE.IPFV

'Two men are arguing with each other.'

(speaker GeMa, SocCog-mwf01-GM_EB_01_2018-10-31 3:35–3:36)

Twelve languages in our corpus have three or more sessions with annotations of predicative descriptions of quotation and cognition. We included the 141 language users who produced seven or more predicates, either quotative or non-quotative, the latter of which includes cognitive and emotional predicates. These predicates may optionally frame quoted content.[8]

Table 8 presents summary statistics for the average number of cognitive and quotative predicates for each language, as well as the range of averages of each language user by language. Unlike with the grammatical linguistic domains in Sections 4.1 and 4.2, there are very few instances of strong language patterning. We see no shading going across the language mean and minimum and maximum language user means. We also see few overall language preferences: Dalabon and Sanzhi Dargwa have more quotative predication, but also some speakers that use more cognitive predicates. Russian and Auslan have overall more non-quotative predicates, but with some users who use both more or less equally. For Matukar Panau, there is one speaker that uses only non-quotative predicates, but for the other languages, all users vary in their preferred strategy for this task.

Our quantitative results below investigate the likelihood of languages being more or less associated with either cognitive (inner world) or quotative (outer world) predicates. In each model our fixed effects are (i) language, (ii) genre (descriptive or narrative), (iii) whether the predicate frames a quotation or not, and (iv) whether or not the picture card stimuli discussed at the time of expression includes a thought bubble, speech bubble, or neither. As above

8 We excluded some less frequent quotation-framing predicates such as perceptual predicates (*The man saw him, "What are you doing?!"*). We also excluded nominal (*The man "Phew!"*) and simulative (*The man was like "No I didn't!"*) framing strategies that did not indicate whether or not the quotation was reflecting inner- or outer-world focus.

language	sessions ▼	users	tokens	predicate type	language mean	min. user mean	max. user mean
Auslan	5	10	190	quotative	35.85%	21.15%	48.53%
			340	non-quotative	64.15%	51.47%	78.85%
Avatime	4	9	127	quotative	48.47%	27.78%	68.42%
			135	non-quotative	51.53%	31.58%	72.22%
Balinese	13	37	649	quotative	57.95%	31.82%	91.11%
			471	non-quotative	42.05%	8.89%	68.18%
Dalabon	3	4	93	quotative	80.17%	35.71%	100.00%
			23	non-quotative	19.83%	0.00%	64.29%
Japanese	3	6	104	quotative	45.81%	23.53%	71.43%
			123	non-quotative	54.19%	28.57%	76.47%
Jinghpaw	3	6	116	quotative	45.49%	30.56%	54.90%
			139	non-quotative	54.51%	45.10%	69.44%
Kogi	3	6	140	quotative	46.67%	27.78%	57.14%
			160	non-quotative	53.33%	42.86%	72.22%
Matukar Panau	10	25	543	quotative	55.24%	0.00%	75.00%
			440	non-quotative	44.76%	25.00%	100.00%
Murrinhpatha	3	6	149	quotative	57.09%	36.84%	74.07%
			112	non-quotative	42.91%	25.93%	63.16%
Russian	3	7	183	quotative	29.61%	18.42%	40.15%
			435	non-quotative	70.39%	59.85%	81.58%
Sanzhi Dargwa	4	10	412	quotative	60.32%	39.53%	100.00%
			271	non-quotative	39.68%	27.27%	60.47%
Yurakaré	7	15	871	quotative	45.18%	13.04%	68.04%
			1057	non-quotative	54.82%	31.96%	86.96%

Table 8 Distribution of predicate semantics in the domain of reported speech, thought, and action constructions across languages and language users. Shaded cells indicate over 60% usage for a structure in the language, or minimum and maximum user means.

we ran three models, *Mod-L* which lacks a random effect for individuals, *Mod-I* which lacks a fixed effect for language, and *Mod-LI* which includes a fixed effect for language and the nested random effects for individuals. The results are reported in Table 9, with significant effects in highlighted in colour.

Our results in Table 9 show that the model with both individuals and languages is the most likely to be the best one (for *Mod-LI*, $w \approx 1$) and that the AICc Δ between it and the other models are high, especially between it and *Mod-L* ($\Delta = 192.42$), where the difference is quite large. We take this as support that including individuals and languages makes for the best possible model, especially in the case of a semantic variable.

Regarding our first fixed effect, most languages are not significantly different from our median/baseline language of Avatime in any of the models. Only Dalabon and Balinese have more outer-world (quotative) predicate usage, and only Russian has consistently more inner-world (cognitive and emotional) predicate usage. Note, though, that if language user variation was not taken into account with a mixed-effects model, additional significant language differences would appear to be present, including with Auslan, Japanese, Matukar Panau, and Sanzhi Dargwa.

For the second fixed effect of genre, we see it has no significant effect across any of the models. This indicates that both inner- and outer-world predicates are occurring in both the descriptive and narrative portions of the task. Whether the predicate is framing or not is significant in all models. A follow-up analysis of this finds that emotional predicates are overwhelmingly non-framing. Most languages have very few, if any, emotional predicates that frame thoughts. On the upper end, Japanese, Matukar Panau, and Russian have between 10–17% of their emotional predicates framing thought content.

The effects of the picture card stimuli are strong across all models, with significantly more inner-world predicates being used while discussing cards with thought bubbles and significantly more outer-world predicates being used while discussing cards with speech bubbles. This shows us that the stimuli contexts are important for this semantic variable.

Our main take-away from this study is that there are fewer language-specific patterns for this semantic variable and language users are much more varied. For typological studies of semantic variables, it is essential to include language user variability in our models. The language patterning that remains is interesting and will be explored in more depth in future

	Mod-L: language only		*Mod-LI:* both		*Mod-I:* individual only	
	Δ = 192.42, $w \approx 0$		AICc = 7599.10, $w \approx 1$		Δ = 21.86, $w \approx 0$	
	$\beta \pm$ SE	p	$\beta \pm$ SE	p	$\beta \pm$ SE	p
(intercept)	0.84 ± 0.28	< 0.01	0.89 ± 0.40	0.03	0.77 ± 0.28	< 0.01
Dalabon	−2.17 ± 0.32	< 0.01	−2.01 ± 0.56	< 0.01		
Balinese	−0.63 ± 0.17	< 0.01	−0.72 ± 0.35	0.04		
Sanzhi Dargwa	−0.40 ± 0.18	0.02	−0.42 ± 0.41	0.31		
Murrinhpatha	−0.41 ± 0.21	0.05	−0.39 ± 0.46	0.39		
Matukar Panau	−0.44 ± 0.17	< 0.01	−0.27 ± 0.36	0.45		
Jinghpaw	0.18 ± 0.21	0.40	0.26 ± 0.46	0.58		
Kogi	0.30 ± 0.20	0.14	0.31 ± 0.46	0.50		
Japanese	0.74 ± 0.21	< 0.01	0.33 ± 0.48	0.49		
Yurakaré	0.25 ± 0.16	0.12	0.36 ± 0.37	0.33		
Auslan	0.45 ± 0.18	0.01	0.48 ± 0.40	0.23		
Russian	1.10 ± 0.18	< 0.01	1.21 ± 0.44	< 0.01		
narrative genre	0.00 ± 0.06	0.94	−0.05 ± 0.07	0.46	−0.05 ± 0.07	0.43
framing predicate	−0.95 ± 0.06	< 0.01	−0.99 ± 0.06	< 0.01	−0.99 ± 0.06	< 0.01
stimulus card: speech bubble	−1.15 ± 0.41	< 0.01	−1.19 ± 0.44	< 0.01	−1.19 ± 0.44	< 0.01
stimulus card: thought bubble	2.02 ± 0.59	< 0.01	2.14 ± 0.62	< 0.01	2.13 ± 0.62	< 0.01

Table 9 Model comparison for inner- vs. outer-world predicates in the domain of reported speech and thought. The models include 7283 observations. Positive coefficients (β) are associated with more inner-world predicate usage. Significant results are shaded. The reference levels are Avatime for language, descriptive for genre, non-framing for framing predicate, and no bubble for stimulus card. For *Mod-L*, the variance for stimuli cards ($n = 17$) is $s^2 = 0.55$ (SD = 0.74). For *Mod-LI*, the variance for stimuli cards is $s^2 = 0.62$ (SD = 0.79), for language users ($n = 141$) $s^2 = 0.23$ (SD = 0.48), and for session groups ($n = 61$) $s^2 = 0.16$ (SD = 0.40). For *Mod-I*, the variance for stimuli cards is $s^2 = 0.63$ (SD = 0.79), for language users $s^2 = 0.24$ (SD = 0.49), and for session groups $s^2 = 0.58$ (SD = 0.76).

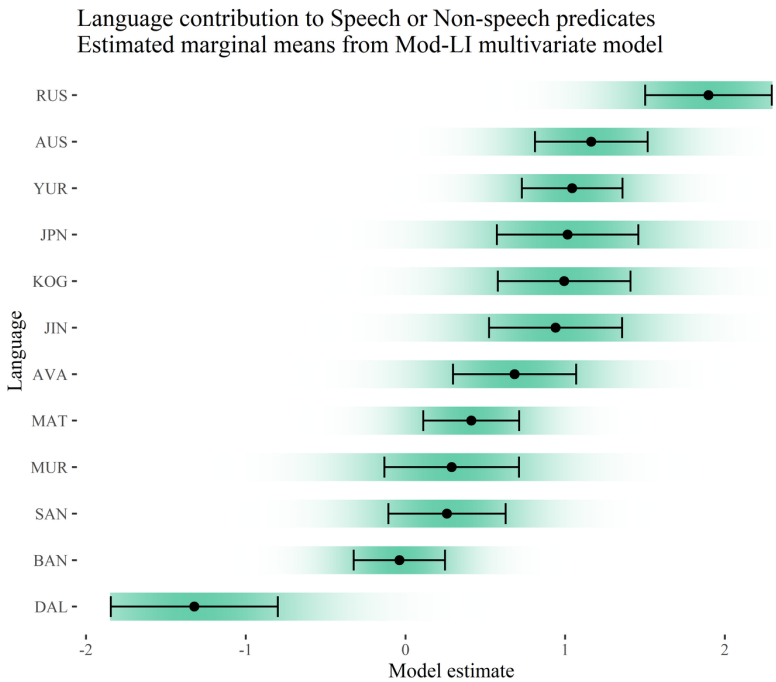

Figure 7 Model estimates with confidence intervals for the contribution of language to inner- or outer-world predicate choice.

studies. Finally, the picture card stimuli are stronger driving forces for our semantic/lexical variable than in our grammatical variable associated with reported speech, thought, and action, where they made little to no difference. Figure 7 shows the estimated means of language for our best model with confidence intervals to assess the contribution of language to predicate choice in our multivariate model.

4.4 Human reference (semantic/lexical)

Our final case study addresses the semantic domain of lexical terms for referring to humans. One of the most important things people do with language is talk about other people. Through reference to ourselves and to other humans we share social information and encode, enact, and alter social relations. However, while human reference is a universal behaviour, it is also a system with many choices (Stivers et al. 2007). The choices that a speaker has in any given moment in discourse (how will I refer to this person who is simultaneously *a father*, *a policeman*, *a Māori*, *Jack*, *my husband*, *your son*?) are greatly diversified across languages in regard to their role in linguistic and cultural practices.

Our data are coded for several categories of lexical reference terms, with the most frequent being kinship reference terms (*mother*, *his father*, *their child*) (12). The most common non-kinship strategy is generic terms (woman, man, people) (14). Some other strategies used to a minor degree include role nouns (*police*, *chief*, *doctor*), descriptive terms (*victim*, *good.one*, *bad.one*), numerical terms (*those three*), or making reference to a person based on their pictorial representation in the stimuli (*the one on the left*, *the one on the bottom*). This last strategy is minor in all but one of the languages in our sample. For Auslan, it is one of the primary human reference strategies, because many of the signers use the visual space available and point at referents depicted on the picture cards which were placed on the ground. Two examples can be found in (13), visualized in Figure 8.

(12) Kogi

hẽki	*hate-dweba*	*hẽki*	*a-skwá*	*ezhi*	*a-hwäsgwi*
DEM=SW	grandfather-old	DEM=SW	3SG.POSS-son	or	3SG.POSS-father.in.law

hálde=ki	*ahí*	*munzhi*
DEM=SW	3SG.POSS	woman/wife

'This is the grandfather. This is his [the old man's] son. Or his [the young man's] father in law. This one is his [the young man's] wife.'

(speaker Daniel Nuvita, SocCog_kog01-CNC_130619_1 00:11–00:18)

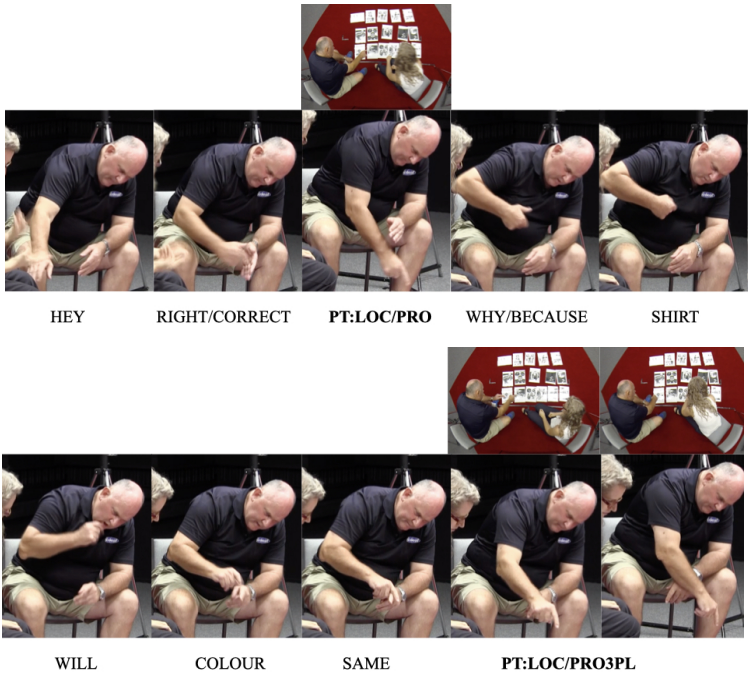

Figure 8 Example (13) – pictorial location/indexation. Figure produced by Gabrielle Hodge.

(13) Auslan

HEY RIGHT/CORRECT PT:LOC/PRO WHY/BECAUSE SHIRT WILL COLOUR SAME PT:LOC/PRO3PL

'Hey, that's-right (it's) **that-one-him-there**, because (the) shirt colour will-be coloured (the) same (for) **all-of-them-there**.'

(speaker DP, SocCog-asf01-DPBe12b 05:39–05:46)

(14) Japanese

nanka	*dansei-to*	*hanasi-te-(i)-ru*
apparently	male-COM	talk-GERUND(be)-NPST

'It seems... [she is] talking to a man.'

(speaker Takeshi Nakamoto, SocCog-jpn02-S_20130117_1HN 7:59–08:01)

Within the domain of human reference, one fundamental choice is whether or not to use a kinship term. Every language has words for kin. Kin terms are inherently relational (one cannot be a mother without having a child, or a sister without having a sibling) and the use of kinship terms highlights relationality as the key to formulating reference to someone else. This is relevant for social cognition since a person is identified through another person (e.g. *Mary's mother*, *my sister*) rather than through an independent characteristic they possess as an individual (i.e. *woman*, *teacher*, *tall person*, *Mary*). Kinship terms are central to sociality, with the power to define, track, regulate, and propagate networks of relationships both in the moment and across spatial and temporal divides. Our analysis of human reference compares the likelihood of using a kinship term versus a non-kinship term in reference.

For ten languages in our corpus there are three or more sessions with annotations for human reference. Across these languages, we included the 120 language users who produced 15 or more lexical references, excluding pronominal reference and any term that was ambiguous to whether or not it was a kinship term (i.e. a word similar to *Frau* in German that, depending on context, could mean either 'wife' or 'woman').

Table 10 presents summary statistics for the average amount of kinship terms versus non-kinship terms for each language, as well as the range of averages of each language user by language. As with the semantic linguistic domains in Section 4.3, there are very few instances of strong language patterning. Only for Auslan do we see consistent shading that goes across the language mean and minimum and maximum language user means. This means that all ten Auslan signers used more non-kinship than kinship terms. We see some overall language preferences: Balinese, Dalabon, Japanese, Jinghpaw, and Kogi have more kinship term usage, though some individual speakers

language	sessions ▼	users	tokens	semantic category	language mean	min. user mean	max. user mean
Auslan	5	10	238	kin	16.10%	12.32%	22.95%
			1240	other	83.90%	77.05%	87.68%
Avatime	4	8	292	kin	51.59%	30.56%	70.13%
			274	other	48.41%	29.87%	69.44%
Balinese	13	38	1929	kin	61.55%	33.33%	84.16%
			1205	other	38.45%	15.84%	66.67%
Dalabon	3	4	219	kin	69.52%	46.67%	83.87%
			96	other	30.48%	16.13%	53.33%
Japanese	3	8	190	kin	47.62%	26.14%	82.93%
			209	other	52.38%	17.07%	73.86%
Jinghpaw	3	6	455	kin	60.19%	52.78%	69.57%
			301	other	39.81%	30.43%	47.22%
Kogi	3	6	216	kin	62.79%	55.17%	80.95%
			128	other	37.21%	19.05%	44.83%
Matukar Panau	10	21	905	kin	55.18%	10.71%	84.62%
			735	other	44.82%	15.38%	89.29%
Murrinhpatha	3	4	54	kin	37.76%	22.73%	66.67%
			89	other	62.24%	33.33%	77.27%
Sanzhi Dargwa	4	6	305	kin	53.70%	39.36%	66.07%
			263	other	46.30%	33.93%	60.64%
Yurakaré	7	15	1207	kin	58.96%	44.72%	78.30%
			840	other	41.04%	21.70%	55.28%

Table 10 Distribution of kinship or other human referents across languages and language users. Shaded cells indicate over 60% usage for a structure in the language, or minimum and maximum user means.

use more non-kinship terms. Other languages in our sample do not have a strong language pattern. For some languages such as Matukar Panau, Murrinhpatha, and Sanzhi Dargwa, there are language users in the sample that have predominantly both kinship and non-kinship strategies.

As discussed above, language users are looking at picture stimuli cards

Figure 9 The stimulus card 'Family talking together', which depicts an intergenerational configuration.

while they formulate expressions. Our picture cards can be grouped into four broad categories relevant to person reference, based on how characters are configured. Seven have intergenerational configurations (cf. Figure 9), three have configurations with peer (or same-age) characters (cf. Figure 10), and two have solitary configurations (cf. Figure 11) or cards with police figures (cf. Figure 4 above).

We model the likelihood of languages using kinship versus non-kinship terms below. We ran three models, *Mod-L* which lacks a random effect for individuals, *Mod-I* which lacks a fixed effect for language, and *Mod-LI* which includes a fixed effect for language and the nested random effects for individuals. The results are reported in Figure 10, with significant effects in highlighted in colour. In each model our fixed-effects are (i) languages, (ii) genre (descriptive or narrative), and (iii) picture card stimuli configuration

Figure 10 The stimulus card 'Sitting and drinking', which depicts a peer configuration.

(intergenerational, peer, solitary, police, or non-specific/general).

The results in Table 11 show that *Mod-LI* is most likely to be the best one (for *Mod-LI*, $w \approx 1$) and that the AICc Δ values between it and the other models are high, especially between it and *Mod-L* ($\Delta = 377.98$). These results show that a model that does not take into account individual variability would incorrectly assign significant differences between our reference level and five of the languages in our sample. However, when we include group and individual random intercepts, only one significant language difference remains: This is Auslan, which has significantly less kinship usage than the rest of the languages in our sample. This is visibly clear in Figure 12, which shows the estimated means of language for *Mod-LI* with confidence intervals.

In all three models, genre has a significant role to play: narratives feature more kinship reference as participants put characters together, relate them to

Figure 11 The stimulus card 'Alone in cell', which depicts a solitary configuration.

each other, and motivate their actions. The fixed-effect of picture card stimulus has a significant effect for *Mod-L* only. In the other models it is a marginal effect: Intergenerational depictions with people from different age groups have more kinship reference while depictions with peers, police, or solitary characters have less. When confronted with a stimulus involving characters of different ages and genders, the inclination is to put these characters into a multigenerational family. The police context dramatically increases the proportion of other referent strategies, namely role nouns. When participants see the uniform of the policemen, all other dimensions of the person are diminished. Their gender or familial status is no longer relevant, simply their role – the intended purpose of uniforms is to diminish differences.

From this case study, we see that much variability is coming from individuals, however, some language differences remain and are helpful for modelling. We are currently exploring motivations for linguistic differences,

	Mod-L: language only		*Mod-LI:* both		*Mod-I:* individual only	
	Δ = 377.98, $w \approx 0$		AICc = 13092.11, $w \approx 1$		Δ = 29.83, $w \approx 0$	
	$\beta \pm$ SE	*p*	$\beta \pm$ SE	*p*	$\beta \pm$ SE	*p*
(intercept)	−0.03 ± 0.40	0.93	0.01 ± 0.51	0.98	−0.14 ± 0.42	0.74
Auslan	−2.00 ± 0.14	< 0.01	−2.06 ± 0.39	< 0.01		
Avatime	−0.36 ± 0.15	0.01	−0.47 ± 0.41	0.25		
Murrinhpatha	−0.54 ± 0.22	0.01	−0.43 ± 0.48	0.38		
Japanese	−0.56 ± 0.16	< 0.01	−0.19 ± 0.44	0.67		
Sanzhi Dargwa	−0.12 ± 0.15	0.40	−0.17 ± 0.42	0.69		
Matukar Panau	−0.06 ± 0.13	0.67	0.01 ± 0.36	0.98		
Jinghpaw	0.06 ± 0.14	0.68	0.06 ± 0.43	0.88		
Balinese	0.10 ± 0.12	0.40	0.15 ± 0.34	0.65		
Dalabon	0.67 ± 0.18	< 0.01	0.20 ± 0.37	0.59		
Yurakaré	0.12 ± 0.13	0.35	0.55 ± 0.46	0.24		
narrative genre	0.51 ± 0.05	< 0.01	0.48 ± 0.05	< 0.01	0.48 ± 0.05	< 0.01
card type: solitary	−0.84 ± 0.48	0.08	−0.89 ± 0.51	0.08	−0.89 ± 0.51	0.08
card type: police	−0.78 ± 0.42	0.07	−0.84 ± 0.45	0.06	−0.85 ± 0.45	0.06
card type: peers	−0.20 ± 0.44	0.64	−0.22 ± 0.46	0.64	−0.22 ± 0.47	0.63
card type: intergenerational	0.81 ± 0.40	0.04	0.82 ± 0.43	0.06	0.81 ± 0.43	0.06

Table 11 Model comparison for kinship vs. other human referents. The models include 11390 observations. Positive coefficients (β) are associated with more kinship referents. Significant results are shaded. The reference levels are Kogi for language, descriptive for genre, non-framing for framing predicate, and all/non-specific for stimulus card. For *Mod-L*, the variance for stimuli cards ($n = 17$) is $s^2 = 0.14$ (SD = 0.37). For *Mod-LI*, the variance for stimuli cards is $s^2 = 0.15$ (SD = 0.39), for language users ($n = 126$) $s^2 = 0.11$ (SD = 0.33), and for session groups ($n = 58$) $s^2 = 0.19$ (SD = 0.44). For *Mod-I*, the variance for stimuli cards is $s^2 = 0.16$ (SD = 0.40), for language users $s^2 = 0.11$ (SD = 0.33), and for session groups $s^2 = 0.61$ (SD = 0.78).

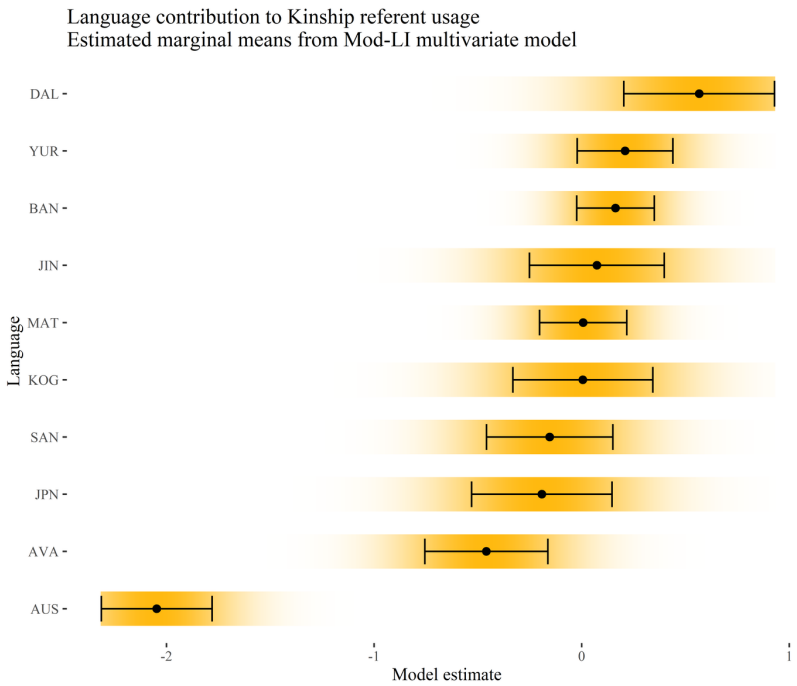

Figure 12 Model estimates with confidence intervals for the contribution of language to the lexical choice of kinship or non-kinship terms for human reference.

including the presence and types of kin-based grammar (Evans 2003) present in the languages in our sample (Barth et al., in preparation). Because we have included multiple additional sources of variability in our model, we can feel confident that the linguistic differences in our sample are real ones. The context variables of genre and picture card stimuli condition most of the variability in our model, an effect we saw play out for our other semantic domain as well, and opposed to the two grammatical domains.

5 Discussion

A number of clear findings emerge from our study:

a. Our models show that some significant differences between languages get reduced when we can account for the variable contribution from participants. Those of our models which include the effect from participants are a more accurate representation of where variability stems from, making better models. However, including language also makes for better models, despite our model comparison method having a built in penalisation for complexity. The extra complexity is outweighed by the more accurate predictions. We take this as support for inter-linguistic differences, but also a strong signal that we also must take individual difference into account in typological endeavours.

b. The relative effect of language on formulation choice varies with the domain. We have differing sets of languages (and language users) for our four test cases, but we can still assess the role of language by looking at how many language contributions are significantly different from our baseline language (i.e. reference level in statistical terminology) in each model. We see several significant contributions from languages in our grammatical variables, but far fewer for the semantic variables. We predict that even if we added data from more language users or more languages, we would continue to see such patterns: stronger linguistic effects for grammatical variables, stronger individual effects for semantic variables.

c. For all the domains investigated, we saw some language-user and session-group differences for grammatical variables and some linguistic differences for semantic variables, so both must be taken into account before making broad generalisations. In all domains, the effect of language is only partially determinate: Language users are strongly or weakly channelled towards certain expressive outcomes, but they nonetheless possess some choice in how they phrase things. This flexibility is, of course, what makes language change possible. An interesting next step for our data would be to investigate the kinds of participants that deviate from overall language patterns. Is

there anything special about these participants? Why do they do something different from other members of their community? Do contact effects – whether being multilingual oneself, or having frequent interactions with others who are – play a role here? If we were able to engage in the time-intensive endeavour of collecting data from hundreds of people from each linguistic community, would we see equal amounts of variation? Would the variation stabilise at some threshold of data volume? More data are always better, but are difficult to collect, particularly when consistent data need to be collected and annotated across multiple languages in a consistent way.

d. For at least some of the domains investigated there are differences in the degree of conformity among the users of each language. Taking the languages that have data for all four domains (Balinese, Dalabon, Kogi, and Matukar Panau) we see that some languages seem to have more inter-speaker variability than others. Dalabon showed fairly consistent speaker behaviour in both semantic and grammatical domains. Balinese has a great deal of difference between speakers; that may be at because of the large number of speakers in our sample, or because the Balinese data were deliberately collected from different dialects across the island. However, compare Matukar Panau and Kogi: There is less variability across the larger Matukar Panau sample of speakers than the smaller Kogi in the grammatical domain of reported speech. We leave as an open question the relative contribution of the number of grammatical options in the language, the sociolinguistic variability in the language community, and additional factors.

e. Language-user variability is not the only additional factor we need to take into account when describing inter-linguistic tendencies. For our semantic domains, context features such as genre and the picture card stimuli strongly drive linguistic behaviour. For our grammatical domains, these factors are less relevant.

These results emphasise the great opportunity that corpus studies offer to typology, by allowing us to incorporate individual variability when doing cross-linguistic analyses. They are also a salutary caution for more standard

approaches, since as we have seen, looking at data from one speaker or signer per language is not enough to assess how individual languages pattern (cf. Montero-Melis et al. 2017 in the experimental domain). Without considering multiple exemplars per language (Wälchli 2009), we risk essentialising our view of how particular languages work (cf. Haspelmath et al. 2005) due to a reduction or simplification of data (cf. Levshina 2019). While current corpus linguistics regularly addresses variability of speaker and signer, as well as register and genre (i.e. recommendations from Baayen et al. 2008 and Gries 2015), and has done this even for some lesser studied languages (Barth 2019; Calude et al. 2019; Lester et al. 2020; Meyerhoff & Klaere 2017; Schnell & Barth 2018; inter alia), this is much less common in typology – though see Heller et al. (2017), Torres Cacoullos & Travis (2019), and Seifart et al. (2018), among others. Engaging with individual variability is especially essential when looking at semantic variability, where it is well-known that individuals create variable categories of meaning based on diverse experiences (cf. Dor 2015; Labov 1973).

Finally, it is reassuring that existing methods such as regression modelling can handle corpus data when cross-linguistic typological variability is added in. As shown in our paper, they allow us to assess both the shape and the scope of variability in the data, and to carry out statistical hypothesis testing. The corpus approach allows us to include additional factors in assessing whether cross-linguistic variability exists and where that variability comes from. By being able to include inter-individual variability of language users in our models, we can produce more robust models which make us more confident about genuine interlanguage differences when we do find these soaring above the turbulence of individual differences.

References

Akaike, Hirotogu. 1973. Information theory and an extension of the maximum likelihood principle. In Petrox, Boris N. & Caski, Frigyes (eds.), *Second international symposium on information theory*, 267–282. Budapest: Akademiai Kiado.

Anand, Pranav & Chung, Sandra & Wagers, Matthew. 2010. *Widening the net: Challenges for gathering linguistic data in the digital age*. Response to *NSF SBE 2020:*

Future Research in the Social, Behavioral and Economic Sciences planning activity. (https://people.ucsc.edu/~schung/anandchungwagers.pdf).

Arnold, Jeffrey B. 2021. *ggthemes: Extra themes, scales and geoms for 'ggplot2'*. R package version 4.2.4. (https://cran.r-project.org/package=ggthemes).

Baayen, R. Harald & Davidson, Douglas J. & Bates, Douglas M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4). 390–412.

Barth, Danielle. 2019. Variation in Matukar Panau kinship terminology. *Asia-Pacific Language Variation* 5(2). 138–170.

Barth, Danielle & Evans, Nicholas. 2017. SCOPIC design and overview. In Barth, Danielle & Evans, Nicholas (eds.), *The Social Cognition Parallax Interview Corpus (SCOPIC): A cross-linguistic resource* (*Language Documentation & Conservation* special publication 12), 1–23. Honolulu, HI: University of Hawai'i Press. (https://hdl.handle.net/10125/24742).

Barth, Danielle & Evans, Nicholas. 2021. Social cognition in Dalabon. In Barth, Danielle & Evans, Nicholas (eds.), *The Social Cognition Parallax Interview Corpus (SCOPIC): A cross-linguistic resource* (Language Documentation & Conservation special publication 12), 22–79. Honolulu, HI: University of Hawai'i Press. (https://hdl.handle.net/10125/24743).

Barth, Danielle & Evans, Nicholas & Schalley, Andrea & San Roque, Lila & Mansfield, John & Gipper, Sonja & Hodge, Gabrielle & Rumsey, Alan & Arka, Wayan & Bergqvist, Henrik & Dickson, Greg & Döhler, Chrisitan & Pratiwi, D. P. Eka & Forker, Diana & Gast, Volker & Guntsetseg, Dolgor & Kashima, Eri & Kelly, Barbara & Kimoto, Yukinori & Knuchel, Dominique & Kogura, Norikazu & Kurabe, Keita & Narrog, Heiko & Schnell, Stefan & Senge, Chikako & Skribnik, Elena & van Putten, Saskia. In preparation. Does word choice mirror grammar? The case of kinship.

Barth, Danielle & Kapatsinski, Vsevolod. 2018. Evaluating logistic mixed-effects models of corpus-linguistic data in light of lexical diffusion. In Speelman, Dirk & Heylen, Kris & Geeraerts, Dirk (eds.), *Mixed-effects regression models in linguistics*, 99–116. Dordrecht: Springer.

Bates, Douglas M. & Maechler, Martin & Bolker, Ben & Walker, Steve. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. (https://doi.org/10.18637/jss.v067.i01).

Bresnan, Joan & Cueni, Anna & Nikitina, Tatiana & Baayen, Harald R. 2007. Predicting the dative alternation. In Boume, Gerlof & Krämer, Irene & Zwarts, Joost (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Burnham, Kenneth P. & Anderson, David R. 2002. *Model selection and multimodel inference: A practical Information-Theoretic approach*. 2nd edn. New York: Springer.

Calude, Andreea S. & Harper, Sally & Miller, Steven & Whaanga, Hemi. 2019. Detecting language change: Māori loanwords in a diachronic topic-constrained corpus of New Zealand English newspapers. *Asia-Pacific Language Variation* 5(2). 109–137.

Chafe, Wallace (ed.). 1980. *The Pear Stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood, NJ: Ablex.

Clark, Herbert. 1996. *Using Language*. Cambridge: Cambridge University Press.

Coulmas, Florian. 1986. Reported speech: Some general issues. In Coulmas, Florian (ed.), *Direct and indirect speech*, 1–28. Berlin: Mouton de Gruyter.

Cristofaro, Sonia. 2003. *Subordination*. Oxford, New York: Oxford University Press. (https://doi.org/10.1093/acprof:oso/9780199282005.001.0001).

Croft, William. 2010. The origins of grammaticalization in the verbalization of experience. *Linguistics* 48(1). 1–48.

Dor, Daniel. 2015. *The instruction of imagination: Language as a social communication technology*. Oxford: Oxford University Press.

Enfield, Nick. 2014. *Natural causes of language: Frames, biases and cultural transmission*. Berlin: Language Science Press.

Enfield, Nick & Levinson, Steven C. (eds.). 2006. *Roots of human sociality: Culture, cognition and human interaction*. Oxford: Berg.

Evans, Nicholas. 2003. Context, culture, and structuration in the languages of Australia. *Annual Review of Anthropology* 32(1). 13–40.

Evans, Nicholas. 2010. *Dying words: Endangered languages and what they have to tell us*. Malden, MA: Wiley.

Evans, Nicholas. 2013. Some problems in the typology of quotation: a canonical approach. In Brown, Dunstan & Chumakina, Marina & Corbett, Greville G. (eds.), *Canonical morphology and syntax*, 66–98. Oxford: Oxford University Press.

Ferrara, Lindsay & Johnston, Trevor. 2014. Elaborating who's what: A study of constructed action and clause structure in Auslan (Australian Sign Language). *Australian journal of linguistics* 34(2). 193–215.

Gries, Stefan T. 2015. The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora* 10(1). 95–125.

Haspelmath, Martin & Dryer, Matthew S. & Gil, David & Comrie, Bernard. 2005. *World Atlas of Language Structures*. Oxford: Oxford University Press.

Heller, Benedikt & Szmrecsanyi, Benedikt & Grafmiller, Jason. 2017. Stability and fluidity in syntactic variation world-wide: The genitive alternation across varieties of English. *Journal of English Linguistics* 45(1). 3–27.

Hodge, Gabrielle & Johnston, Trevor. 2014. Points, depictions, gestures and enactment: Partly lexical and non-lexical signs as core elements of single clause-like units in

Auslan (Australian Sign Language). *Australian Journal of Linguistics* 34(2). 262–291.

Kimoto, Yukinori & Kogura, Norikazu & Barth, Danielle & Evans, Nicholas & Shiohara, Asako & Arka, Wayan & Kashima, Eri & Kasuga, Yuki & Kawakami, Carine & Kurabe, Keita & Narrog, Heiko & Nomoto, Hirkoi & Ono, Hitomi & Pratiwi, D. P. Eka & Rumsey, Alan & Yokoyama, Akiko. In preparation. Framing a proposition in language use: Typological study of complement clauses and their alternatives.

Kuznetsova, Alexandra & Brockhoff, Per B. & Christensen, Rune H. B. 2017. *lmerTest: Tests in linear mixed effects models.* R package version 2.0-3.3. (`https://cran.r-project.org/package=lmerTest`).

Labov, William. 1973. The boundaries of words and their meanings. In Bailey, Charles-James N. & Shuy, Roger W. (eds.), *New ways of analyzing variation in English.* Washington, D.C.: Georgetown University Press.

Lenth, Russell V. 2021. *emmeans: Estimated marginal means, aka least-squares means.* R package version 1.7.0. (`https://cran.r-project.org/package=emmeans`).

Lester, Nicholas & Bickel, Balthasar & Moran, Steven & Stoll, Sabina. 2020. Speech rates differentiate nouns and verbs in child-surrounding and child-produced speech: Evidence from Chintang. In Brown, Megan M. & Kohout, Alexandra (eds.), *Proceedings of the 44th Annual Boston University Conference on Language Development*, 280–293. Somerville, MA: Cascadilla Press.

Levinson, Stephen C. 2003. *Space in language and cognition: Explorations in cognitive diversity.* Cambridge University Press. (`https://doi.org/10.1017/CBO9780511613609`).

Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Languages in Contrast* 23(3). 533–572. (`https://doi.org/10.1515/lingty-2019-0025`).

Lucy, John. 1992. *Language diversity and thought: A reformulation of the linguistic relativity hypothesis.* Cambridge: Cambridge University Press.

Meyerhoff, Miriam & Klaere, Steffen. 2017. A case for clustering speakers and linguistic variables. In Buchstaller, Isabella & Siebenhaar, Beat (eds.), *Language Variation – European Perspectives VI*: *Selected papers from the Eighth International Conference on Language Variation in Europe (ICLaVE 8), Leipzig, May 2015*, 23–46. Amsterdam: John Benjamins.

Montero-Melis, Guillermo & Eisenbeiß, Sonja & Narasimhan, Bhuvana & Ibarretxe-Antuñano, Iraide & Kita, Sotaro & Kopecka, Anetta & Lüpke, Friederike & Nikitina, Tatiana & Tragel, Ilona & Jaeger, T. Florian & Bohnemeyer, Jürgen. 2017. Satellite vs. verb-framing underpredicts nonverbal motion categorization: Insights from a large language sample and simulations. *Cognitive Semantics* 31(1). 36–61.

Noonan, Michael. 1985. Complementation. In Shopen, Timothy (ed.), *Language typology and syntactic description*, 42–140. Cambridge: Cambridge University Press.

Nuñez, Rafael E. & Celik, Kenan & Nakagawa, Natsuko. 2019. Absolute spatial frames of reference in bilingual speakers of endangered Ryukyuan languages: An assessment via a novel gesture elicitation paradigm. *Proceedings of the Annual Meeting of the Cognitive Science Society* 41. 890–896.

Quené, Hugo & van den Bergh, Huub. 2008. Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language* 59(4). 890–896.

R Core Team. 2021. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. (`https://r-project.org`).

San Roque, Lila & Rumsey, Alan & Gawne, Lauren & Spronck, Stef & Hoenigman, Darja & Carroll, Alice & Miller, Julia & Evans, Nicholas. 2012. Getting the story straight: Language fieldwork using a narrative problem-solving task. *Language Documentation and Conservation* 6. 134–173.

Schnell, Stefan & Barth, Danielle. 2018. Discourse motivations for pronominal and zero objects across registers in Vera'a. *Language Variation and Change* 30(1). 51–81.

Seifart, Frank & Strunk, Jan & Danielsen, Swintha & Hartmann, Iren & Pakendorf, Brigitte & Wichmann, Søren & Witzlack-Makarevich, Alena & de Jong, Nivja H. & Bickel, Balthasar. 2018. Nouns slow down speech across structurally and culturally diverse languages. *Proceedings of the National Academy of Sciences of the United States of America* 115(22). 5720–5725. (`https://doi.org/10.1073/pnas.1800708115`).

Spiess, Anrej-Nikolai. 2018. *qpcR: Modelling and analysis of real-time PCR data*. R package version 1.4-1. (`https://cran.r-project.org/package=qpcR`).

Stivers, Tanya & Enfield, Nick J. & Levinson, Stephen C. 2007. Person reference in interaction. In Enfield, Nicholas & Stivers, Tanya (eds.), *Person reference in interaction*: *Linguistic, cultural, and social perspectives*, 1–20. Cambridge: Cambridge University Press.

Sugiura, Nariaki. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics – Theory and Methods* 7(1). 13–26.

Tagliamonte, Sali A. & Baayen, R. Harald. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.

Tannen, Deborah. 1989. *Talking voices: Repetition, dialogue, and imagery in conversational discourse*. Cambridge: Cambridge University Press.

Tomasello, Michael. 2014. *A natural history of human thinking*. Harvard: Harvard University Press.

Torres Cacoullos, Rena & Travis, Catherine E. 2019. Variationist typology: Shared probabilistic constraints across (non-)null subject languages. *Linguistics* 57(3). 653–692.

Wagenmakers, Eric-Jan & Farrell, Simon. 2004. AIC model selection using Akaike weights. *Psychonomic Bulletin & Review* 11(1). 192–196.

Wälchli, Bernhard. 2009. Data reduction typology and the bimodal distribution bias. *Linguistic Typology* 13(1). 77–94.

Whorf, Benjamin L. 1956. *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. Cambridge, MA: MIT Press.

Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. New York: Springer.

Wilke, Claus O. 2021. *ungeviz: Tools for visualizing uncertainty with ggplot2*. R package version 0.1.0. (`https://github.com/wilkelab/ungeviz`).

Winawer, Jonathan & Witthoft, Nathan & Frank, Michael C. & Wu, Lisa & Wade, Alex R. & Boroditsky, Lera. 2007. The Russian blues: Effects of language on color discrimination. *Proceedings of the National Academy of Sciences* 108. 7780–7785.

Winter, Bodo. 2020. *Statistics for linguists: An introduction using R*. New York: Routledge.