Tense systems across languages support efficient communication

Geoff Bacon[1] (bacon@berkeley.edu) Yang Xu[1] (yang_xu_ch@berkeley.edu)
Terry Regier[1,2] (terry.regier@berkeley.edu)

Department of Linguistics[1] and Cognitive Science Program[2]
University of California, Berkeley, CA 94720 USA

Abstract

All languages have ways of expressing location in time, but they differ widely in their grammatical tense systems. At the same time, there are tense systems that recur across unrelated languages. What explains this wide but constrained variation? Here, we propose that tense systems are shaped by the need to support *efficient communication*–a need that has recently been shown to explain cross-language semantic variation in other domains. We test this proposal computationally against the tense systems of 64 diversely sampled languages. We find that most languages in the sample support near-optimally efficient communication, but with some interesting and potentially illuminating exceptions. We conclude that efficient communication may play an important role in explaining why tense systems vary across languages in the ways they do.

Keywords: grammatical variation; time; tense; mental time line; efficient communication.

Time and tense systems

Time is one of the most fundamental aspects of human experience, and it occupies a significant position in the grammars and lexicons of natural languages (Quine, 1960; Hornstein, 1993; Klein & Li, 2009). However, linguistic systems of temporal expression differ substantially (Dahl, 1985; Bybee & Dahl, 1989; Bybee, Perkins, & Pagliuca, 1994). Klein (2009) describes six major ways in which languages express time: tense, aspect, Aktionsart (lexical aspect), temporal adverbials, temporal particles, and discourse principles. We focus on variation in *tense*, which is one of the most well-documented means of temporal expression (Binnick, 2012).

Tense is "the grammaticalised expression of location in time" (Comrie, 1985: 9). In some ways, tense systems are strikingly similar across languages. For example, there is a well-documented cross-language preference for more elaborate past tense categories than future tense categories (Comrie, 1985: 85). Yet in other ways, they vary considerably. For instance, English has grammatical categories to express the past, present and future. To locate an event of walking in the past, English uses the morphologically marked form "walked" to distinguish from "walk" in the present tense. To locate the same event in the future, English employs the auxiliary "will" to form the periphrastic "will walk". However, some languages have more elaborate tense systems than English, that specify not only whether an event is in the past or future, but also *how far* in the past or future it is. Kikuyu, for example, a Bantu language spoken in Kenya, uses different grammatical categories depending on whether an event took place very recently or a long time ago. Intuitively, Kikuyu's tense system is more precise than that of English at locating the time of an event. In contrast to languages like Kikuyu and English, some languages are tenseless, in that their grammars do not locate events in time at all. An example of a tenseless language is Cebuano, an Austronesian language of the Philippines. To express the same event of walking in Cebuano does not require any reference to when the walking takes place.

What explains this wide but constrained cross-linguistic variation? We seek general principles that explain why tense systems vary as they do, and why many logically possible tense systems are not attested.

Efficient communication

A recent proposal has the potential to explain variation in tense systems. By this account, systems of semantic categories across languages are shaped by the need to support *efficient communication*. This communicative principle has been shown to account for cross-linguistic variation in the semantic domains of color (Regier, Kay, & Khetarpal, 2007; Regier, Kemp, & Kay, in press), kinship (Kemp & Regier, 2012), space (Khetarpal, Neveu, Majid, Michael, & Regier, 2013) and numerosity (Xu & Regier, 2014). It also reflects a more general recent interest in communicative pressure as a source of explanation for linguistic structure (e.g. Piantadosi, Tily, & Gibson, 2011; Fedzechkina, Jaeger, & Newport, 2012; Smith, Tamariz, & Kirby, 2013). We hypothesize that this drive for efficient communication may also explain the variation we find in grammatical tense systems across languages.

The notion of efficient communication involves two competing forces: *informativeness* and *simplicity*. A communicative system is informative to the extent that it communicates precisely, whereas it is simple to the extent that its cognitive representation is compact. These two forces compete against each other. For example, the most informative tense system would have a unique linguistic form (e.g. a word or grammatical morpheme) to denote each temporal location. However, such a system would be highly complex, not simple. In contrast, the simplest system would have one linguistic form for all temporal locations. This would be simple, but would not support precise communication. The hypothesis of efficient communication proposes that languages reflect a near-optimal tradeoff between these two competing constraints.

Figure 1 illustrates a simple communicative scenario. Here, the speaker is thinking of a particular occasion of her having gone somewhere, which took place in the immediate past, e.g. earlier that morning. We represent time in terms of a discretized time line divided into seven units, spanning from the distant past to the distant future: remote past, recent past, immediate past, present (time of speech), immediate future, recent or intermediate future, and remote future.[1] Because the speaker is certain that the event took place in the immediate past, her mental representation of the time of the event is a discrete probability distribution with all probability mass on the immediate past. The speaker then attempts to communicate this event to the listener, using the English past tense "I went". The listener, having access only to this linguistic form, must *mentally reconstruct* when the event took place. Because the speaker used a broad past tense category, the listener's reconstruction of the time of the event is necessarily uncertain. Concretely, the listener has no way of knowing whether the event took place in the immediate, recent or remote past, because the English past tense category does not make such fine distinctions. We represent this uncertainty in the listener's mind as probability masses over these possible points in the past, that sum to 1. We take the informativeness of communication to be the extent to which the listener's reconstruction closely approximates the speaker's intended message.

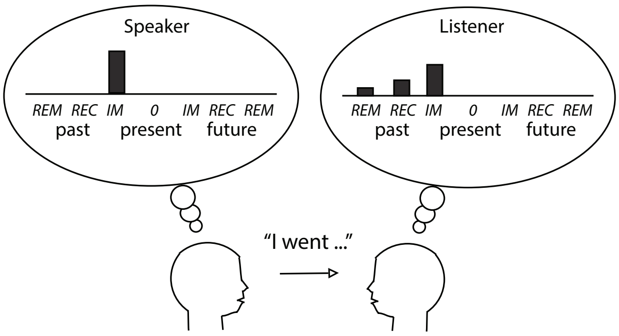

Figure 1: A communicative scenario about time.

We have seen that the tense system of English is relatively coarse and leads to temporal uncertainty. In contrast, a tense system like that of Kikuyu is more precise, because of the fine grained distinctions it makes in its past tense categories. However, Kikuyu's system is also less simple than English's system, by virtue of having these additional categories. Thus there is a tradeoff between a preference for informativeness and one for simplicity. We ask whether the tense systems we find in the world's languages reflect an optimal tradeoff between these two preferences.

In what follows, we first describe the cross-linguistic data that we use for our analysis. We then present the theory of efficient communication in formal terms, building on the informal sketch given above. Finally, we test our theory against the data.

Data

We used data from Dahl (1985), the most comprehensive cross-linguistic survey of tense, mood and aspect systems currently available. These data represent a diverse genetic and geographic distribution against which to test our hypothesis. Of the 64 languages in the sample, the most well-represented families are Indo-European (21 languages), Afro-Asiatic (8), Niger-Congo (8) and Austronesian (7). The remaining 17 languages are well-spread, with at least two languages from each inhabited continent.

For all languages in the sample except Latin, Dahl (1985) uses primary data collected through a questionnaire designed specifically for the survey. Each speaker was presented with 197 standardized sentences in English, with accompanying linguistic and extralinguisic context, and asked to translate them into the target language. Dahl coded the responses for language-specific categories, then classified those categories into major cross-linguistic categories on the basis of similarity of distribution. It is possible that some subtle differences between languages may not have been fully captured in these cross-language categories, but for this initial test of our hypothesis we took Dahl's coding into cross-language tense categories as definitive. The categories we consider as tense are PAST, PRESENT, FUTURE, and finer-grained subdivisions of those expressing degree of remoteness, as in our discretized timeline above. In our initial analyses reported here, we restrict attention to absolute tense and do not consider relative tense or aspect, which we leave for future work.

Dahl's classification of tense systems displays three broad classes. The first class consists of tenseless systems like Cebuano discussed above, in which no tense category is expressed. The second class uses absolute tense, in which events being communicated are temporally located with respect to the present, but without expressing degree of remoteness. The third and final class are systems that encode both absolute tense and degree of remoteness. Degrees of remoteness encode a magnitude associated with the temporal location of events, as explained in the Kikuyu example above. Dahl's data present a maximally three-way distinction in degrees of remoteness: *immediate*, *recent*, and *remote*. Note that languages are not consistent in the precise meanings of *immediate*, *recent* and *remote* past and future. *Recent* past for one language may specify up to a week ago, while for another it may specify up to a month ago. However, cross-linguistic tendencies do exist, with the distinction between *immediate* and *recent* past most commonly specifying 'today' and 'before today' (Comrie, 1985: 87). On this basis we chose to define *immediate* as occurring today. Another common tendency is for languages to distinguish between 'a few days ago' and 'more than a few days ago' (Comrie, 1985: 88). On this basis we chose to define *remote* as occurring more than a week from today, with *recent* categories sitting between *immediate* and *remote* categories.

The three classes are summarized in Table 1, along with example languages. The numbers in parentheses represent the number of languages in that qualitative class with the same *number* of categories, but not the same *categories*. For example, within the class of absolute tense systems are 22 lan-

[1] In theory, the time line should be continuous. However, grammatical tense systems never treat time in such detail, so we discretize it into intervals to account for the most fine-grained representation available in the tense systems that we analyze.

guages with systems of two categories. However, these may be any combination of PAST, PRESENT and FUTURE.

Table 1: The three qualitative classes of tense systems in Dahl (1985). Parentheses indicate multiple languages that have the same number of categories within a class.

Class	# of categories	Language (total #)
Tenseless	1	Cebuano (1)
Absolute	1	Hawaiian (3)
	2	Maltese (22)
	3	English (33)
Absolute and remoteness	4	Zulu (3)
	5	Sotho (2)

Formal presentation of theory

The notion of efficient communication involves two competing forces: informativeness and simplicity. We describe each of these in turn in the specific case of tense systems, building on the informal presentation above. Our presentation here follows that of Kemp and Regier (2012) and Regier et al. (in press).

Informativeness

We assume a communicative scenario such as that depicted in Figure 1, in which a speaker is communicating with a listener. As in that figure, we assume that the shared mental representation of the time line consists of seven ordered temporal locations, which we denote as: *-REM* (remote past), *-REC* (recent past), *-IM* (immediate past), t_0 or 0 (present), *+IM* (immediate future), *+REC* (recent or intermediate future), and *+REM* (remote future). We model the speaker's and listener's mental representations as probability distributions, $S(\cdot)$ and $L(\cdot)$ respectively, over these temporal locations. We assume that the speaker wishes to communicate an event that occurred at a particular temporal location i (e.g. -IM: immediate past), and that the speaker is certain of this location: $S(i) = 1$ and $S(j) = 0, \forall j \neq i$. In order to convey this location, the speaker produces an utterance (e.g. "I went") that is marked for the tense category (here, PAST for English) in which the target location falls. The listener then attempts to reconstruct the speaker's intended meaning, creating a mental representation $L(\cdot)$ based on the tense category c used by the speaker:

$$L(i) \propto f(i|c) \tag{1}$$

We assume that $f(i|c)$ is determined by how mentally accessible each temporal location i within the category c is. Previous work on the mental representation of time has suggested that in general, "recent items ... are more retrievable than distant items" (Brown et al., 2007: 541). For this reason we distribute mass within the category c according to the similarity of each item in the category to the present (t_0):

$$f(i|c) = \begin{cases} sim(i,t_0) & \text{if } i \in c \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Following Brown et al. (2007: 544), we also assume that the psychological similarity between any two temporal locations i and j is an exponentially decaying function of temporal distance between them:

$$sim(i,j) = e^{-dist(i,j)} \tag{3}$$

Finally, we assume that the mental distance $dist(\cdot,\cdot)$ between any two neighboring temporal locations on our idealized 7-location mental time line is 1. Given these assumptions, the listener reconstruction for the English PAST category would assign the most mass to -IM (immediate past), less to -REC (recent past), and still less to -REM (remote past), as in Figure 1 above.

Given these definitions of the speaker S and listener L distributions, we define the *communicative cost* of communicating a mental object i under a given semantic system to be the Kullback-Leibler divergence between S and L. Intuitively, this is the amount of information that is lost when using L to approximate S. In the case of speaker certainty as assumed here, this quantity reduces to surprisal:

$$C(i) = D_{KL}(S||L) = \sum_j S(j)\log_2 \frac{S(j)}{L(j)} = \log_2 \frac{1}{L(i)} \tag{4}$$

We then define the communicative cost of a tense system as a whole as the expected communicative cost it incurs over all seven temporal locations on the discretized time line:

$$E[C] = \sum_{i=1}^{7} C(i)N(i) \tag{5}$$

Here $N(i)$ is the *need probability* for location i; that is, the probability that the speaker will need to refer to location i rather than any other temporal location. We estimated these probabilities using data from the Google Ngram English corpus (Michel et al., 2011) for the year of publication of Dahl's book: 1985. This involved two steps. First, we found the 10 most common verbs according to the Corpus of Contemporary American English (Davies, 2008-): *be*, *have*, *do*, *say*, *go*, *think*, *know*, *want*, *get* and *make*, which account for over 50% of verb tokens in a 17,000 sentence spoken corpus (Ota, 1963). We conjugated each of these verbs to express present, past or future tense. For instance, *be* becomes *am*, *are* and *is* for PRESENT, *will be* and *shall be* for FUTURE, *was* and *were* for PAST. We then individually searched for these conjugated verb forms in the corpus, and summed the frequencies to obtain aggregated frequencies for the coarse categories past, present and future. Second, we used frequencies of specific temporal adverbs to approximate the fine-grained *remote*, *recent* and *immediate* categories for PAST and FUTURE. The specific temporal adverbs we searched for are shown in Table 2. Since both the immediate past and immediate future are expressed through *today* in English, we assigned half of *today*'s frequency to each of the two stimuli. We used this second set

Degree of remoteness	Temporal adverb
IMMEDIATE PAST/FUT.	*today*
PAST RECENT	*yesterday*
FUTURE RECENT	*tomorrow*
PAST REMOTE	*last week/month/year/decade/century*
FUTURE REMOTE	*next week/month/year/decade/century*

Table 2: Temporal adverbs used to estimate the need probabilities for varying degrees of remoteness.

of frequencies to distribute probability mass within past and future categories.

The resulting need probabilities are shown in Figure 2, and follow the rank order *present* > *past* > *future*. We confirmed this rank order in an independent corpus of spoken English (Du Bois et al., 2000-2005) by randomly sampling 100 sentences and categorizing them into *present*, *past*, and *future* based on conjugated verbs in these sentences.

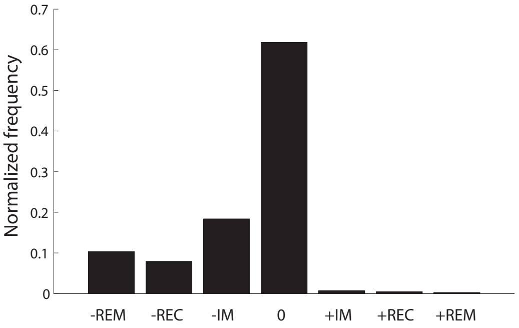

Figure 2: Need probabilities of 7 temporal locations.

Given these definitions and quantities, we take a semantic system to be informative to the extent that it exhibits low communicative cost $E[C]$, as defined in Equation 5.

Simplicity

Simplicity is the opposite of complexity, and we take the complexity of a tense system to be the number of grammatical categories in it—whether marked morphologically or periphrastically, as coded by Dahl (1985). For example, English has morphologically marked categories for PAST and PRESENT, and a periphrastic category for FUTURE, so it has a total complexity of 3. For those systems that do not include all seven temporal locations within their tense categories, we added a null category that groups together the otherwise uncategorized temporal locations.

Procedure and results

We tested the proposal of efficient communication by comparing tense systems from Dahl (1985) to hypothetical systems that partition the seven temporal locations of the idealized time line in all possible ways. We considered an attested system to be communicatively efficient to the extent that it is more informative (has lower communicative cost) than most hypothetical systems of the same complexity.

Figure 3a-b summarizes the results. The two axes of panel (a) are complexity and communicative cost. Gray dots denote hypothetical systems, and colored circles denote attested systems. It can be seen that most attested systems are near-optimally efficient, in that they exhibit near-minimal cost (near-maximal informativeness) for their level of complexity–with some exceptions. For the tenseless class, there is only one hypothetical system, hence this system is necessarily and trivially most informative. The class of tense systems without degrees of remoteness (shown in blue) is near-optimally informative when compared with hypothetical systems of matching complexity ($p < 10^{-15}$ using Fisher's method). However, within this class, Greenlandic Eskimo is clearly not efficient. The class of tense systems with degrees of remoteness (shown in red) is also near-optimally informative when taken as a whole ($p < 0.001$), although Zulu is further away from the minimal cost system than other languages in this class.

Why are most languages efficient on this analysis, and a few languages not? The distribution of need probabilities shown in Figure 2 suggests an answer. Past and present locations have high need probability, therefore any information loss concerning those temporal locations is heavily weighted in Equation 5. Information loss results from broad, uninformative categories; in consequence, categories in the past and present are under especially great pressure to minimize information loss by being semantically precise or narrow. To the extent that this usage pattern appears across languages, it helps to explain why languages are more likely to subdivide PAST than FUTURE into finer-grained categories (Comrie, 1985: 85).[2] Most of the languages in the sample we tested specify past and present tenses–but Greenlandic Eskimo does not and is penalized for it. Figure 3c confirms this line of reasoning by showing the theoretically optimal tense systems: those systems that exhibit minimum cost at different complexities. Note that at complexity $k = 2$, the optimal system is one that assigns a category to present, and a second category to the remainder of the time line, reflecting the importance of present in contributing to communicative efficiency. This optimal system is attested in Hawaiian, as shown in panels (a) and (b).

Why then are there languages that appear inefficient on this analysis? One possibility is that our theory is simply inadequate, but there are also other possibilities. In some instances, there appears to be a discrepancy between Dahl's coding and other reports in the literature. For example, Dahl codes Mandarin Chinese as having the same suboptimal tense system as Greenlandic Eskimo, but other works have suggested that Mandarin Chinese is tenseless (Lin, 2012), which would render it (trivially) optimal, like Cebuano. Another possibility is that tense and aspect are inseparable dimensions of temporal

[2] Regier and Kemp (2012) used analogous reasoning to explain markedness asymmetries in kinship terminologies.

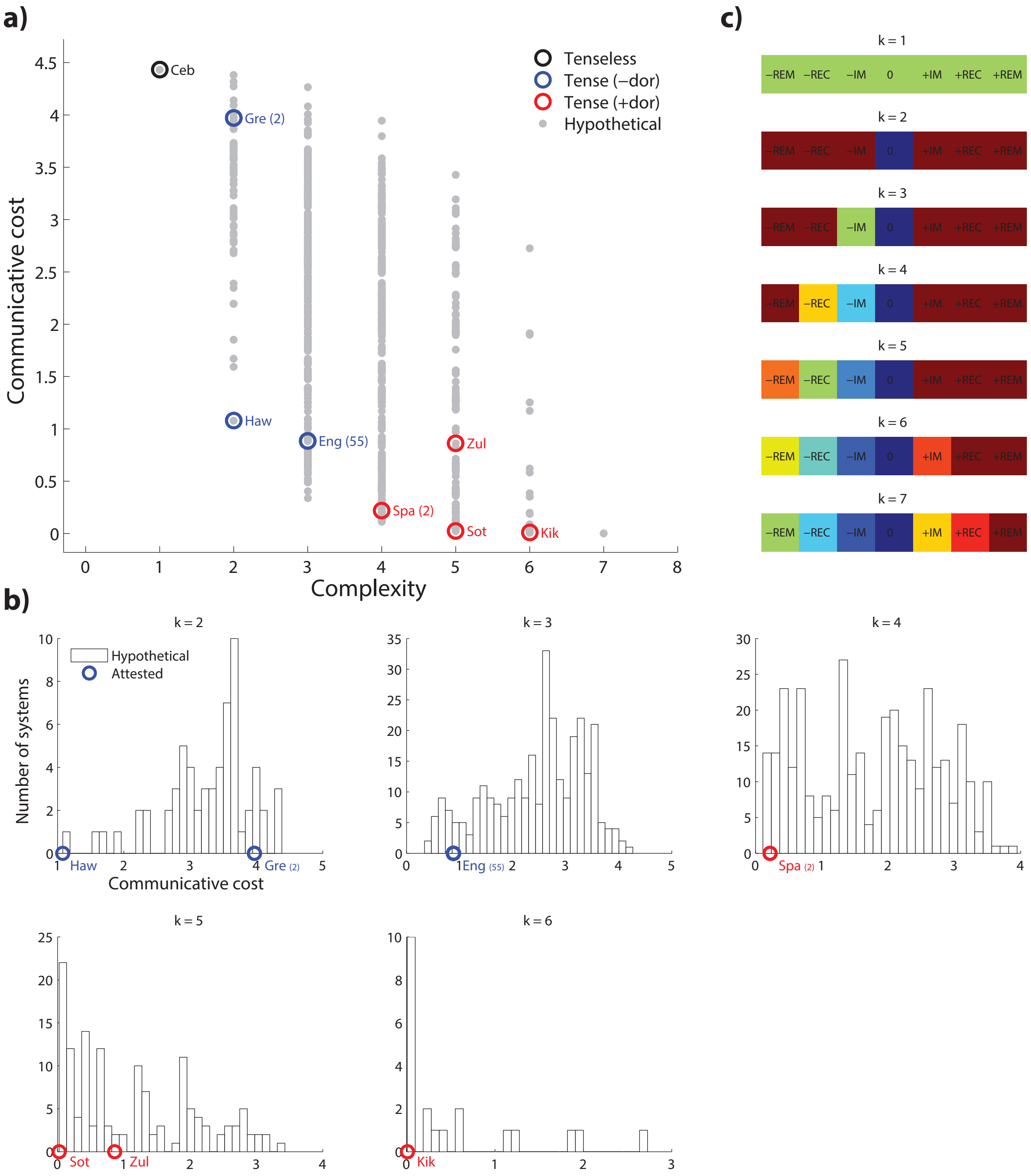

Figure 3: Efficiency analyses of tense systems. a) Near-optimal tradeoff between communicative cost and complexity. Attested languages are circled with 3-letter abbreviations and correspond to: Cebuano, Hawaiian, Greenlandic Eskimo, English, Spanish, Sotho, Zulu and Kikuyu; parentheses indicate multiple languages that have identical categorizations of the time line. "-dor" and "+dor" correspond to tense systems without and with degrees of remoteness respectively. b) Densities of hypothetical systems juxtaposed with attested systems of equal complexities. c) Theoretically optimal systems at different complexities. Categories are indicated by different colors.

situations (Binnick, 2012), and that some languages appear inefficient only because we are considering an isolated part of this larger system of temporal reference. A final possibility is that need probabilities may vary substantially across cultures.

Calculating need probabilities on a per-language basis could change the efficiency assessment of many languages–either toward greater efficiency, or away from it. Exploring these possibilities is a topic for future research.

Conclusion

We have presented evidence that tense systems across languages support efficient communication, and that this principle may explain cross-language variation in tense systems. Notably, our analysis has the potential to explain the tendency of languages to have finer-grained categories in the past than in the future (Comrie, 1985). Our present findings theoretically align the study of tense with existing work that explains cross-language semantic variation in the domains of color, kinship, space and number in terms of the same principles.

We find that a small number of tense systems are not communicatively efficient. It is presently unclear whether these systems represent a challenge to the central principles of our theory, merely reflect less critical implementational details, or highlight the need for the same theoretical notions to be applied in a more language-specific and culture-specific way. Further research will be needed to settle these matters. For now, however, our initial analyses suggest a communicative basis for the ways in which tense systems vary across languages, and also suggest ways to further explore and refine that idea.

Acknowledgments

We thank Östen Dahl for making his book publicly available, Charles Kemp for pointers to the data source, and Alex Carstensen for discussions of need probabilities. Work on this project was funded by NSF award SBE-1041707 to the Spatial Intelligence and Learning Center (SILC).

References

Binnick, R. (Ed.). (2012). *The Oxford Handbook of Tense and Aspect*. New York: Oxford University Press.

Brown, G., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539-576.

Bybee, J., & Dahl, Ö. (1989). The Creation of Tense and Aspect Systems in the Languages of the World. *Studies in Languages*, *13*(1), 51-103.

Bybee, J., Perkins, R., & Pagliuca, W. (1994). *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: University of Chicago Press.

Comrie, B. (1985). *Tense*. Cambridge: Cambridge University Press.

Dahl, Ö. (1985). *Tense and Aspect systems*. Oxford: Basil Blackwell.

Davies, M. (2008-). *The Corpus of Contemporary American English: 450 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/.

Du Bois, J., Chafe, W., Meyer, C., Thompson, S., Englebretson, R., & Martey, N. (2000-2005). *Santa Barbara corpus of spoken American English, Parts 1-4*. Philadelphia: Linguistic Data Consortium.

Fedzechkina, M., Jaeger, T., & Newport, E. (2012). Language learners restructure their input to facilitate efficient communication. *PNAS*, *109*(44), 17897-17902.

Hornstein, N. (1993). *As Time Goes By: Tense and Universal Grammar*. Cambridge, MA: MIT Press.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*, 1049–1054.

Khetarpal, N., Neveu, G., Majid, A., Michael, L., & Regier, T. (2013). Spatial terms across languages support near-optimal communication: Evidence from Peruvian Amazonia, and computational analyses. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society.*

Klein, W. (2009). How time is encoded. In W. Klein & P. Li (Eds.), *The Expression of Time.* Berlin: De Gruyter.

Klein, W., & Li, P. (Eds.). (2009). *The Expression of Time*. Berlin: De Gruyter.

Lin, J.-W. (2012). Tenselessness. In R. Binnick (Ed.), *The Oxford Handbook of Tense and Aspect.* New York: Oxford University Press.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*, 176–182.

Ota, A. (1963). *Tense and aspect of present-day American English*. Tokyo: Kenkyusha.

Piantadosi, S., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *PNAS*, *108*(9), 3526-3529.

Quine, W. (1960). *Word and Object*. Cambridge, MA: MIT Press.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *PNAS*, *104*, 1436–1441.

Regier, T., Kemp, C., & Kay, P. (in press). Word meanings across languages support efficient communication. In B. MacWhinney & W. O'Grady (Eds.), *The Handbook of Language Emergence.* Hoboken, NJ: Wiley.

Smith, K., Tamariz, M., & Kirby, S. (2013). Linguistic structure is an evolutionary trade-off between simplicity and expressivity. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society.*

Xu, Y., & Regier, T. (2014). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. In P. Bello et al. (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society.*