Running head:

SATELLITE- VS. VERB-FRAMING UNDERPREDICTS MOTION CATEGORIZATION

Title:

Satellite- vs. verb-framing underpredicts nonverbal motion categorization: Insights from a large language sample and simulations

Authors and affiliations:

Guillermo Montero-Melis (Stockholm University)

Sonja Eisenbeiss (University of Essex)

Bhuvana Narasimhan (University of Colorado)

Iraide Ibarretxe-Antuñano (University of Zaragoza)

Sotaro Kita (University of Warwick)

Anetta Kopecka (CNRS & University of Lyon)

Friederike Lüpke (School of Oriental and African Studies, University of London)

Tatiana Nikitina (CNRS)

Ilona Tragel (University of Tartu & Beijing Foreign Studies University)

T. Florian Jaeger (University of Rochester)

Juergen Bohnemeyer (University at Buffalo - SUNY)

Corresponding author:

Guillermo Montero-Melis
Centre for Research on Bilingualism
Stockholm University
SE-106 91 Stockholm
Sweden

Acknowledgements:

The present work was funded by the Max Planck Society for the Advancement of Science (SE, FL, BN, JB), by a European Commission Marie Curie Fellowship (AK) and partly funded by the Spanish Government grant MovEsII: FFI2013-45553-C3-1-P (IIA) and by the NSF CAREER Award IIS-1150028 (TFJ).

Author contributions:

JB, SE and BN designed the study. AK, BN, FL, IIA, IT, JB, SE, SK and TN collected the data. GMM and TFJ designed and conducted the analyses and simulations with input from SE. GMM wrote the manuscript with active contributions by JB, SE, BN and TFJ, and comments from the rest of the authors.

Satellite- vs. verb-framing underpredicts nonverbal motion categorization: Insights from a large language sample and simulations

Abstract

Is motion cognition influenced by the large-scale typological patterns proposed in Talmy's (2000) two-way distinction between verb-framed (V) and satellite-framed (S) languages? Previous studies investigating this question have been limited to comparing two or three languages at a time and have come to conflicting results. We present the largest cross-linguistic study on this question to date, drawing on data from nineteen genealogically diverse languages, all investigated in the same behavioral paradigm and using the same stimuli. After controlling for the different dependencies in the data by means of multilevel regression models, we find no evidence that S- vs. V-framing affects nonverbal categorization of motion events. At the same time, statistical simulations suggest that our study and previous work within the same behavioral paradigm suffer from insufficient statistical power. We discuss these findings in the light of the great variability between participants, which suggests flexibility in motion representation. Furthermore, we discuss the importance of accounting for language variability, something which can only be achieved with large cross-linguistic samples.

Keywords:

motion events; cross-linguistic comparison; semantic typology; multilevel models; power analysis; statistical simulations; linguistic relativity; event categorization

1. Introduction

Talmy’s (1991; 2000) finding that most languages fall into one of two types which systematically differ in their linguistic encoding of motion (satellite-framed, S-, or verb-framed, V-languages) soon triggered a related question: Do speakers of different languages also form distinct conceptual representations of motion? The idea that relativistic effects may be found in the domain of motion was probably first inspired by Slobin's (1987; 1991) finding that speakers of S-framed languages are more likely to mention Manner of motion than speakers of V-framed languages.[1] Yet the results of a series of studies designed to address this question are not conclusive, as reviewed below. Most evidence points towards weak effects of language on non-verbal conceptualization of motion events (Finkbeiner et al. 2002; Gennari et al. 2002; Kersten et al. 2010; Montero-Melis & Bylund in press; Papafragou & Selimis 2010), while one study has not found any evidence of such an effect (Papafragou, Massey & Gleitman 2002). There is now increasing evidence that differences in the experimental task explain some of the differences in results (e.g. Kersten et al. 2010; Montero-Melis & Bylund in press). Here we focus on another major limitation that has received less attention: previous investigations have focused on just a handful of largely related languages.

The limitations of a small language sample are aggravated by the fact that each study typically only involves comparison of two languages at a time, and different studies employ different paradigms whose results are not straightforwardly comparable. As we outline in more detail below, these factors render problematic any generalization about the effect of language type (S or V) on motion representation. In the present work, we use what to our knowledge

[1] Slobin distanced himself from the Whorfian stance that language influences thought in general. Instead he formulated his more subtle “thinking-for-speaking” hypothesis, which postulates that language has an effect on the way we think while speaking or preparing to speak (Slobin 1996).

constitutes the largest sample of languages investigated to date to address cross-linguistic differences in the conceptualization of motion events.

For more than half a century a vivid debate has surrounded the question of whether language-specific patterns of grammar, lexicon, or usage affect how speakers come to see the world. Famously, Benjamin Lee Whorf suggested in his principle of linguistic relativity that "We cut nature up, organize it into concepts, and ascribe significances as we do, largely because we are parties to an agreement to organize it in this way—an agreement that holds throughout our speech community and is codified in the patterns of our language" (Whorf 1956:213). In other words, the language we speak shapes the way we think. At the other extreme, universalist accounts stress that humans share a common cognitive structure that permeates all fundamental aspects of cognition. According to this view, language is merely a means of communicating universal conceptual categories, and differences in how languages encode reality represent mere accidents of how these concepts are mapped onto different linguistic units (Gleitman & Papafragou 2012; Pinker 1994). Motion represents an interesting conceptual domain to investigate the relation between language and thought: perception of space and motion is rooted in a cognitive architecture common to all humans and even shared with other species (e.g., Snowden & Freeman 2004), yet we find systematic cross-linguistic variability in how motion is expressed in different languages.

Talmy's (1991; 2000) framing typology provides a stimulating starting point to examine this domain. The typology is based on where the main semantic component of a motion event—the *path* followed by the figure with respect to a ground—is linguistically encoded. Languages can be classified depending on how motion is characteristically expressed. S-languages such as

English encode Path information outside of the main verb root, typically in a verb satellite, such as *out* in (1); Manner is typically encoded in the main verb root, *rolled* in (1):

(1)	The ball	rolled	**out**	of	the box.
	FIGURE	MANNER	PATH		GROUND

In contrast, V-languages like Spanish lexicalize path information in the verb root (*salió*, 'exited' in (2)). Consequently, they require a separate expression for the Manner of the motion event. In Spanish, this information requires minimally a gerund (*rodando* 'rolling' in (2)), which can however be omitted without affecting the grammaticality of the sentence:

(2)	La pelota	**salió**	de	la caja	(rodando).
	the ball	exited	of	the box	rolling
	FIGURE	PATH		GROUND	(MANNER)

'The ball moved out of the box (rolling).'

The additional and optional syntactic position is taken to render Manner less codable in V-languages, while the open verb slot in S-languages results in Manner being encoded more routinely in discourse (Özçalışkan & Slobin 2003; Slobin 1996; Slobin 2003). This difference has led researchers to ask whether speakers of S-languages also pay more attention to Manner than V-language speakers when comparing motion events in a nonverbal task.[2]

A common paradigm to test relative attention to Path versus Manner has been to elicit forced-choice similarity judgments in triads, where participants have to compare a target motion event to one variant altering the Manner and one altering the Path (Finkbeiner et al. 2002; Gennari et al. 2002; Papafragou, Massey & Gleitman 2002; Papafragou & Selimis 2010). Participants' choices are taken to indicate their preference for categorizing events in terms of

[2] There are potential issues with this hypothesis, to which we return in the discussion section.

either Path or Manner. Finkbeiner and colleagues examined monolingual English (S) and Japanese (V) speakers and Japanese–English and Spanish–English bilinguals in their respective first languages, Spanish and Japanese (both V). They found a relativistic effect in a forced choice similarity task when the targets were presented prior to their variants: monolingual English speakers showed a significantly stronger tendency than the other groups to judge event similarity on the basis of Manner. However, this effect was not found when targets and variants were presented simultaneously so that there was no need for linguistic encoding as a way of committing the events to memory. Similarly, Gennari et al. (2002) found a significantly stronger same-manner bias in English than in Spanish speakers, but only when participants verbally described the targets before their similarity judgments were recorded. The effect disappeared when participants did not describe the events or were given a linguistic interference task while judging event similarity. Papafragou et al. (2002) found that native speakers of English (S) and Greek (considered V by the authors) both made same-manner choices at about chance level, while Papafragou and Selimis (2010) reported a stronger same-manner bias in English speakers than in Greek speakers only when the prompt encouraged linguistic encoding.[3] Finally, Kersten et al. (2010) found that English speakers attended more strongly to Manner than Spanish speakers in a supervised learning task in which participants did not overtly describe the events before or during the task.

Taken together, studies so far have shown some influence of linguistic encoding on motion representation, mostly under conditions that favor the online use of language (cf. thinking-for-speaking, Slobin 1996; Slobin 2003). While some of the conflicting findings might be due to

[3] We note that Papafragou and colleagues' treatment of Greek as a V–language conflicts with Talmy's (2000:66) characterization of Greek as a language in which V-type and S-type descriptions of most events are equally colloquial.

differences in task (e.g. Papafragou, Massey & Gleitman 2002 vs. Kersten et al. 2010), the literature also contains conflicting results even when the same experimental task is used (e.g. Papafragou, Massey & Gleitman 2002 vs. Gennari et al. 2002; or Finkbeiner et al. 2002). This raises a major question: to what extent are conflicting findings artifacts of the particular languages chosen?

Broad language samples are important in view of the now well-documented degree of linguistic variation in motion event framing within languages and between languages supposedly belonging to the same type (Bohnemeyer et al. 2007; Beavers, Levin & Tham 2010; Croft et al. 2010; Kopecka & Narasimhan 2012; Goschler & Stefanowitsch 2013). More generally, broader language samples will increase the ability to detect potentially existing effects of S- vs. V-framing (i.e., they will increase statistical power), because any effect of language type on motion cognition will necessarily be accompanied by some variability between languages. Therefore, adequate tests of hypotheses about typological categories based on behavioral data should account, not only for participant- and item-specific variability, but also for language-specific variability. Yet previous studies could not account for language variability because each ‘sample’ of S- or V-languages consisted of only one or two observations (i.e., one or two languages per type). In sum, statistical power for questions like the one pursued here will depend on a) the degree of variability between participants (and experimental stimuli) from the same language community, b) the degree of variability between languages from the same typological category, c) the number of participants per language, and d) the number of languages per typological category.

To shed more light on this concern, we conducted a forced-choice similarity judgment task—analogous in design to those reviewed above—with native speakers of nineteen genetically and typologically diverse languages. Our design and analyses let us gauge different sources of variability in the data: crucially, language-specific and participant-specific variability. We chose to test the Whorfian claim that language may affect "habitual thought" (Whorf 1956:134–159), i.e. that language may affect non-verbal behavior even when linguistic representations are not overtly evoked, because thinking-for-speaking type of effects (Slobin 1996) have received wide support in previous work, as reviewed above. For that reason, participants did not provide descriptions of the events prior to, or during, the similarity judgement task (they provided descriptions *after* the similarity judgement task, but this data is not treated in the current paper). In addition, we conducted Monte Carlo simulations to assess the power of our study and, by extension, of designs similar to ours.

2. Method

2.1 Participants

The participants were 12 adult native speakers of each language; see Table 1 for an overview of languages, genetic affiliation, home country of the population tested, the collaborators who collected the data, S/V-classification and source of the classification. The sample comprised 7 S-languages and 12 V-languages.

[Table 1]

2.2 Materials

The materials consisted of 72 motion event video-animations arranged in triads, each triad consisting of a target item and two variants (Figure 1).[4] The targets were 24 animations which systematically varied four manners of motion (SPIN, ROLL, BOUNCE, SLIDE), three scenarios

[4] Stimulus materials and the corresponding field manual entry (Bohnemeyer, Eisenbeiss & Narasimhan 2001) are freely accessible at http://fieldmanuals.mpi.nl/volumes/2001/event-triads/ .

with different ground objects (inclined ramp; field with tree and rock; field with hut and cave), and two directed paths (motion UP/RIGHT, DOWN/LEFT). For each of these targets (e.g. tomato-ROLLs-UP-RAMP, see Figure 1), we created a same-manner (and different-path) variant (e.g. tomato-ROLLs-DOWN-RAMP), and three types of same-path (and different-manner) variants (e.g., BOUNCE/SLIDE/SPIN-UP-RAMP). This resulted in 72 triads with a target clip, a same-manner variant and one of the three same-path variants. The variants were presented side by side, one second after the target-clip presentation ended (see Figure 1).

[Figure 1]

The 72 triads were distributed across 6 randomized presentation lists in a Latin-square design. Each list was given to two participants per language (in reverse presentation order). Each list contained 12 triads, with the target clips combining the four manners of motion with the three scenarios so that each participant saw all 12 combinations in the target clip. The number of items showing UP/RIGHT and DOWN/LEFT motion in the target- and variant-clips, as well as the manners of motion in the different-manner variants, was counterbalanced across the lists, as was the position in which the variants were presented on the screen. The position of the ground objects remained the same in all clips. These minimal variations in the triad clips allow us to take into account the effects of different manners, paths and scenarios, but they also make the test triads quite similar. Therefore, we added 38 filler triads to each list, which involved other types of events and variations (e.g. replacing either the agent or the goal in a possession-transfer event with another character) and aimed at preventing the participants from settling into a fixed response pattern.

2.3 Procedure

2.3.1 Similarity-judgment task

The tasks were performed on a PC with color screen. The triads were stored as individual files in ordered lists on the experimenter's PC and the experimenter started the presentation of each triad with a mouse-click when participants were ready. Participants were instructed to carefully watch the first clip of each triad, then to watch the two following scenes all the way to the end, and then point to "the one which is more similar to the first clip" (Bohnemeyer, Eisenbeiss & Narasimhan 2001:103–104). Instructions to participants were translated into their native languages (see pre-experimental elicitation task below). Instructions were presented verbally and five practice trials gave participants the chance to get familiarized with the procedure and to ask questions. Halfway through the experiment, participants were allowed a brief break. The experimenter noted down the response on a separate coding sheet.

2.3.2 Pre-experimental elicitation task

Cross-linguistic differences in the expression of the concept of similarity might influence how participants interpret the task (cf. Loucks & Pederson 2011). For example, one of the constructions used to express similarity in Tiriyó involves pretense (Sergio Meira, p.c.). When somebody says "B is more like A than C" what they mean is 'B is only pretending to be like A, but C is really like A'. Participants interpreting the task in the sense of detecting pretenders might systematically identify the *less* similar variants. Hence, before running the task, each contributor/experimenter was asked to determine with a different set of native speakers how the concept of graded similarity is expressed in the respective language. A brief questionnaire with instructions for evaluation was provided to the experimenter for this purpose (cf. Bohnemeyer, Eisenbeiss & Narasimhan 2001:109–110).

2.4 Analysis approach: modelling S- and V-type as populations of languages

The present study addresses Whorf's hypothesis in the motion domain by sampling observations at the level of language (see Pederson et al. 1998; Bohnemeyer et al. 2014; Bohnemeyer et al. under revision for similar approaches in a different semantic domain). The rationale is that one needs to consider several observations (i.e. languages) in order to draw conclusions about the larger populations of *all* S- and V-languages. Meanwhile, one can only study a given language through its speakers. Therefore we will have to account for two sources of variation when testing the effects of language type (S or V) on motion conceptualization: variation between languages of the same type (henceforth *language variability*) and variation between participants of the same language (henceforth *participant variability*).

The following thought experiment illustrates this perspective (see Figure 2). Assume there is a Whorfian effect, such that speakers of S-languages have a higher mean probability than speakers of V-languages of categorizing events in terms of Manner rather than Path. Let the mean probabilities of Manner categorization be .77 and .59 for S-language and V-language speakers, respectively. These numbers would be true for the two populations; however, any given sample would show some amount of deviation with respect to the population from which it was drawn. How much on average it would deviate crucially depends on the amount of language and participant variability.

[Figure 2]

If participants and languages both showed relatively low variability (Figure 2, scenario A), languages would tend to cluster around their respective type means and participants would tend to cluster around their language means. In this scenario, it would be fairly easy to detect the Whorfian effect: type II errors—failures to detect a true effect—would be unlikely (i.e., statistical power would be high). If, in contrast, variability were high at both levels (Figure 2, scenario D), one would expect to observe languages that deviate from their respective type means simply because of chance. Hence, effects of S- vs. V-framing would be harder to detect and type II errors expected to be frequent (i.e., statistical power would be low).

Comparison of the relative amount of language and participant variability also offers one more insight (Figure 2, scenarios B and C). If language variability is high compared to participant variability (scenario B), this would suggest that Talmy’s two-way typology was missing out on important language-specific effects. Such a scenario could theoretically be suggestive of some kind of Whorfian effect (since speakers of different languages would systematically behave differently), but researchers would have to refine their linguistic account to capture language-specific variance not explained by the two-way typology. If, on the other hand, we found that participant variability was high compared to language variability (scenario C), it would suggest that participants’ nonverbal behavior was only weakly constrained by their language. Hence, Whorfian effects would not be strong and researchers would be well advised to further explore what explains variability at the *participant* (rather than language) level.

Multilevel regression models provide a suitable statistical framework for these questions, allowing researchers to control for various grouping factors that contribute to the overall variability in the behavioral responses (see Baayen, Davidson & Bates 2008; Gelman & Hill 2007; Jaeger 2008; Johnson 2009 for general introductions; Jaeger et al. 2011 for an application to linguistic typology). All analyses were conducted in R (R Core Team 2015) using the *lme4* package (Bates et al. 2015). Data and R Scripts to replicate all analyses are freely available at [REF to be provided after acceptance].

3. Results

3.1 S- versus V-framed languages

Figure 3 shows the proportion of same-manner choices by language and participant. Although there was a numerically higher proportion of same-manner choices among S-language than V-language speakers (0.63 vs. 0.58 respectively, see horizontal lines in Figure 3), the bootstrap estimated confidence intervals already suggest that the amount of language and participant variability drowns out these population differences.

[Figure 3]

To statistically assess the effect of language type we fitted a multilevel logistic regression model (Breslow & Clayton 1993; Jaeger 2008) predicting response type (same-manner = 1, same-path = 0) as the binary dependent variable from language type (S or V) as the single fixed-effects predictor. Language type was contrast-coded (S = 1, V = -1). The model included crossed random intercepts by language (accounting for language variability), by participant (accounting for participant variability), as well as for the scene shown in each triad and the contrast shown in that scene (both accounting for variability between items). The model formula in R was "SameMannerResponse ~ 1 + LanguageType + (1 | Language) + (1 | Participant) + (1 | ItemScene) + (1 | ItemWithinScene)".[5]

The intercept of the model indicated an overall reliable preference for Manner over Path categorizations across languages ($\hat{\beta}_0$ = .78, z = 2.77, p < .01). Critically, however, there was no significant difference in event categorization between language types, that is, speakers of S- and V-languages were equally likely to categorize events by Manner ($\hat{\beta}_{S\text{-}vs\text{-}V}$ = .17, z = 0.86, p > .3).

To better understand the sources of variability in response behavior, consider

[5] ItemScene refers to the three different scenarios or ground objects, ItemWihtinScene to the 72 different triads (see Method – Materials).

Table 2. It shows an estimate of the variability for each of the random effects included in the model. By far the largest source of variability comes from participants, whose standard deviation is about four times larger than that of the next largest random effect, language (reminiscent of scenario C in Figure 2). A standard deviation of about 2 logit units informally means that, while the average participant had a probability of about .7 of choosing a same-manner alternate (this is the estimate of the intercept, β_0 = .78 log-odds, converted to probability of a same-manner response), it would not be the least surprising to find participants who chose same-manner alternates with a probability of .94 or .23 (this is the mean +/− 1 standard deviation of the by-subject random intercept). Indeed, such extreme responses were very common in our sample, as can also be seen in Figure 3. Next, we analyze participant-specific patterns of responses.

[Table 2]

3.2 Effect of first choice on subsequent trials

The random effects structure in the main analysis revealed that the largest source of variability came from individual participants. That is, some participants were very likely to choose same-manner alternates whereas others were very unlikely to do so, even once language and language type were accounted for. An intriguing question is whether there was also large variability *within* subjects. Did subjects haphazardly switch between Path and Manner choices or did they largely settle on one categorization criterion? We examined this by gauging to what extent the first choice of a participant predicted the rest of their choices in the experiment. This is plotted in Figure 4, which shows that the choice on the first trial indeed was a good predictor of responses in the rest of the experiment: Participants who initially chose the same-manner alternate had a mean probability of 0.71 of continuing with this choice throughout the rest of the experiment, while the corresponding probability for participants who initially chose the same-path alternate was only 0.44.

[Figure 4]

We assessed the statistical significance of this effect by fitting a logistic multilevel regression model similar to our main model, but now adding choice on the first trial and its interaction with language type as fixed-effect predictors. To avoid redundancy in the data, all observations corresponding to responses to first trials were removed, as this information was already encoded in the predictors. Language type was contrast-coded as above (S = 1, V = -1) and choice on first trial was centered by subtracting the mean from the vector of observations. The model formula in R was "SameMannerResponse ~ 1 + LanguageType * SameMannerResponse_FirstTrial + (1 | Language) + (1 | Participant) + (1 | ItemScene) + (1 | ItemWithinScene)".

This new model also had a positive intercept ($\hat{\beta}_0 = .75$, $z = 2.74$, $p < .01$), reflecting the overall preference for same-manner responses. The effect of language type was not significantly different from zero, and actually became smaller compared to the main model ($\hat{\beta}_{S\text{-}vs\text{-}V} = .07$, $z = 0.40$, $p > .6$). Critically, the effect of choice on first trial was large and highly significant as Figure 4 suggested ($\hat{\beta}_{manner\text{-}on\text{-}first\text{-}trial} = 2.01$, $z = 6.22$, $p < .001$). In other words, participants did not appear to haphazardly switch between Path and Manner choices; rather their overall categorization preference could be predicted from the first trial. Finally, there was no interaction between language type and choice on first trial ($\hat{\beta}_{S\text{-}vs\text{-}V*manner\text{-}on\text{-}first\text{-}trial} = 0.06$, $z = 0.32$, $p > .8$), indicating that choice on first trial predicted subsequent choices for speakers of both language types in the same way.

3.3 Type I and type II error assessment using statistical simulations

The main analysis indicated that there was no reliable difference in event categorization between speakers of S- and V-languages, despite having a sample of 19 languages and 228 participants, a large sample judged by current standards in cross-linguistic experiments in psycholinguistics. Next we put this result into context by estimating how likely we were to falsely reject the null hypothesis (type I error) or to fail to find a truly existing effect (type II error). To this end, we conducted Monte Carlo simulations (see Mooney 1997 for a general introduction to Monte Carlo methods; Johnson 2009; Jaeger et al. 2011 for examples in linguistics). In brief, we generated a very large number of random data sets based on a range of parameters estimated from our original analysis (the exact parameters are detailed below), and we fitted new models to these simulated data sets. By aggregating the results of the simulations, one can estimate type I and type II errors under different scenarios, notably for different effect sizes and for different sample sizes. The approach we take here can be applied to any similar cross-linguistic question.

3.3.1 False rejections of the null hypothesis (type I errors)

We first establish that our main analysis does not lead to high type I error rates. Type I errors occur when the null hypothesis is true (i.e., no difference between S- and V-language speakers), yet the analysis yields a spurious significant effect. The desired type I error rate of an analysis is equal to its α-level, typically .05 in the behavioral sciences. Intuitively, this means researchers are ready to incur one type I error out of twenty times when there is no true effect in the population. However, analyses that do not take into account the structure of the data risk inflating type I error rates (cf. Jaeger, Pontillo & Graff 2012). Thus, the first simulation compares type I errors of four different multilevel regression models: our full model, which accounts for random variability by language, participant and item; a second model that does not account for by-participant variability; a third model that does not account for by-language variability; and a fourth model that does neither account for by-participant nor by-language variability.

We generated 10000 random data sets based on a null effect of language type; that is, assuming no population difference between S- and V-speakers. Otherwise, the characteristics of simulated data sets were like in our study: they consisted of unbalanced samples of 19 languages (7 S, 12 V), with 12 subjects per language and 12 observations per participant; there were 72 target items administered to participants following a Latin square design. The by-language, by-participant and by-item variability in the simulated data was as estimated from the random effects in our original analysis.[6] In all simulations we kept the intercept constant and identical to that observed in our sample of 19 languages, i.e. a probability of .69 across languages to choose manner over path alternates. Each of the four models was fitted on each data set. Convergence failures were excluded (2.6% of the fitted models).

[Figure 5]

Figure 5 shows the results of these simulations. Each bar shows the proportion of analyses that yielded spurious significant effects, as a function of the model used to analyze the simulated data. The type I error rate of 7.7% for the full model (leftmost bar) suggests that our analyses stayed close to the intended α-level of .05; in other words, they did not suffer from severe anti-conservativity. Removal of the random by-subject intercept did not notably increase the Type I error (7.9%, second bar); however, removal of the random by-language intercept increased type I error rate to 12.8% (third bar), and removal of both by-participant and by-language intercepts

[6] This is a simplifying assumption: just like our main analysis leaves uncertainty about the actual difference between S- and V-languages, it leaves uncertainty about the actual variances associated with items, participants and languages. All our simulations ignore this uncertainty for the sake of computational feasibility.

increased it to 47.5% (rightmost bar), leading to type I errors about half of the time. These simulations strongly suggest that studies that fail to account for by-language variability risk spurious significances. We now turn to the issue of power.

3.3.2 Power of our analysis (type II error)

Statistical power is the probability of detecting a significant difference when there really is one (Cohen 1988). It is equal to 1 minus the type II error rate. A general recommendation for the behavioral sciences is to run studies with power at least at .80; however, insufficient power is far from infrequent (cf. Cohen 1988). Our solution to assessing power is also general and should thus be relevant to researchers beyond those working on the Whorfian debate.

We conducted simulations following the same general logic as above: we generated random data sets and for each of them we fitted a single model analogous to the model of our main analysis; from each sample we recorded if the result yielded a significant effect of language type or not. All parameters were the same as in the type I error simulations, except for the value of the critical effect, namely the difference in likelihood of manner-choice by speakers of S- and V-languages. We let this effect vary between three values that are consistent with our data: the lower bound, the mean estimate and the upper bound of the confidence interval estimated from our main model (see Table 3). In other words, we generated data using a range of effect sizes that was likely to contain the *actual* effect. For example, our upper estimate (last row in Table 3) corresponded to a positive difference of 1.12 log-odds of choosing a same-manner alternate by speakers of S-languages compared to speakers of V-languages. Converted to the more familiar scale of probabilities, this upper estimate would imply that S-language speakers choose the manner alternate 79% of the time, whereas V-language speakers make this choice only 56% of the time.[7]

[Table 3]

Figure 6 shows the power of our analysis with our current sample, as a function of the effect of language type. Bar heights indicate the proportion of simulations that yielded differences between language types at a significance level of .05, out of a total of 10000 simulations per cell from which convergence failures were removed (Appendix A reports the proportion of convergence failures). For the lower and mean estimates (first two panels), power was very low (<.25). Only if we assume the most extreme effect still consistent with our data does the power increase slightly above the minimum conventionally recommended level of .80 (third panel). This shows that, despite having a data set that is much larger than any previous study on this question, our study was likely underpowered. Hence, we may ask how many languages would be required in future work to achieve reasonable power.

[Figure 6]

3.3.3 How many languages are needed to achieve reasonable power?

How large a sample of languages would be required to increase power to at least .80? This last question was addressed conducting simulations with the same parametrizations as in the previous analysis, but now changing the number of languages to a sample of 20, 40 or 80 languages, of which half were S- and the other half V-languages. (All other things being equal, such balanced designs maximize the power to detect significant effects.)

[7] Upper and lower bounds were obtained by calculating the 95% confidence intervals of the coefficient for language type in our original model. We assume that our dependent variable is normally distributed in log-odds space, and so the two bounds correspond to the mean estimate +/− 1.96 * Standard Error of the mean.

The results are shown in Figure 7. Again, bars show the proportion of simulated samples that yielded significant effects of S- vs. V-framing, as a function of effect size (different panels) and sample size (x-axis within panels). We exemplify with the 2nd panel, which represents an effect size equal to the mean estimate of our original analysis. In this scenario, the probability of Manner categorization is .72 for S-language speakers and.65 for V-language speakers. If this were the true difference between language types, then even with an enormous sample of 80 languages and 960 participants, power would still not reach .50. In other words, we would detect an existing difference less than half of the time. Power is above .80 only if we assume the effect to be equal to our upper estimate (i.e., the most extreme effect still consistent with our data). Indeed, for the upper estimate, balanced samples of 20 languages have a power of .86, while samples of 40 or 80 languages reach very high power (>.98). But of course it is unlikely that the true effect is equal to the upper bound, which again suggests that even analyses based on balanced samples of as much as 40 or 80 languages could in fact remain underpowered.

[Figure 7]

4. Discussion

The present work tested the hypothesis that language type with respect to motion encoding (S or V) biases speakers toward categorizing events in terms of either Path or Manner. To this end, data was collected from speakers of nineteen genealogically and typologically diverse languages. We reasoned that the mixed evidence found in previous studies could in part have been an artefact of the small sample of languages tested in each study. We found no significant effect of language type (S or V) on event categorization: variability within language types was greater than variability between types. The two greatest sources of variability in the data came from languages and participants, with the latter being larger than the former, a result to which we return below. Overall, our main analysis suggests that being a speaker of an S- or V-language does not per se lead to a bias toward categorizing motion events in terms of Path or Manner, at least not for the paradigm employed in the present study (a point to which we return below).

In contrast to the large variability *between* participants, we found that individual participants were fairly consistent in their choices: after the first target trial, they had largely settled on a strategy for the whole experiment. This was revealed by a second analysis showing that a participant's choice on the first target trial was a good predictor of their choice in all subsequent trials. This might mean that the task, and perhaps each individual participant's interpretation of the task, has a strong influence on participants' responses.

Finally, we assessed type I and type II errors in a series of Monte Carlo simulations. The rate of type I errors showed that our analyses—which took into account both language- and participant-level variability—largely avoided the considerable anti-conservativity of the approach used in previous studies. Indeed, models that failed to account for language variability did substantially inflate type I errors. From this we conclude that modelling by-language intercepts—something which previous studies did not, and could not, do—is critical in avoiding false rejections of the null hypothesis.

Type II error analyses were of interest because a null result like the one obtained here is informative to the extent one can be reasonably sure that one *could* have found an effect had there really been one. We estimated the probability of detecting an effect that was consistent with our data, i.e. an effect of language type that fell within the 95% confidence interval obtained from our main analysis. With our current design consisting of 19 languages, the study appears to be under-powered. Given the large variability in the categorization data, even 80 languages would not be a guarantee for reasonable power to detect an effect. An even better approach (to be pursued in future work) might be to take into account the full uncertainty about the variability estimates that mixed models provide and to conduct Bayesian approaches to the null hypothesis, such as the Bayes Factor (e.g. Kass & Raftery 1995; Gelman et al. 2013).

In the reminder of this section, we further discuss the insights afforded by the approach and results of the present study—focusing in particular on the different sources of variability and what they tell us about motion event categorization.

4.1 Sampling languages to test cross-linguistic hypotheses

An important methodological contribution of this study has been to apply a statistically informed approach to testing the hypothesis that the dominant pattern of encoding motion events in language (S- or V-framing) is related to the conceptualization of motion events. This approach can in principle be applied to any hypothesis that connects a typological feature to any other domain of interest, be it other linguistic features or conceptual structure as in studies on linguistic relativity. Simply put, the design involves a two-stage sampling recipe: first, sample languages from the different typological categories; second, sample individual participants from these languages; let all participants carry out the same task under the same experimental conditions. The data thus obtained can be analyzed using multilevel regression models that properly account for the dependencies in the data as we have done here (see Gelman & Hill 2007 for a general introduction; Cysouw 2010; Atkinson 2011; Jaeger et al. 2011; Bohnemeyer et al. 2014; Bohnemeyer et al. under revision, for examples of similar analyses for typological data). Such an approach will give researchers firmer ground to conclude that it is the typological feature of interest that is related to the effect, rather than other aspects of speech communities which might accidentally co-vary with language type.

An additional benefit of using the current design and analysis comes from the informative output of multilevel regression models. Indeed, multilevel modelling allows researchers to provide quantified intuitions about what contributes most of the variability in the phenomenon under study. In the present case, we can gain important insights by considering what contributed most of the variability in the categorization data beyond the effect of language type (S or V). We next discuss two sources that are of critical theoretical importance: language variability and participant variability, keeping in mind that the latter was by a wide margin larger than the former.[8]

4.2 Language variability in the light of Talmy's two-way typology

Being able to assess variability between languages of the same type is a distinctive feature of our study. Since previous work has focused on just two or three languages at a time, it could not tease apart random variability between languages from the hypothesized systematic variability between language types. In general, language-specific differences in the tendency to focus on Path vs. Manner are expected by mere chance. However, the amount of variability is informative in the context of Talmy's (2000) binary typological distinction on which Whorfian studies on motion event cognition tend to rest. A conclusion to be drawn from our main analysis is that testing a *single pair* of S- and V-languages is not enough to be able to make inferences about S- and V-languages *in general*. To illustrate this point, consider again Figure 3. Had we randomly chosen a pair of S- and V-languages from the current sample, we could have found a language effect in the expected direction (Jalonke and Polish), a null result (Japanese and German) or an outcome that at least qualitatively would go in the opposite direction than expected (Estonian and

[8] The model also estimated variability at the item-level. Overall, by-item variability was small and, as additional analyses not reported here revealed, not critical to our conclusions. At the same time, item variability is unlikely to generalize to other studies with different stimuli. We thus do not discuss item variability any further.

French). We could even have found significant differences among languages of the same type (Jalonke and French). Thus, inferences based on small sets of languages should be treated with caution.

Maybe Talmy's binary typology provides the wrong framework to predict cross-linguistic differences in motion cognition? It has previously been argued that choices of languages should be based on a more nuanced understanding of the particular way in which a language encodes motion (e.g., Loucks & Pederson 2011). This argument is in line with the bulk of work on language variability in motion event descriptions showing that within-type variability is large and may be better described as a cline than as a binary distinction (Beavers, Levin & Tham 2010; Bohnemeyer et al. 2007; Croft et al. 2010; Filipović 2007; Ibarretxe-Antuñano 2009; Kopecka 2006; Matsumoto 2003; Nikitina 2008; Slobin 2004; Slobin et al. 2011). The present study does not directly speak to this question, since we treated variability between languages as random. What we can say, however, is that an inaccurate typological account is certainly not the whole story.

Had we observed a scenario of high language variability and low participant variability (scenario B in Figure 2), this would have supported the idea that Whorfian effects exist, but are not captured by Talmy's two-way typology. Indeed, under a strong Whorfian effect one would expect low participant variability in non-verbal performance among speakers of the same language (cf. Lucy 1992). Our results, however, suggest a scenario of relatively large participant variability and low language variability (like scenario C in Figure 2). We now focus on participant variability.

4.3 Participant variability suggests flexibility in motion representation

Possibly the most striking result of the present study is the great individual variability in categorization responses (unfortunately, individual response patterns are typically not reported in previous studies, but see Loucks & Pederson 2011:127). Figure 3 illustrated this point: the shapes representing individual participants covered the whole range of responses along the y-axis in virtually all of the languages. Why this remarkable individual variation? The argument we put forward is that motion event categorization in terms of either Path or Manner is flexible; that is, participants do not permanently prefer Path over Manner categorizations or vice versa and their preferences can easily be tweaked.

One piece of evidence for the flexibility in representing motion events comes from the lack of an inherent, language-independent bias towards Path or Manner across studies. For instance, Gennari et al. (2002) found a Path bias for all groups; Finkbeiner et al. (2002), as well as the present study, found a Manner bias; finally, Papafragou et al. (2002) found no bias. Since these studies all employed different experimental items, the conflicting biases strongly suggest that attention to Path and Manner is affected by the nature of the contrast shown in the scenes (see also Bohnemeyer under review; Zlatev, Blomberg & David 2010 for discussion).

Further evidence for the flexibility of mental representations comes from work on late bilinguals, which have shown that the language in which a task is carried out can bias responses toward patterns typical of speakers of that language (Kersten et al. 2010; Lai, Garrido Rodriguez & Narasimhan 2014), and also from the fact that attention to Path and Manner can be linguistically primed (Billman, Swilley & Krych 2000; Montero-Melis, Jaeger & Bylund 2016). These studies suggest a high degree of malleability in the conceptual representation of motion participants form during non-verbal tasks. Incidentally, the large individual variability in path vs.

manner categorization preferences suggests that there is no universal bias for one or the other, in which case we would expect considerably less variation between speakers.

While variability was high between participants, it was relatively low within participants: a participants' first response was largely predictive of the rest of responses. Self-predictiveness from one trial to another from the same participant is, of course, expected (it is in fact, part of the motivation for modelling participant variability). However, it is possible that the forced choice paradigm leads to higher intra-participant consistency in responses than other paradigms. It seems reasonable to assume that both components are salient, but that the dichotomous nature of the task, forcing participants to choose either Path or Manner, leads to equally dichotomous, and possibly conscious, strategies that would have no counterpart outside of the experimental situation (see Loucks & Pederson 2011 for discussion). If so, forced-choice tasks would provide a poor measure of habitual motion conceptualization. Other experimental paradigms that do not explicitly contrast Path and Manner choices should be preferred, such as supervised learning paradigms (Kersten et al. 2010), eye-tracking paradigms (Flecken, Carroll, et al. 2015) or similarity arrangement tasks (Montero-Melis & Bylund in press; Montero-Melis, Jaeger & Bylund 2016). These tasks might be more likely to conceal the experimental manipulation between Path and Manner, thus avoiding conscious strategies.

4.4 Future directions

This study has shown how a large-scale cross-linguistic approach is informative despite the failure to find an effect of language type. Being able to quantify variability at different levels provides valuable theoretical insights. Furthermore, the simulations we report illustrate how future cross-linguistic work (on motion or any other domain) can address the question of how many languages and how many participants per language should be sampled to achieve reasonable statistical power . While not yet common in cross-linguistic studies, power analyses are all the more important there because testing different populations is costly and entails considerable practical challenges. To facilitate similar approaches, all our analysis scripts are made publicly available [REF to be provided after acceptance].

In the present case, we found that the substantial variability at the levels of language and participant led to low power. One way forward for relativistic research on motion cognition is therefore to conceive of manipulations within subject or at least within language, so as to block these sources of variation. A means to achieve the former is to test bilingual speakers in their two languages (e.g. Filipović 2011; Athanasopoulos et al. 2015), while the latter can be achieved with between-subject manipulations in training studies (e.g., Casasanto 2008; Montero-Melis, Jaeger & Bylund 2016). Additionally, the choice of languages should not simply be based on status as S- or V-framed, but anchored on a more detailed understanding of how motion events are encoded in a given language.

Finally, future research will have to more carefully consider what type of cognitive processing is at work in different tasks. Forced choice tasks like the one used here and in previous research might be mediated by conscious and strategic thinking (as also pointed out by a reviewer). It has not been common to use tasks that tap onto more automatic, less conscious processing (but see e.g. Flecken, Athanasopoulos, et al. 2015 for such a paradigm in a related domain), and hypotheses about language effects at different levels of cognitive processing remain open for future work.

5. Conclusion

Over the last 25 years, Talmy's typology of motion event lexicalization has inspired several hypotheses about the possible effect of grammatical structure on the conceptual categories we form. Studies to date, however, have focused on a small sample of languages making it difficult to draw conclusions about language type in general. The present study tested the effect of language type by choosing a varied sample of languages within each type. We found no evidence that being a speaker of a satellite-framed as opposed to a verb-framed language led to a difference in the likelihood of categorizing motion events in terms of either Path or Manner. Languages of both types formed a continuum that spanned from a weak Path categorization preference to a clear Manner categorization preference. In addition, we found great individual variation between participants, even among those of the same language. This suggests that the specific lexicalization pattern of a language in Talmy's sense affects at most weakly motion conceptualization, at least when linguistic representations are not explicitly activated. Based on this and previous studies, we conclude that nonverbal event categorization is dynamic and effects of language on motion conceptualization are flexible.

Appendix A. Convergence failures in simulations

[Table 4]

References

Aksu-Koç, Ayhan. 1994. Development of linguistic forms: Turkish. In Ruth A. Berman & Dan I. Slobin (eds.), *Relating events in narrative: A crosslinguistic developmental study*, 329–388. Hillsdale, NJ: Lawrence Erlbaum Associates.

Athanasopoulos, Panos, Emanuel Bylund, Guillermo Montero-Melis, Ljubica Damjanovic, Alina Schartner, Alexandra Kibbe, Nick Riches & Guillaume Thierry. 2015. Two languages, two minds: Flexible cognitive processing driven by language of operation. *Psychological Science*. doi:10.1177/0956797614567509. http://pss.sagepub.com.ezp.sub.su.se/content/early/2015/03/06/0956797614567509 (22 March, 2015).

Atkinson, Quentin D. 2011. Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa. *Science* 332(6027). 346–349. doi:10.1126/science.1199295.

Baayen, R. Harald, D. J. Davidson & D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4). 390–412. doi:10.1016/j.jml.2007.12.005.

Bates, Douglas M., D. Maechler, Benjamin M. Bolker & S. Walker. 2015. *lme4: Linear mixed-effects models using Eigen and S4*. http://CRAN.R-project.org/package=lme4.

Beavers, John, Beth Levin & Shiao-Wei Tham. 2010. The typology of motion expressions revisited. *Journal of Linguistics* 46(2). 331–377. doi:10.1017/S0022226709990272.

Berthele, Raphael. 2006. *Ort und Weg: die sprachliche Raumreferenz in Varietäten des Deutschen, Rätoromanischen und Französischen*. . Vol. 16. (Linguistik — Impulse & Tendenzen). Berlin: Walter de Gruyter.

Billman, Dorrit, Angela Swilley & Meredyth Krych. 2000. Path and manner priming: verb production and event recognition. In Lila Gleitman & Aravind K. Joshi (eds.), *Proceedings of the twenty-second annual conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

Bohnemeyer, Jürgen. under review. Linguistic relativity: From Whorf to now. In Lisa Matthewson, Cécile Meier, Hotze Rullmann & Thomas Ede Zimmermann (eds.), *The Blackwell Companion to Semantics*.

Bohnemeyer, Jürgen. 2007. The pitfalls of getting from here to there: Bootstrapping the syntax and semantics of motion event expressions in Yucatec Maya. In Melissa Bowerman & Penelope Brown (eds.), *Cross-linguistic perspectives on argument structure: Implications for learnability*, 49–68. Mahwah, NJ: Lawrence Erlbaum. http://ubir.buffalo.edu/xmlui/handle/10477/38631 (16 June, 2016).

Bohnemeyer, Jürgen, Elena Benedicto, Katharine T. Donelson, A. Eggleston, C. K. O'Meara, G. Pérez Báez, R. E. Moore, et al. under revision. The cultural transmission of spatial cognition: Evidence from a large-scale study.

Bohnemeyer, Jürgen, Katharine T. Donelson, Randi E. Tucker, Elena Benedicto, A. Eggleston, A. Capistrán Garza, N. Hernández Green, et al. 2014. The cultural transmission of spatial cognition: Evidence from a large-scale study. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, 212–217.

Bohnemeyer, Jürgen, Sonja Eisenbeiss & Bhuvana Narasimhan. 2001. Event triads. In Stephen C. Levinson & Nick J. Enfield (eds.), *Manual for the field season 2001*, 100–114. Nijmegen: Max Planck Institute for Psycholinguistics.

Bohnemeyer, Jürgen, Nicholas J. Enfield, James Essegbey, Iraide Ibarretxe-Antuñano, Sotaro Kita, Friederike Lüpke & Felix K. Ameka. 2007. Principles of event segmentation in language: The case of motion events. *Language* 83(3). 495–532. doi:10.1353/lan.2007.0116.

Breslow, N. E. & D. G. Clayton. 1993. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* 88(421). 9–25. doi:10.2307/2290687.

Canty, Angelo & Brian Ripley. 2014. *boot: Bootstrap R (S-Plus) Functions*.

Casasanto, Daniel. 2008. Who's afraid of the Big Bad Whorf? Crosslinguistic differences in temporal language and thought. *Language Learning* 58. 63–79. doi:10.1111/j.1467-9922.2008.00462.x.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J.: L. Erlbaum Associates.

Croft, William, Jóhanna Barðdal, Willem Hollmann, Violeta Sotirova & Chiaki Taoka. 2010. Revising Talmy's typological classification of complex event constructions. In Hans Christian Boas (ed.), *Contrastive studies in construction grammar*, 201–235. Amsterdam: John Benjamins. http://public.eblib.com/EBLPublic/PublicView.do?ptiID=623335 (20 August, 2014).

Cysouw, Michael. 2010. Dealing with diversity: Towards an explanation of NP-internal word order frequencies. *Linguistic Typology* 14(2/3). 253–286. doi:10.1515/LITY.2010.010.

Filipović, Luna. 2007. *Talking about motion: a crosslinguistic investigation of lexicalization patterns*. Amsterdam: John Benjamins Pub.

Filipović, Luna. 2011. Speaking and remembering in one or two languages: bilingual vs. monolingual lexicalization and memory for motion events. *International Journal of Bilingualism* 15(4). 466–485. doi:10.1177/1367006911403062.

Finkbeiner, Matthew, Janet Nicol, Delia Greth & Kumiko Nakamura. 2002. The role of language in memory for actions. *Journal of Psycholinguistic Research* 31(5). 447–457.

Flecken, Monique, Panos Athanasopoulos, Jan Rouke Kuipers & Guillaume Thierry. 2015. On the road to somewhere: Brain potentials reflect language effects on motion event perception. *Cognition* 141. 41–51. doi:10.1016/j.cognition.2015.04.006.

Flecken, Monique, Mary Carroll, Katja Weimar & Christiane Von Stutterheim. 2015. Driving Along the Road or Heading for the Village? Conceptual Differences Underlying Motion Event Encoding in French, German, and French-German L2 Users. *Modern Language Journal* 99. 100–122. doi:10.1111/j.1540-4781.2015.12181.x.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari & Donald B. Rubin. 2013. *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall/CRC Press.

Gelman, Andrew & Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge; New York: Cambridge University Press.

Gennari, Silvia P., Steven A. Sloman, Barbara C. Malt & W. Tecumseh Fitch. 2002. Motion events in language and cognition. *Cognition* 83(1). 49–79. doi:16/S0010-0277(01)00166-4.

Gleitman, Lila & Anna Papafragou. 2012. New perspectives on language and thought. In Keith J. Holyoak & Robert G. Morrison (eds.), *The Oxford handbook of thinking and reasoning*, 543–568. Oxford: Oxford University Press.

Goschler, Juliana & Anatol Stefanowitsch (eds.). 2013. *Variation and change in the encoding of motion events*. . Vol. 41. (Human Cognitive Processing). Amsterdam: John Benjamins.

Hijazo-Gascón, Alberto & Iraide Ibarretxe-Antuñano. 2013. Las lenguas románicas y la tipología de los eventos de movimiento. *Romanische Forschungen* 125(4). 467–494.

Ibarretxe-Antuñano, Iraide. 2004. Language typologies in our language use: the case of Basque motion events in adult oral narratives. *Cognitive Linguistics* 15(3). 317–349.

Ibarretxe-Antuñano, Iraide. 2009. Path salience in motion events. In Jiansheng Guo, Elena Lieven, Nancy Budwig, Susan Ervin-Tripp, Keiko Nakamura & Seyda Ösçaliskan (eds.), *Crosslinguistic approaches to the psychology of language: research in the tradition of Dan Isaac Slobin*, 403–414. New York: Routledge.

Ibarretxe-Antuñano, Iraide. 2015. Going beyond motion events typology: The case of Basque as a verb-framed language. *Folia Linguistica* 49(2). 307–352. doi:10.1515/flin-2015-0012.

Ibarretxe-Antuñano, Iraide, Alberto Hijazo-Gascón & M. T. Moret. in press. The importance of minority languages in semantic typology: the case of Aragonese and Catalan. In Iraide Ibarretxe-Antuñano (ed.), *Motion and space across languages and applications*. Amsterdam: John Benjamins.

Jaeger, T. Florian. 2008. Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of memory and language* 59(4). 434–446. doi:10.1016/j.jml.2007.11.007.

Jaeger, T. Florian, Peter Graff, William Croft & Daniel Pontillo. 2011. Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology* 15(2). 281–319. doi:10.1515/LITY.2011.021.

Jaeger, T. Florian, Daniel Pontillo & Peter Graff. 2012. Comment on "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa." *Science* 335(6072). 1042–1042. doi:10.1126/science.1215107.

Johnson, Daniel Ezra. 2009. Getting off the GoldVarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis. *Language and Linguistics Compass* 3(1). 359–383. doi:10.1111/j.1749-818X.2008.00108.x.

Kass, Robert E. & Adrian E. Raftery. 1995. Bayes Factors. *Journal of the American Statistical Association* 90(430). 773–795. doi:10.1080/01621459.1995.10476572.

Kersten, Alan W., Christian A. Meissner, Julia Lechuga, Bennett L. Schwartz, Justin S. Albrechtsen & Adam Iglesias. 2010. English speakers attend more strongly than Spanish speakers to manner of motion when classifying novel objects and events. *Journal of Experimental Psychology: General* 139(4). 638–653. doi:http://dx.doi.org/10.1037/a0020507.

Kopecka, Anetta. 2006. The semantic structure of motion verbs in French: Typological perspectives. In Maya Hickmann & Stéphane Robert (eds.), *Space in languages: linguistic systems and cognitive categories*, vol. 66, 83–101. (Typological Studies in Language 0167–7373). Amsterdam: J. Benjamins.

Kopecka, Anetta & Bhuvana Narasimhan (eds.). 2012. *Events of putting and taking: a crosslinguistic perspective*. Amsterdam: John Benjamins. http://public.eblib.com/choice/publicfullrecord.aspx?p=901975 (25 January, 2016).

Lai, Vicky Tzuyin, Gabriela Garrido Rodriguez & Bhuvana Narasimhan. 2014. Thinking-for-speaking in early and late bilinguals. *Bilingualism: Language and Cognition* 17(1). 139–152. doi:http://dx.doi.org/10.1017/S1366728913000151.

Loucks, Jeff & Eric Pederson. 2011. Linguistic and non-linguistic categorization of complex motion events. In Jürgen Bohnemeyer & Eric Pederson (eds.), *Event Representation in Language and Cognition*, vol. 11, 108–133. (Language Culture and Cognition).

Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511782039.

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

Lüpke, Friederike. 2005. A grammar of Jalonke argument structure. Nijmegen: Radboud University Nijmegen Doctoral dissertation.

Matsumoto, Yo. 2003. Typologies of lexicalization patterns and event integration: Clarifications and reformulations. In Shuji Chiba (ed.), *Empirical and theoretical investigations into language: a festschrift for Masaru Kajita*, 403–418. Tokyo: Kaitakusha.

Meira, Sérgio. 2006. Approaching space in Tiriyó grammar. In Stephen C Levinson & David Wilkins (eds.), *Grammars of space: explorations in cognitive diversity*, 311–358. Cambridge, UK: Cambridge University Press.

Montero-Melis, Guillermo & Emanuel Bylund. in press. Getting the ball rolling: the cross-linguistic conceptualization of caused motion. *Language and Cognition*. doi:10.1017/langcog.2016.22.

Montero-Melis, Guillermo, T. Florian Jaeger & Emanuel Bylund. 2016. Thinking is modulated by recent linguistic experience: Second language priming affects perceived event similarity. *Language Learning*. n/a-n/a. doi:10.1111/lang.12172.

Mooney, Christopher Z. 1997. *Monte Carlo simulation*. . Vol. 116. (Quantitative Applications in the Social Sciences 99-143917–5). London: SAGE.

Narasimhan, Bhuvana. 2003. Motion events and the lexicon: a case study of Hindi. *Lingua* 113(2). 123–160. doi:10.1016/S0024-3841(02)00068-2.

Nikitina, Tatiana. 2008. Pragmatic factors and variation in the expression of spatial goals. In Anna Asbury, Jakub Dotlacil, Berit Gehrke & Rick Nouwen (eds.), *Syntax and semantics of spatial P*, 175–195. Amsterdam: John Benjamins. http://public.eblib.com/choice/publicfullrecord.aspx?p=623177 (6 November, 2015).

Nikitina, Tatiana. 2010. Variation in the encoding of endpoints of motion in Russian. In Viktoria Hasko & Renee Perelmutter (eds.), *New approaches to Slavic verbs of motion*, 267–290. Amsterdam: John Benjamins.

Özçalışkan, Seyda & Dan I. Slobin. 2003. Codability effects on the expression of manner of motion in English and Turkish. In A. Sumru Özsoy (ed.), *Studies in Turkish linguistics: proceedings of the Tenth International Conference on Turkish Linguistics*, 259–270. İstanbul: Boğaziçi Üniv.

Papafragou, Anna. 2008. Space and the language-cognition interface. In Peter Carruthers, Stephen Laurence & Stephen P Stich (eds.), *The Innate mind. Volume 3: Foundations and the future*, 272–289. (Evolution and Cognition). Oxford: Oxford University Press. http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195332834.001.0001/acprof-9780195332834 (27 October, 2011).

Papafragou, Anna, Christine Massey & Lila Gleitman. 2002. Shake, rattle, “n” roll: the representation of motion in language and cognition. *Cognition* 84(2). 189–219. doi:16/S0010-0277(02)00046-X.

Papafragou, Anna & Stathis Selimis. 2010. Event categorisation and language: A cross-linguistic study of motion. *Language and Cognitive Processes* 25(2). 224–260. doi:10.1080/01690960903017000.

Pederson, Eric, Eve Danziger, David G. Wilkins, Stephen C. Levinson, Sotaro Kita & Gunter Senft. 1998. Semantic typology and spatial conceptualization. *Lanugage* 74(3). 557–589.

Pinker, Steven. 1994. *The language instinct: the new science of language and mind*. London: Penguin.

R Core Team. 2015. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Sebastián, Eugenia & Dan I. Slobin. 1994. Development of linguistic forms: Spanish. In Ruth A. Berman & Dan I. Slobin (eds.), *Relating events in narrative: A crosslinguistic developmental study*, 239–284. Hillsdale, NJ: Lawrence Erlbaum Associates.

Slobin, Dan I. 1987. Thinking for speaking. *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*, 435–445. Berkeley, CA: Berkeley Linguisitcs Society.

Slobin, Dan I. 1991. Learning to think for speaking: native language, cognition, and rhetorical style. *Pragmatics* 1(1). 7–25.

Slobin, Dan I. 1996. From "thought and language" to "thinking for speaking." In John J. Gumperz & Stephen C. Levinson (eds.), *Rethinking linguistic relativity*, 70–96. Cambridge: Cambridge Univ. Press.

Slobin, Dan I. 2003. Language and thought online: cognitive consequences of linguistic relativity. In Dedre Gentner & Susan Goldin-Meadow (eds.), *Language in mind: advances in the study of language and thought*, 157–191. Cambridge, Mass.: MIT Press.

Slobin, Dan I. 2004. The many ways to search for a frog: Linguistic typology and the expression of motion events. In Sven Strömqvist & Ludo Th. Verhoeven (eds.), *Relating events in narrative. Vol. 2, Typological and contextual perspectives*, 219–257. Mahwah, N.J.: Lawrence Erlbaum.

Slobin, Dan I., Melissa Bowerman, Penelope Brown, Sonja Eisenbeiss & Bhuvana Narasimhan. 2011. Putting things in places: developmental consequences of linguistic typology. In Jürgen Bohnemeyer & Eric Pederson (eds.), *Event Representation in Language and Cognition*, vol. 11, 134–165. (Language, Culture and Cognition). Cambridge, UK: Cambridge University Press.

Snowden, Robert J. & Tom C. A. Freeman. 2004. The visual perception of motion. *Current Biology* 14(19). R828–R831. doi:10.1016/j.cub.2004.09.033.

Sugiyama, Yukiko. 2005. Not all verb-framed languages are created equal: The case of Japanese. *Proceedings of the 31st Annual Meeting of the Berkeley Linguistics Society*, vol. 31, 299–310. Berkeley, CA: Berkeley Linguisitcs Society.

Talmy, Leonard. 1991. Path to realization: a typology of event conflation. *Proceedings of the seventeenth annual meeting of the BLS*, 480–519. Berkeley: Berkeley Linguisitcs Society.

Talmy, Leonard. 2000. *Toward a cognitive semantics: Typology and process in concept structuring*. Cambridge, Mass.: MIT Press.

Tragel, Ilona & Ann Veismann. 2008. Kuidas horisontaalne ja vertikaalne liikumissuund eesti keeles aspektiks kehastuvad? [Embodiment of the Horizontal and Vertical Dimensions in Estonian Aspect]. *Keel ja Kirjandus* 7. 515–530.

Whorf, Benjamin Lee. 1956. *Language, thought, and reality. Selected writings of Benjamin Lee Whorf*. (Ed.) John B. Carroll. Cambridge, Mass.: MIT Press.

Zlatev, Jordan, Johan Blomberg & Caroline David. 2010. Translocation, language and the categorization of experience. In Vyvyan Evans & Paul A Chilton (eds.), *Language, cognition and space the state of the art and new directions*, 389–418. London: Equinox. http://public.eblib.com/choice/publicfullrecord.aspx?p=1069088 (27 May, 2016).

In-text tables & Figures

Table 1

Overview of languages.

Language	Affiliation	Country	Contributor	Type	Source
Basque	Isolate	Spain	Ibarretxe-Antuñano	V	Ibarretxe-Antuñano (2004; 2015)
Catalan	Romance	Spain	M. Martínez / M. Sauret / Bohnemeyer	V	Ibarretxe-Antuñano, Hijazo-Gascón, and Moret (in press)
Dutch	Germanic	Nether-lands	D. v. Exel/ Bohnemeyer	S	Talmy (2000)
English	Germanic	USA	M. Dixson	S	Talmy (2000)
Estonian	Finno-Ugric	Estonia	Tragel	S	Tragel and Veisman (2008)
French	Romance	France	Kopecka	V	Talmy (2000)
German	Germanic	Germany	K. Samland / Eisenbeiss	S	Berthele (2006)
Hindi	Indo-Iranian	India	Narasimhan	V	Narasimhan (2003)
Italian	Romance	Italy	M. Martínez / M. Sauret / Bohnemeyer	V	Hijazo-Gascón and Ibarretxe-Antuñano (2013)
Jalonke	Mande	Guinea	Lüpke	V	Lüpke (2005)
Japanese	Isolate	Japan	Kita	V	Sugiyama (2005)
Polish	Slavic	Poland	Kopecka	S	Talmy (2000)
Russian	Slavic	Russia	Nikitina	S	Nikitina (2010)
Spanish	Romance	Spain	M. Martínez / M. Sauret / Bohnemeyer	V	Sebastián and Slobin (1994)
Tamil	Dravidian	India	Narasimhan	V	Talmy (2000)
Tidore	West Papuan	Indonesia	M. v. Staden	V	M. v. Staden, pc
Tiriyó	Carib	Brazil	S. Meira	S	Meira (2006)
Turkish	Altaic	Turkey	A. Özyürek	V	Aksu-Koç (1994)
Yukatek	Mayan	Mexico	Bohnemeyer	V	Bohnemeyer (2007)

Table 2

Random effects for main model predicting same-manner response from language type.

Group	Variance	SD
Participant	4.23	2.06
Language	0.29	0.54
Scene in item	0.11	0.34
Item within scene	0.10	0.31

Number of observations: 2733. Groups: Participant, 228; Language, 19; Scene in item, 3; Item within scene, 72.

Table 3

Effect sizes used in power simulations.

Effect size	Difference in log-likelihood (S vs. V)	Probability of same-manner choice
Lower estimate	−0.43	S: .64; V: .73
Mean estimate	0.35	S: .72; V: .65
Upper estimate	1.12	S: .79; V: .56

Table 4

Model convergence failures during power simulations, broken down by number of languages and size of effect of language type.

Number of languages in simulation	Effect of language type (S vs. V)	Probability of same-manner choice	Samples	Convergence failures (count)	Convergence failures (%)
19	lower esimate	S: prob=0.64; V: prob=0.73	10000	439	4.4
19	mean estimate	S: prob=0.72; V: prob=0.65	10000	791	7.9
19	upper estimate	S: prob=0.79; V: prob=0.56	10000	941	9.4
20	lower esimate	S: prob=0.64; V: prob=0.73	10000	411	4.1
20	mean estimate	S: prob=0.72; V: prob=0.65	10000	669	6.7
20	upper estimate	S: prob=0.79; V: prob=0.56	10000	860	8.6
40	lower esimate	S: prob=0.64; V: prob=0.73	10000	147	1.5
40	mean estimate	S: prob=0.72; V: prob=0.65	10000	236	2.4
40	upper estimate	S: prob=0.79; V: prob=0.56	10000	279	2.8
80	lower esimate	S: prob=0.64; V: prob=0.73	10000	13	0.1
80	mean estimate	S: prob=0.72; V: prob=0.65	10000	34	0.3
80	upper estimate	S: prob=0.79; V: prob=0.56	10000	48	0.5

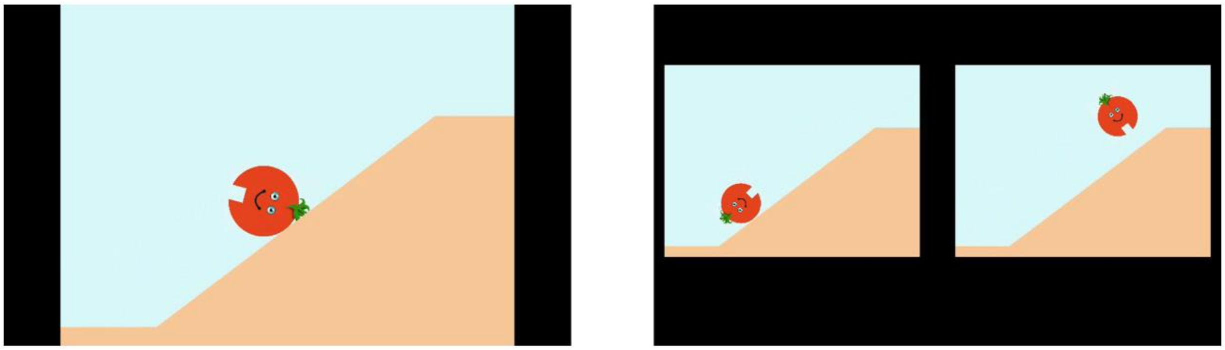

Figure 1 Example item. Left figure: ROLL UP target; right figure: ROLL DOWN same-manner variant (left panel) and BOUNCE UP same-path variant (right panel).

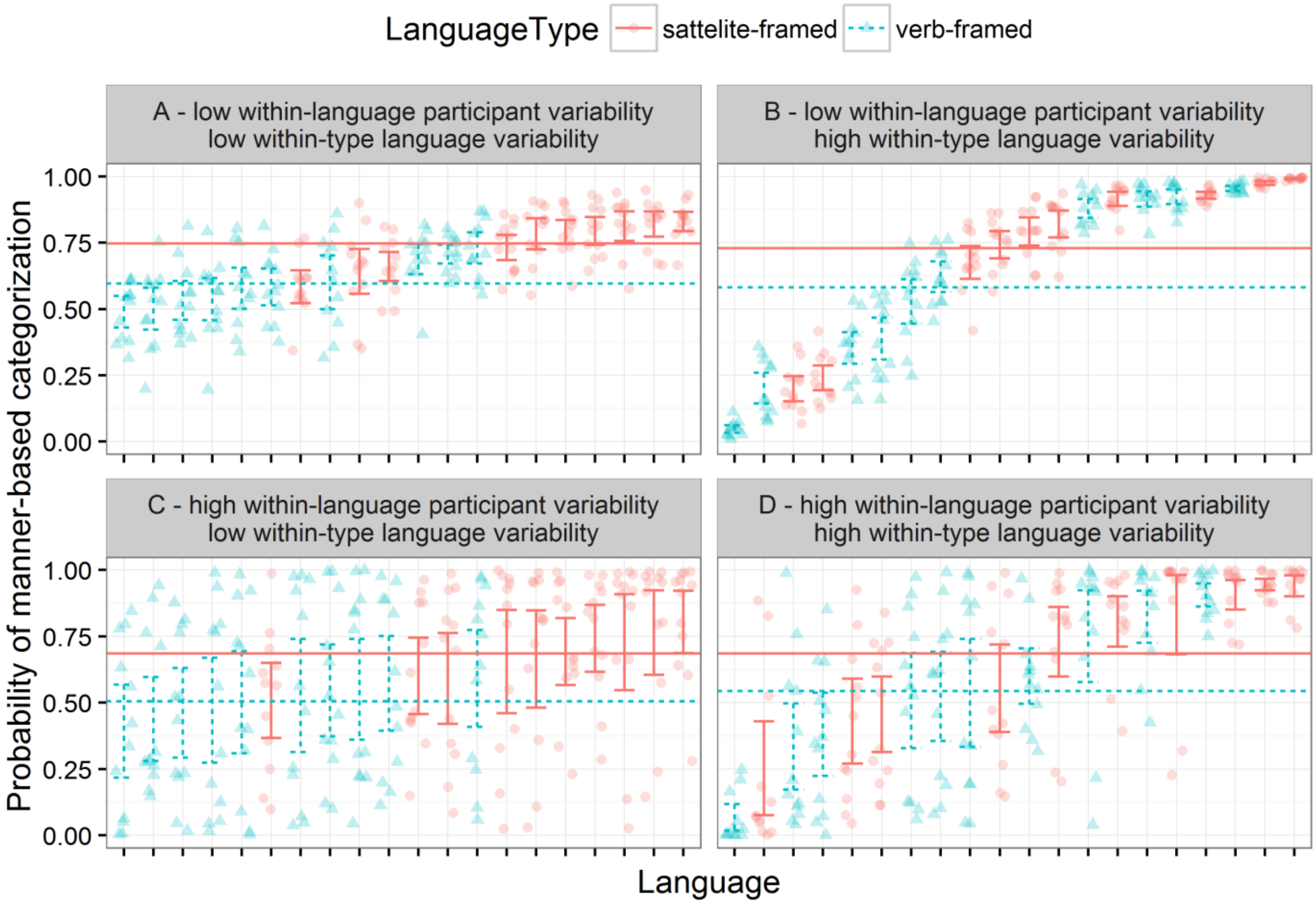

Figure 2 Four hypothetical scenarios under a Whorfian effect of language type (S vs. V). The y-axis shows the probability of categorizing an event in terms of Manner (rather than Path). Ticks along the x-axis represent a random sample of languages (10 S, 10 V); shapes represent participants (12 per language, as in the present study); error bars show confidence intervals per language; horizontal lines indicate empirical by-type means. Each panel is a random simulation from four distributions with the same underlying by-type mean (S > V). The panels differ with respect to the amount of participant variability (low/high) and language variability (low/high).

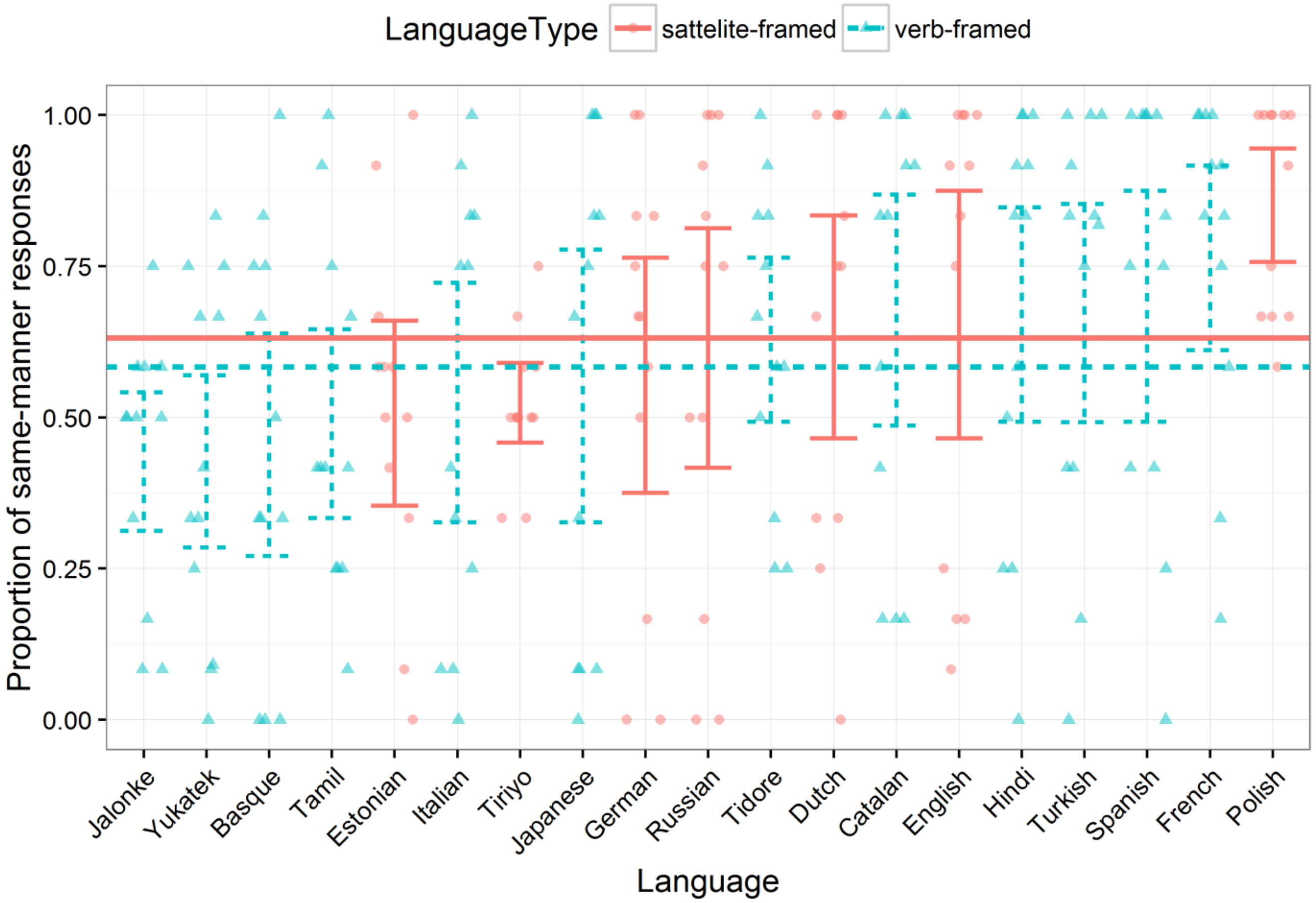

Figure 3 Proportion of same-manner choices by language (x-axis) and language type (S-framed: red dots and solid lines; V-framed: blue triangles and dashed lines). Shapes show by-participant averages, error bars show 95% confidence intervals obtained by non-parametric bootstrap over by-subject means (computed using the boot function, Canty & Ripley 2014). The two horizontal lines show average percentage of same-manner choices for each language type.

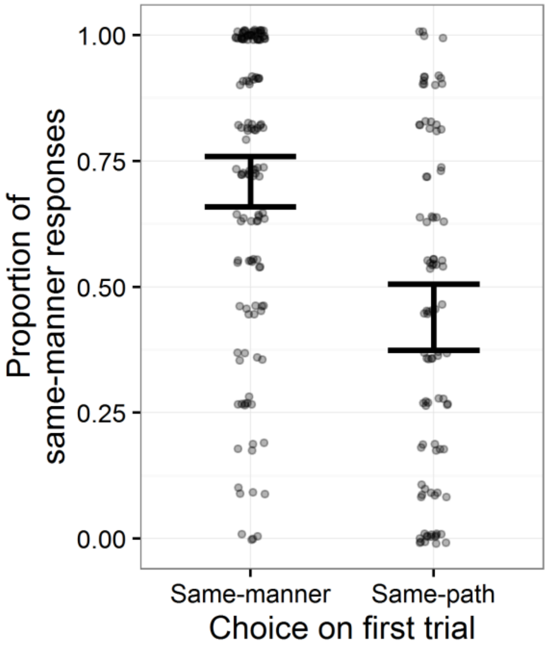

Figure 4 Effect of first trial choice (same-manner vs. same-path) on the proportion of same-manner choices in the rest of the experiment. Dots show jittered by-participant averages, error bars show non-parametric 95% confidence intervals of by-subject means.

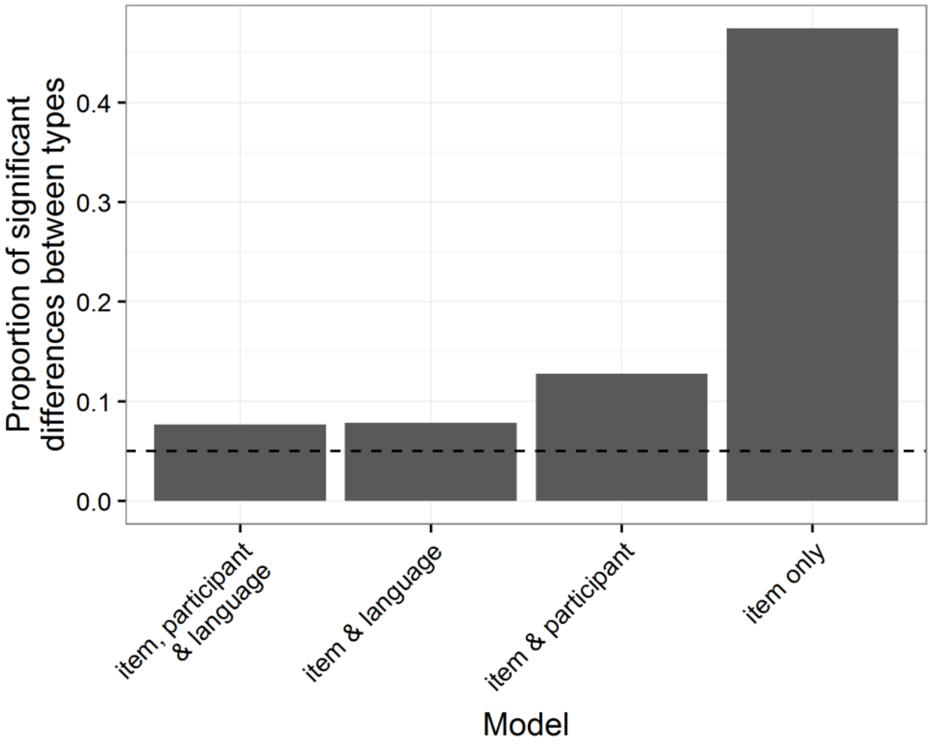

Figure 5 Rate of type I errors based on simulations. Bar heights show the proportion of simulation samples that yielded spurious significances, as a function of the random intercepts included in the model. The dashed horizontal line shows the α-level of .05. The figure is based on 10000 simulation samples from which convergence failures were excluded.

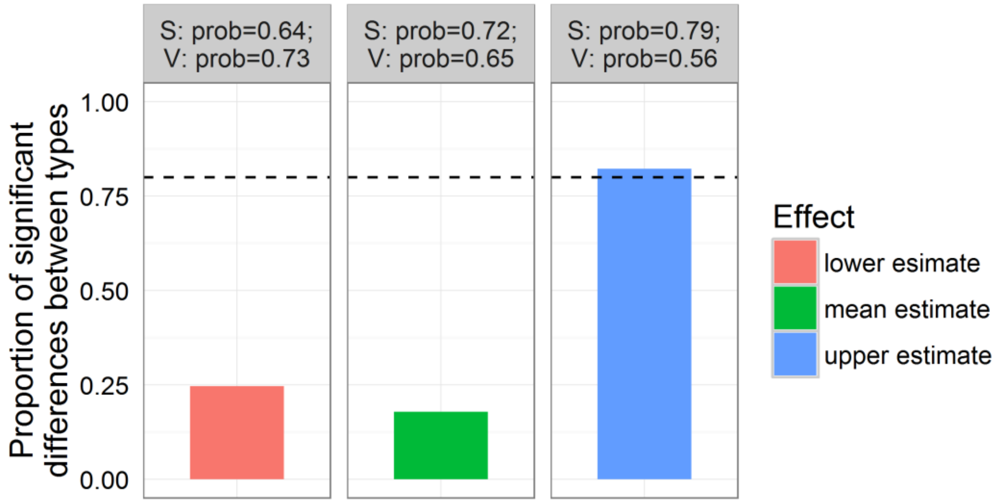

Figure 6 Power analysis based on current sample (7 S-, 12 V-languages) and three estimates of the effect of language type (10000 simulations per cell). Bar heights show the proportion of significant differences between language types at the .05 level. Panel titles and bar colors show the size of the population-level effect. Horizontal dashed lines mark a power of .08.

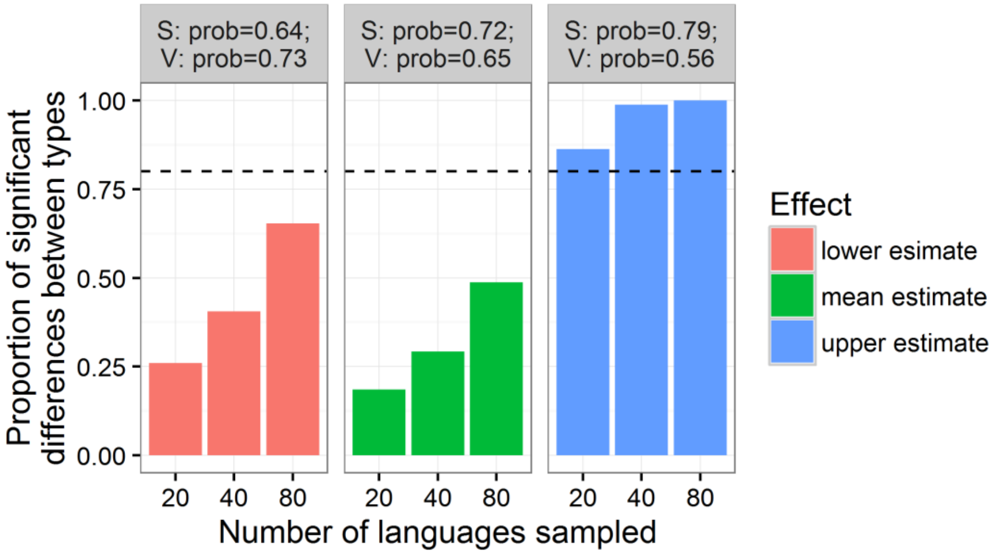

Figure 7 Power analysis based on balanced language samples of 20, 40 and 80 S/V-languages and three estimates of the effect of language type (10000 simulations per cell). Bar heights show the proportion of significant differences between language types at the .05 level. Panel titles and bar colors show the size of the population-level effect. Horizontal dashed lines mark a power of .08.