Supplementary Information for: Crosslinguistic Word Orders Enable an Efficient Tradeoff between Memory and Surprisal

Michael Hahn, Judith Degen, Richard Futrell

2018

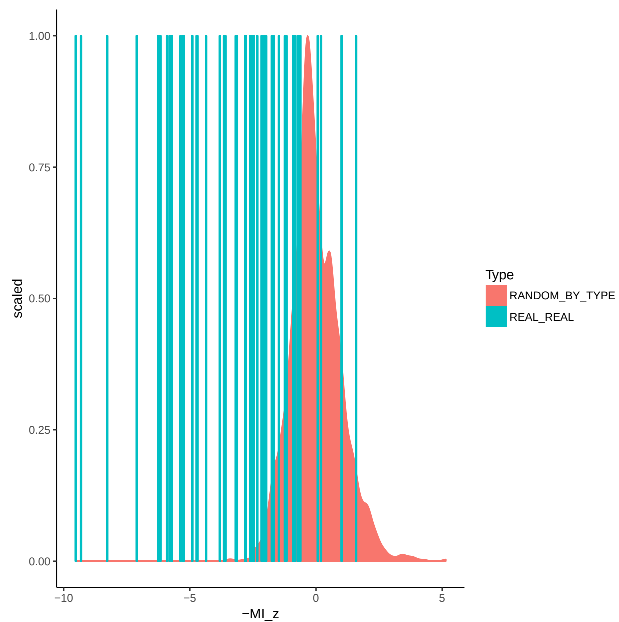


Figure 1: Histogram

{fig:hist-re

1 Formal Analysis and Proofs

In this section, we prove the theorem described above.

1.1 Mathematical Assumptions

We first make explicit how we formalize language processing for proving the theorem.

Ingredient 1: Language as a Stationary Stochastic Process We represent language as a stochastic process of words $\dots w_{-2} w_{-1} w_0 w_1 w_2 \dots$, extending indefinitely both into the past and into the future. The symbols w_i belong to a common set, representing the words of the language.[1]

[1]Could also be phonemes, sentences, ..., any other kind of unit.

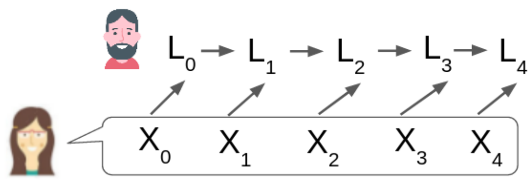


Figure 2: Illustration of (4). As the utterance unfolds, the listener maintains a memory state. After receiving word w_t, the listener computes their new memory state m_t based on the previous memory state m_{t-1} and the new word w_t. {fig:listene

The assumption of infinite length is for mathematical convenience and does not affect the substance of our results: As we restrict our attention to the processing of individual sentences, which have finite length, we will actually not make use of long-range and infinite contexts.

We make the assumption that this process is *stationary*. Formally, this means that the conditional distribution $P(w_t|w_{<t})$ does not depend on t, it only depends on the actual sequence $w_{<t}$. Informally, this says that the process has no 'internal clock', and that the statistical rules of the language do not change at the timescale we are interested in. In reality, the statistical rules of language do change: They change as language changes over generations, and they also change between different situations – e.g., depending on the interlocutor at a given point in time. Given that we are interested in memory needs in the processing of *individual sentences*, at a timescale of seconds or minutes, stationarity seems to be a reasonable assumption to make.

Ingredient 2: Flow of Information There are no assumptions about the memory architecture and the nature of its computations. We only make a basic assumption about the flow of information (Figure 2): At a given point in time, the listener's memory state m_t is determined by the last word w_t, and the prior memory state m_{t-1}:

$$m_t = M(m_{t-1}, w_t) \tag{1}$$

As a consequence, m_t contains no information about the process beyond what is contained in the last word observed w_{t-1} and in the memory state before that word was observed m_{t-1}. As a consequence, the listener has no knowledge of the speaker's state beyond the information provided in their prior communication. This is a simplification, as the listener could obtain information about the speaker from other sources, such as their common environment (weather, ...). (For the study of memory in sentence processing, this seems fair. Discuss this more.)

1.2 Proof of the Theorem

We restate the theorem:

{prop:subopt

Theorem 1. *Let T be any positive integer ($T \in \{1, 2, 3, \ldots\}$), and consider a listener using at most*

$$\sum_{t=1}^{T} t I_t \tag{2}$$

bits of memory on average. Then this listener will incur surprisal at least

$$H[w_t|w_{<t}] + \sum_{t>T} I_t$$

on average.

We formalize a language as a stationary stochastic process $\dots w_{-2}w_{-1}w_0w_1w_2\dots$, extending indefinitely both into the past and into the future. The symbols w_i belong to a common set, representing the words of the language.[2] We denote the listener's memory state at time t, after hearing $w_{<t} = ...w_{t-2}w_{t-1}$ by m_t. As described above, we assume

$$m_t = M(m_{t-1}, w_{t-1}) \tag{3}$$

[3] As a consequence, the listener has no knowledge of the speaker's state beyond the information provided in their prior communication.

The average number of bits required to encode this state is $\mathrm{H}[m_t]$, which by assumption is at most $\sum_{t=1}^{T} tI_t$. As the listener's predictions are made on the basis of her memory state, her average surprisal is at least $\mathrm{H}[w_t|m_t]$. The difference between the listener's surprisal and optimal surprisal is thus at least $\mathrm{H}[w_t|m_t] - \mathrm{H}[w_t|w_{<t}]$. By the assumption of stationarity, we can, for any positive integer T, rewrite this expression as

$$\mathrm{H}[w_t|m_t] - \mathrm{H}[w_t|w_{<t}] = \frac{1}{T}\sum_{t'=1}^{T} \left(\mathrm{H}[w_{t'}|m_{t'}] - \mathrm{H}[w_{t'}|w_{<t'}]\right) \tag{5}$$

Because m_t is determined by $(w_{1\dots t-1}, m_1)$:

$$m_t = M(m_{t-1}, w_{t-1}) = M(M(m_{t-2}, w_{t-2}), w_{t-1}) = M(M(M(m_{t-3}, w_{t-3}), w_{t-2}), w_{t-1}) = \dots \tag{6}$$

the Data Processing inequality entails the following inequality for every positive integer t:

$$H[w_t|m_t] \geq H[w_t|w_{1\dots t-1}, m_1] \tag{7}$$

Plugging this inequality into Equation 5 above:

$$\mathrm{H}[w_t|m_t] - \mathrm{H}[w_t|w_{<t}] \geq \frac{1}{T}\sum_{t=1}^{T}(\mathrm{H}[w_t|w_{1\dots t-1}, m_1] - \mathrm{H}[w_t|w_{1\dots t-1}, w_{\leq 0}]) \tag{8}$$

$$= \frac{1}{T}\left(\mathrm{H}[w_{1\dots T}|m_1] - \mathrm{H}[w_{1\dots T}|w_{\leq 0}]\right) \tag{9}$$

$$= \frac{1}{T}\left(I[w_{1\dots T}, w_{\leq 0}] - I[w_{1\dots T}, m_1]\right) \tag{10}$$

The first term $I[w_{1\dots T}, w_{\leq 0}]$ can be rewritten in terms of I_t:

$$I[w_{1\dots T}, w_{\leq 0}] = \sum_{i=1}^{T}\sum_{j=-1}^{-\infty} I[w_i, w_j|w_{j+1}...w_{i-1}] = \sum_{t=1}^{T} tI_t + T\sum_{t>T} I_t \tag{11}$$

Therefore

$$\mathrm{H}[w_t|m_t] - \mathrm{H}[w_t|w_{<t}] \geq \frac{1}{T}\left(\sum_{t=1}^{T} tI_t + T\sum_{t>T} I_t - I[w_{1\dots T}, m_1]\right)$$

[2] Could also be phonemes, sentences, ..., any other kind of unit.

[3] Alternatively we could admit nondeterministic memory encodings, and require

$$p(m_{t+1}|(w_{t'})_{t'\in\mathbb{Z}}, m_t) = p(m_{t+1}|m_t, w_t) \tag{4}$$

that is, m_{t+1} contains no information about the utterances beyond what is contained in m_t and w_t.

$I[w_{1\ldots T}|m_1]$ is at most $\mathrm{H}[m_1]$, which is at most $\sum_{t=1}^{T} tI_t$ by assumption. Thus, the expression above is bounded by

$$\begin{aligned}\mathrm{H}[w_t|m_t] - \mathrm{H}[w_t|w_{<t}] &\geq \frac{1}{T}\left(\sum_{t=1}^{T} tI_t + T\sum_{t>T} I_t - \sum_{t=1}^{T} tI_t\right) \\ &= \sum_{t>T} I_t\end{aligned}$$

Rearranging shows that the listener's surprisal is at least $\mathrm{H}[w_t|m_t] \geq \mathrm{H}[w_t|w_{<t}] + \sum_{t>T} I_t$, as claimed.

1.3 Locality in a model with Memory Retrieval

Here we show that our information-theoretic analysis is compatible with models placing the main bottleneck in the difficulty of retrieval (McElree, 2000; Lewis and Vasishth, 2005; Nicenboim and Vasishth, 2018; Vasishth et al., 2019). We extend our model of memory in incremental prediction to capture key aspects of the models described by Lewis and Vasishth (2005); Nicenboim and Vasishth (2018); Vasishth et al. (2019).

The ACT-R model of Lewis and Vasishth (2005) assumes a small working memory consisting of *buffers* and a *control state*, which together hold a small and fixed number of individual *chunks*. It also assumes a large short-term memory that contains an unbounded number of chunks. This large memory store is accessed via *cue-based retrieval*: a query is constructed based on the current state of the buffers and the control state; a chunk that matches this query is then selected from the memory storage and placed into one of the buffers.

Formal Model We extend our information-theoretic analysis by considering a model that maintains both a small working memory m_t – corresponding to the buffers and the control state – and an unlimited short-term memory s_t. Predictions are made based on working memory m_t, incurring surprisal $H[w_t|m_t]$. When processing a word x_t, there is some amount of communication between m_t and s_t, corresponding to retrieval operations. We model this using a variable r_t representing the information that is retrieved from s_t. In our formalization, r_t reflects the totality of all retrieval operations that are made during the processing of x_{t-1}; they happen after x_{t-1} has been observed but before x_t has.

The working memory state is determined not just by the input x_t and the previous working memory state m_{t-1}, but also by the retrieved information:

$$m_t = f(x_t, m_{t-1}, r_t) \tag{12}$$

The retrieval operation is jointly determined by working memory, short-term memory, and the previous word:

$$r_t = g(x_{t-1}, m_{t-1}, s_{t-1}) \tag{13}$$

Finally, the short-term memory can incorporate any – possibly all – information from the last word and the working memory:

$$s_t = h(x_{t-1}, m_{t-1}, s_{t-1}) \tag{14}$$

While s_t is unconstrained, there are constraints on the capacity of working memory $\mathrm{H}[m_t]$ and the amount of retrieved information $\mathrm{H}[r_t]$. Placing a bound on $\mathrm{H}[m_t]$ reflects the fact that the buffers can only hold a small and fixed number of chunks (Lewis and Vasishth, 2005).

Cost of Retrieval In the model of Lewis and Vasishth (2005), the time it takes to process a word is determined primarily by the time spent retrieving chunks, which is determined by the number of retrieval operations and the time it takes to complete each retrieval operation. If the information content of each chunk is bounded, then a bound on $H[r_t]$ corresponds to a bound on the number of retrieval operations.

In the model of Lewis and Vasishth (2005), a retrieval operation takes longer if more chunks are similar to the retrieval cue, whereas, in the direct-access model (McElree, 2000; Nicenboim and Vasishth, 2018; Vasishth et al., 2019), retrieval operations take a constant amount of time. There is no direct counterpart to differences in retrieval times and similarity-based inhibition as in the activation-based model in our formalization. Our formalization thus more closely matches the direct-access model, though it might be possible to incorporate aspects of the activation-based model in our formalization.

Role of Surprisal The ACT-R model of Lewis and Vasishth (2005) does not have an explicit surprisal cost. Instead, surprisal effects are interpreted as arising because, in less constraining contexts, the parser is more likely to make decisions that then turn out to be incorrect, leading to additional correcting steps. We view this as an algorithmic-level implementation of a surprisal cost $H[x_t|m_{t-1}]$: If the word x_t is unexpected given the current state of the working memory – i.e., buffers and control states – then their current state must provide insufficient information to constrain the actual syntactic state of the sentence, meaning that the parsing steps made to integrate x_t are likely to include more backtracking and correction steps. Thus, we argue that cue-based retrieval models predict that the surprisal $-\log P(x_t|m_{t-1})$ will be part of the cost of processing word x_t.

Theoretical Result We now show an extension of our theoretical result in the setting of the retrieval-based model described above.

Theorem 2. *Let $0 < S \leq T$ be positive integers such that the average working memory cost $\mathrm{H}[m_t]$ is bounded as*

$$\mathrm{H}[m_t] \leq \sum_{t=1}^{T} tI_t \tag{15}$$

and the average amount of retrieved information is bounded as

$$\mathrm{H}[r_t] \leq \sum_{t=T+1}^{S} I_t \tag{16}$$

Then the surprisal cost is lower-bounded as

$$\mathrm{H}[w_t|m_t] \geq \mathrm{H}[w_t|x_{<t}] + \sum_{t>S} I_t \tag{17}$$

Proof. The proof is a generalization of the proof above. For any positive integer t, m_t is determined by $w_{1\ldots t}, m_0, r_0, \ldots, r_t$. Therefore, the Data Processing Inequality entails:

$$\mathrm{H}[w_t|m_t] \geq \mathrm{H}[w_t|w_{1\ldots t}, m_0, r_0, \ldots, r_t] \tag{18}$$

As in (8), this leads to

$$\mathrm{H}[w_t|m_t] - \mathrm{H}[w_t|w_{<t}] \geq \frac{1}{T}\sum_{t=1}^{T}(\mathrm{H}[w_t|w_{1\ldots t}, m_0, r_0, \ldots, r_t] - \mathrm{H}[w_t|w_{1\ldots t-1}, w_{\leq 0}]) \tag{19}$$

$$\geq \frac{1}{T}\left(\mathrm{H}[w_{1\ldots T}|m_0, r_0, \ldots, r_T] - \mathrm{H}[w_{1\ldots T}|w_{\leq 0}]\right) \tag{20}$$

$$= \frac{1}{T}\left(I[w_{1\ldots T}, w_{\leq 0}] - I[w_{1\ldots T}, (m_0, r_0, \ldots, r_T)]\right) \tag{21}$$

Now, using the calculation from (11), this can be rewritten as:

$$\begin{aligned}\mathrm{H}[w_t|m_t] - \mathrm{H}[w_t|w_{<t}] =& \frac{1}{T}\left(\sum_{t=1}^{T} tI_t + T\sum_{t>T} I_t - I[X_1 \ldots X_T, (M_0, R_1, ..., R_T)]\right) \\ =& \frac{1}{T}\left(\sum_{t=1}^{T} tI_t + T\sum_{t>T} I_t - I[X_{1\ldots T}, M_0] - \sum_{t=1}^{T} I[X_{1\ldots T}, R_t|M_0, r_{1\ldots t-1}]\right)\end{aligned}$$

Due to the inequalities $I[X_{1\ldots T}, M_0] \leq \mathrm{H}[M_0]$ and $\mathrm{I}[X_{1\ldots T}, R_t|M_0, r_{1\ldots t-1}] \leq \mathrm{H}[R_t]$, this can be bounded as

$$\mathrm{H}[w_t|m_t] - \mathrm{H}[w_t|w_{<t}] \geq \frac{1}{T}\left(\sum_{t=1}^{T} tI_t + T\sum_{t>T} I_t - H[M_0] - \sum_{t=1}^{T} H[R_t]\right) \tag{22}$$

$$\tag{23}$$

Finally, this reduces as

$$\mathrm{H}[w_t|m_t] - \mathrm{H}[w_t|w_{<t}] \geq \frac{1}{T}(T\sum_{t>T} I_t - T \cdot H[R_t]) \tag{24}$$

$$= \sum_{t>T} I_t - H[R_t] \tag{25}$$

$$\geq \sum_{t>T} I_t - \sum_{t=T+1}^{S} I_t \tag{26}$$

$$= \sum_{t>S} I_t \tag{27}$$

□

Information Locality We now show that this result predicts information locality provided that retrieving information is more expensive than keeping the same amount of information in working memory. For this, we formalize the problem of finding an optimal memory strategy as a multi-objective optimization, aiming to minimize

$$\lambda_1 H[m_t] + \lambda_2 H[r_t] \tag{28}$$

to achieve a given surprisal level, for some setting of $\lambda_1, \lambda_2 > 0$ describing the relative cost of storage and retrieval. What is the optimal division of labor between keeping information in working memory and recovering it through retrieval? The problem

$$\min_T \lambda_1 \sum_{t=1}^{T} tI_t + \lambda_2 \sum_{t=T+1}^{S} I_t \tag{29}$$

has solution $T \approx \frac{\lambda_2}{\lambda_1}$. This means that, as long as retrievals are more expensive than keeping the same amount of information in working memory (i.e., $\lambda_2 > \lambda_1$), the optimal strategy stores information from the last $T > 1$ words in working memory. Due to the factor t inside $\sum_{t=1}^{T} tI_t$, the bound (29) will be reduced when I_t decays faster, i.e., there is strong information locality.

The assumption that retrieving information is more difficult than storing it is reasonable for cue-based retrieval models, as retrieval suffers from similarity-based interference effects due to the unstructured nature of the storage (Lewis and Vasishth, 2005). A model that maintains no information in its working memory, i.e. $H[m_t] = 0$, would correspond to a cue-based retrieval model that stores nothing in its buffers and control states, and relies entirely on retrieval to access past information. Given the nature of representations assumed in models (Lewis and Vasishth, 2005), such a model would seem to be severely restricted in its ability to parse language.

1.4 Results for Language Production

Here we show results linking memory and locality in production.

First, we consider a setting in which a speaker produces sentences with bounded memory, and analyze the deviation of the produced distribution from the actual distribution of the language.

The first setting does not account for the fact that language is produced aiming for some communicative goal. We therefore now assume that the speaker has a communicative goal G in mind. This goal G stays constant during production process for a sentence, and we count how much memory is needed in addition to the goal G. We assume that there is a distribution of sentences expressing goals G:

$$P(sentence|G) \tag{30}$$

and assume that the speaker aims to match this distribution

$$\mathbb{E}_G[D_{KL}((language|G)||(produced|G))] \tag{31}$$

We can analyze this model by adding conditioning w.r.t. G throughout the analysis of the previous case. Specifically, we need $I_t^G := I[X_t, X_0|X_1, \dots, X_{t-1}, G]$.

Take I_t conditioned on G: only count statistical dependencies to the degree that they are not redundant with the goal

2 Example where window model is not optimal

Here we provide an example of a stochastic process where a window-based memory encoding is not optimal, but the bound provided by our theorem still holds.

Let k be some positive integer. Consider a process $x_{t+1} = (v_{t+1}, w_{t+1}, y_{t+1}, z_{t+1})$ where

1. The first two components consist of fresh random bits. Formally, v_{t+1} is an independent draw from $Bernoulli(0.5)$, independent from all preceding observations $x_{\leq t}$. Second, let w_{t+1} consist of $2k$ many such independent random bits (so that $H[w_{t+1}] = 2k$)

2. The third component *deterministically* copies the first bit from $2k$ steps earlier. Formally, y_{t+1} is equal to the first component of x_{t-2k+1}

3. The fourth component *stochastically* copies the second part (consisting of $2k$ random bits) from one step earlier. Formally, each component $z_{t+1}^{(i)}$ is determined as follows: First take a sample $u_{t+1}^{(i)}$ from $Bernoulli(\frac{1}{4k})$, independent from all preceding observations. If $u_{z+1}^{(i)} = 1$, set $z_{t+1}^{(i)}$ to be equal to the second component of $w_t^{(i)}$. Otherwise, let $z_{t+1}^{(i)}$ be a fresh draw from $Bernoulli(0.5)$.

Predicting observations optimally requires taking into account observations from the $2k$ last time steps.

We show that, when approximately predicting with low memory capacities, a window-based approach does *not* in general achieve an optimal memory-surprisal tradeoff.

Consider a model that predicts x_{t+1} from only the last observation x_t, i.e., uses a window of length one. The only relevant piece of information in this past observation is w_t, which stochastically influences z_{t+1}. Storing this costs $2k$ bit of memory as w_t consists of $2k$ draws from $Bernoulli(0.5)$. How much does it reduce the surprisal of x_{t+1}? Due to the stochastic nature of z_{t+1}, it reduces the surprisal only by about $I[x_{t+1}, w_t] = I[z_{t+1}, w_t] < 2k \cdot \frac{1}{2k} = 1$, i.e., surprisal reduction is strictly less than one bit. [4]

We show that there is an alternative model that strictly improves on this window-based model: Consider a memory encoding model that encodes each of $v_{t-2k+1}, \dots, v_t$, which costs $2k$ bits of memory – as the window-based model did. Since $y_{t+1} = v_{t-2k+1}$, this model achieves a surprisal reduction of $H[v_{t-2k+1}] = 1$ bit, strictly more than the window-based model.

This result does not contradict our theorem because the theorem only provides *bounds* across models, which are not necessarily achieved by a given window-based model. In fact, for the process described here, no memory encoding function M can exactly achieve the theoretical bound described by the theorem.

3 Corpus Size per Language

Language	Training	Held-Out	Language	Training	Held-Out
Afrikaans	1,315	194	Indonesian	4,477	559
Amharic	974	100	Italian	17,427	1,070
Arabic	21,864	2,895	Japanese	7,164	511
Armenian	514	50	Kazakh	947	100
Bambara	926	100	Korean	27,410	3,016
Basque	5,396	1,798	Kurmanji	634	100
Breton	788	100	Latvian	4,124	989
Bulgarian	8,907	1,115	Maltese	1,123	433
Buryat	808	100	Naija	848	100
Cantonese	550	100	North Sami	2,257	865
Catalan	13,123	1,709	Norwegian	29,870	4,639
Chinese	3,997	500	Persian	4,798	599
Croatian	7,689	600	Polish	6,100	1,027
Czech	102,993	11,311	Portuguese	17,995	1,770
Danish	4,383	564	Romanian	8,664	752
Dutch	18,310	1,518	Russian	52,664	7,163

[4] We can evaluate $I[z_{t+1}, w_t]$ as follows. Set $l = k/4$. Write z, w for any of the $2k$ components of z_{t+1}, w_t, respectively. First, calculate $p(z=1|w=1) = 1/l + (1-1/l)\frac{1}{2} = 1/(2l) + 1/2 = \frac{1+l}{2l}$ and $p(z=0|w=1) = (1-1/l)\frac{1}{2} = 1/2 - 1/2l = \frac{l-1}{2l}$. Then $I[Z,W] = D_{KL}(p(z|w=1)||p(z)) = \frac{1+l}{2l}\log\frac{\frac{1+l}{2l}}{1/2} + \frac{l-1}{2l}\log\frac{\frac{l-1}{2l}}{1/2} = \frac{1+l}{2l}\log\frac{1+l}{l} + \frac{l-1}{2l}\log\frac{l-1}{l} \leq \frac{1+l}{l}\log\frac{1+l}{l} = (1+1/l)\log(1+1/l) \leq (1+1/l)(1/l) = 1/l + 1/l^2 < 2/l = \frac{1}{2k}$.

English	17,062	3,070	Serbian	2,935	465
Erzya	1,450	100	Slovak	8,483	1,060
Estonian	6,959	855	Slovenian	7,532	1,817
Faroese	1,108	100	Spanish	28,492	3,054
Finnish	27,198	3,239	Swedish	7,041	1,416
French	32,347	3,232	Thai	900	100
German	13,814	799	Turkish	3,685	975
Greek	1,662	403	Ukrainian	4,506	577
Hebrew	5,241	484	Urdu	4,043	552
Hindi	13,304	1,659	Uyghur	1,656	900
Hungarian	910	441	Vietnamese	1,400	800

Table 2: Languages, with the number of training and held-out sentences available. {tab:corpora

4 Samples Drawn per Language

Language	Base.	Real	Language	Base.	Real
Afrikaans	13	10	Indonesian	11	11
Amharic	137	10	Italian	10	10
Arabic	11	10	Japanese	25	15
Armenian	140	76	Kazakh	11	10
Bambara	25	29	Korean	11	10
Basque	15	10	Kurmanji	338	61
Breton	35	14	Latvian	308	178
Bulgarian	14	10	Maltese	30	24
Buryat	26	18	Naija	214	10
Cantonese	306	32	North Sami	335	194
Catalan	11	10	Norwegian	12	10
Chinese	21	10	Persian	25	12
Croatian	30	17	Polish	309	35
Czech	18	10	Portuguese	15	55
Danish	33	17	Romanian	10	10
Dutch	27	10	Russian	20	10
English	13	11	Serbian	26	11
Erzya	846	167	Slovak	303	27
Estonian	347	101	Slovenian	297	80
Faroese	27	13	Spanish	14	10
Finnish	83	16	Swedish	31	14
French	14	11	Thai	45	19
German	19	13	Turkish	13	10
Greek	16	10	Ukrainian	28	18
Hebrew	11	10	Urdu	17	10
Hindi	11	10	Uyghur	326	175

Hungarian	220	109	Vietnamese	303	12

Figure 3: Samples drawn per language according to the precision-dependent stopping criterion. {tab:samples

Language	Mean	Lower	Upper	Language	Mean	Lower	Upper
Afrikaans	1.0	1.0	1.0	Indonesian	1.0	1.0	1.0
Amharic	1.0	1.0	1.0	Italian	1.0	1.0	1.0
Arabic	1.0	1.0	1.0	Japanese	1.0	1.0	1.0
Armenian	0.92	0.87	0.97	Kazakh	1.0	1.0	1.0
Bambara	1.0	1.0	1.0	Korean	1.0	1.0	1.0
Basque	1.0	1.0	1.0	Kurmanji	0.93	0.88	0.98
Breton	1.0	1.0	1.0	Latvian	0.49	0.4	0.57
Bulgarian	1.0	1.0	1.0	Maltese	1.0	1.0	1.0
Buryat	1.0	1.0	1.0	Naija	1.0	0.99	1.0
Cantonese	0.96	0.86	1.0	North Sami	0.37	0.3	0.44
Catalan	1.0	1.0	1.0	Norwegian	1.0	1.0	1.0
Chinese	1.0	1.0	1.0	Persian	1.0	1.0	1.0
Croatian	1.0	1.0	1.0	Polish	0.1	0.04	0.17
Czech	1.0	1.0	1.0	Portuguese	1.0	1.0	1.0
Danish	1.0	1.0	1.0	Romanian	1.0	1.0	1.0
Dutch	1.0	1.0	1.0	Russian	1.0	1.0	1.0
English	1.0	1.0	1.0	Serbian	1.0	1.0	1.0
Erzya	0.99	0.98	1.0	Slovak	0.07	0.03	0.12
Estonian	0.8	0.72	0.86	Slovenian	0.82	0.77	0.88
Faroese	1.0	1.0	1.0	Spanish	1.0	1.0	1.0
Finnish	1.0	1.0	1.0	Swedish	1.0	1.0	1.0
French	1.0	1.0	1.0	Thai	1.0	1.0	1.0
German	1.0	0.91	1.0	Turkish	1.0	1.0	1.0
Greek	1.0	1.0	1.0	Ukrainian	1.0	1.0	1.0
Hebrew	1.0	1.0	1.0	Urdu	1.0	1.0	1.0
Hindi	1.0	1.0	1.0	Uyghur	0.65	0.57	0.73
Hungarian	0.87	0.8	0.93	Vietnamese	1.0	0.98	1.0

Figure 4: Bootstrapped estimates for G. {tab:boot-g}

5 Detailed Results per Language

5.1 Median Surprisal per Memory Budget

Afrikaans Amharic Arabic Armenian

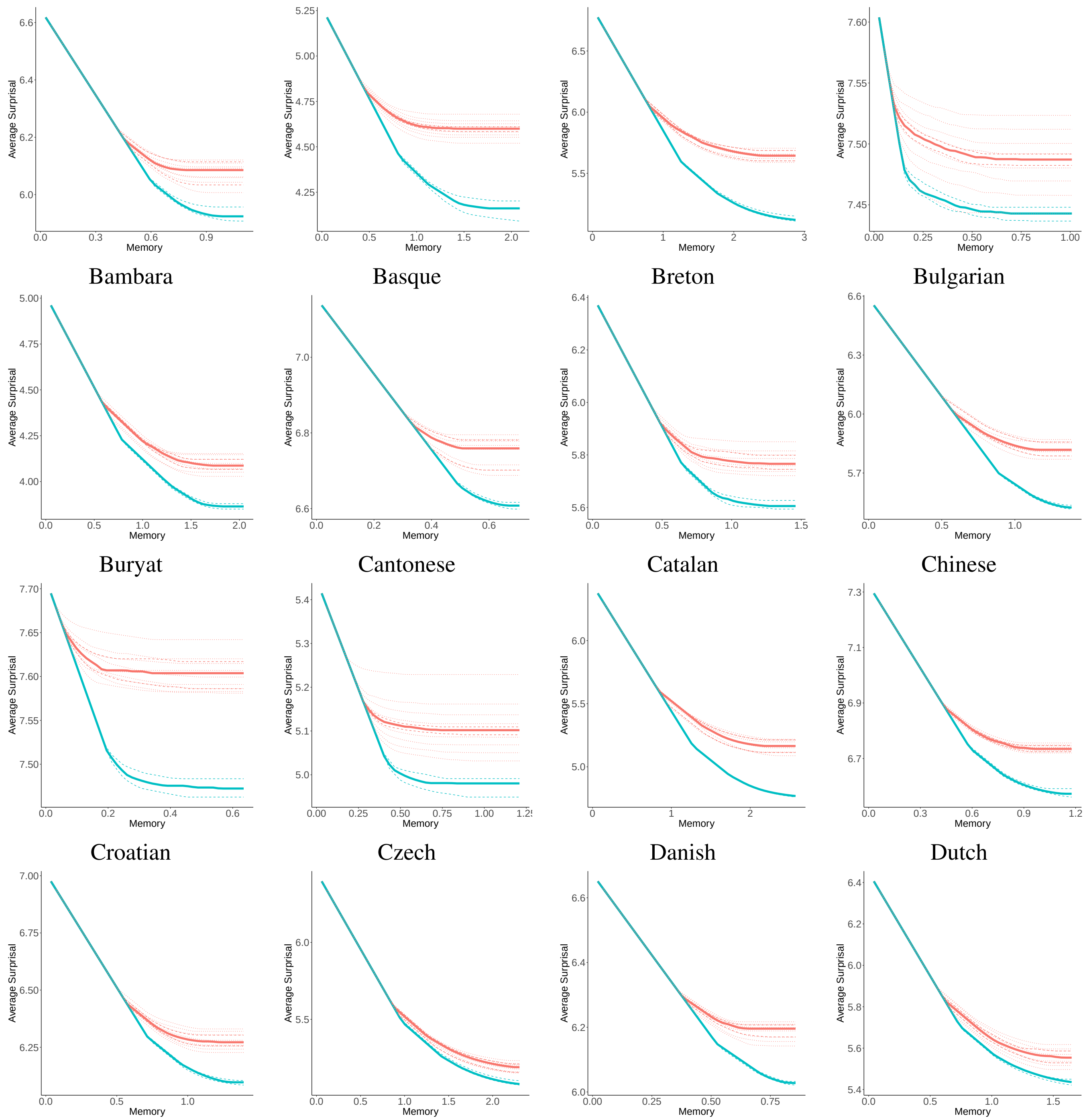


Figure 5: Medians: For each memory budget, we provide the median surprisal for real and random languages. Solid lines indicate sample medians, dashed lines indicate 95 % confidence intervals for the population median, dotted lines indicate empirical quantiles (10%, 20%, . . . , 80%, 90%). Green: Random baselines; blue: real language; red: maximum-likelihood grammars fit to real orderings. {tab:medians

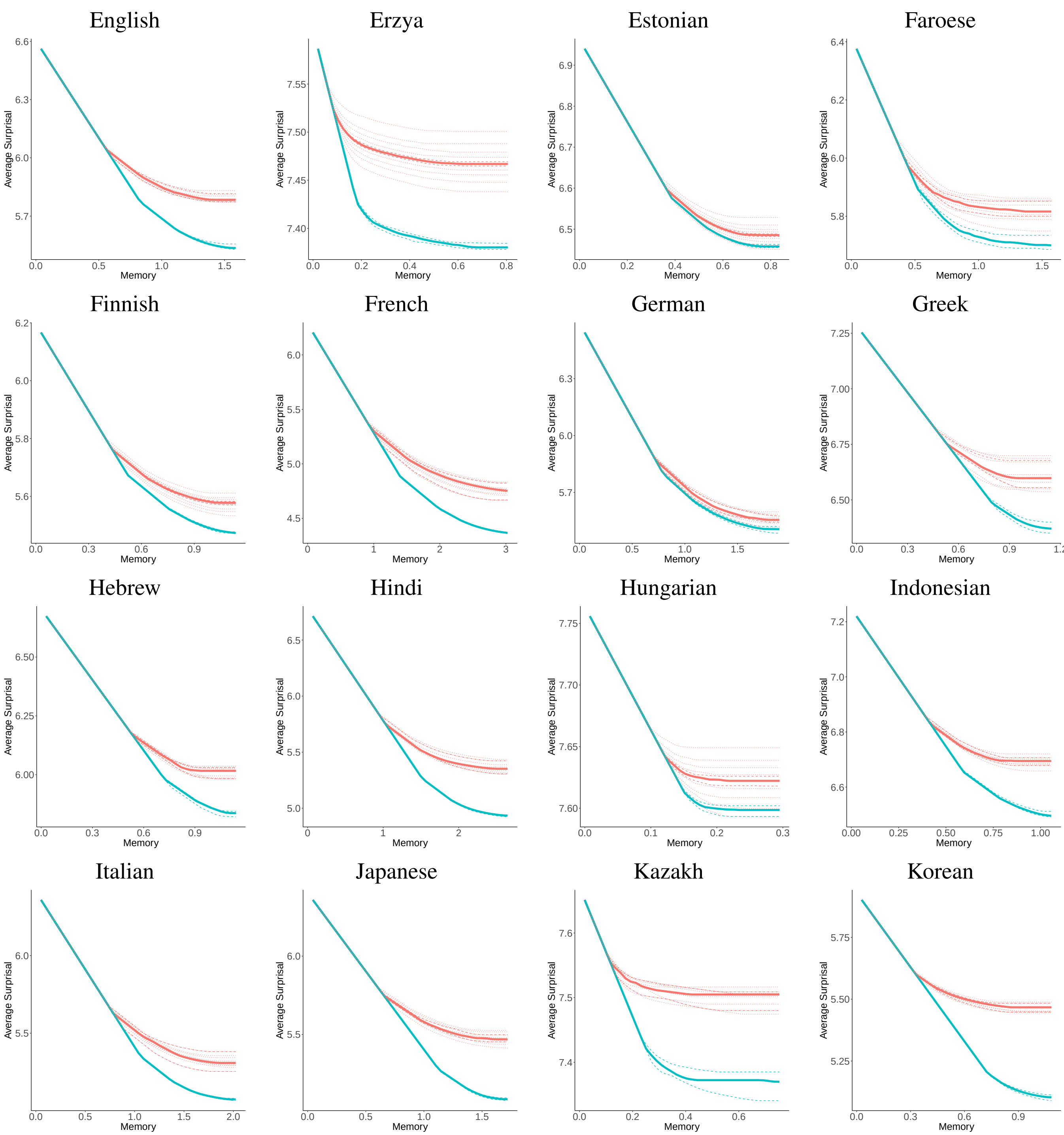
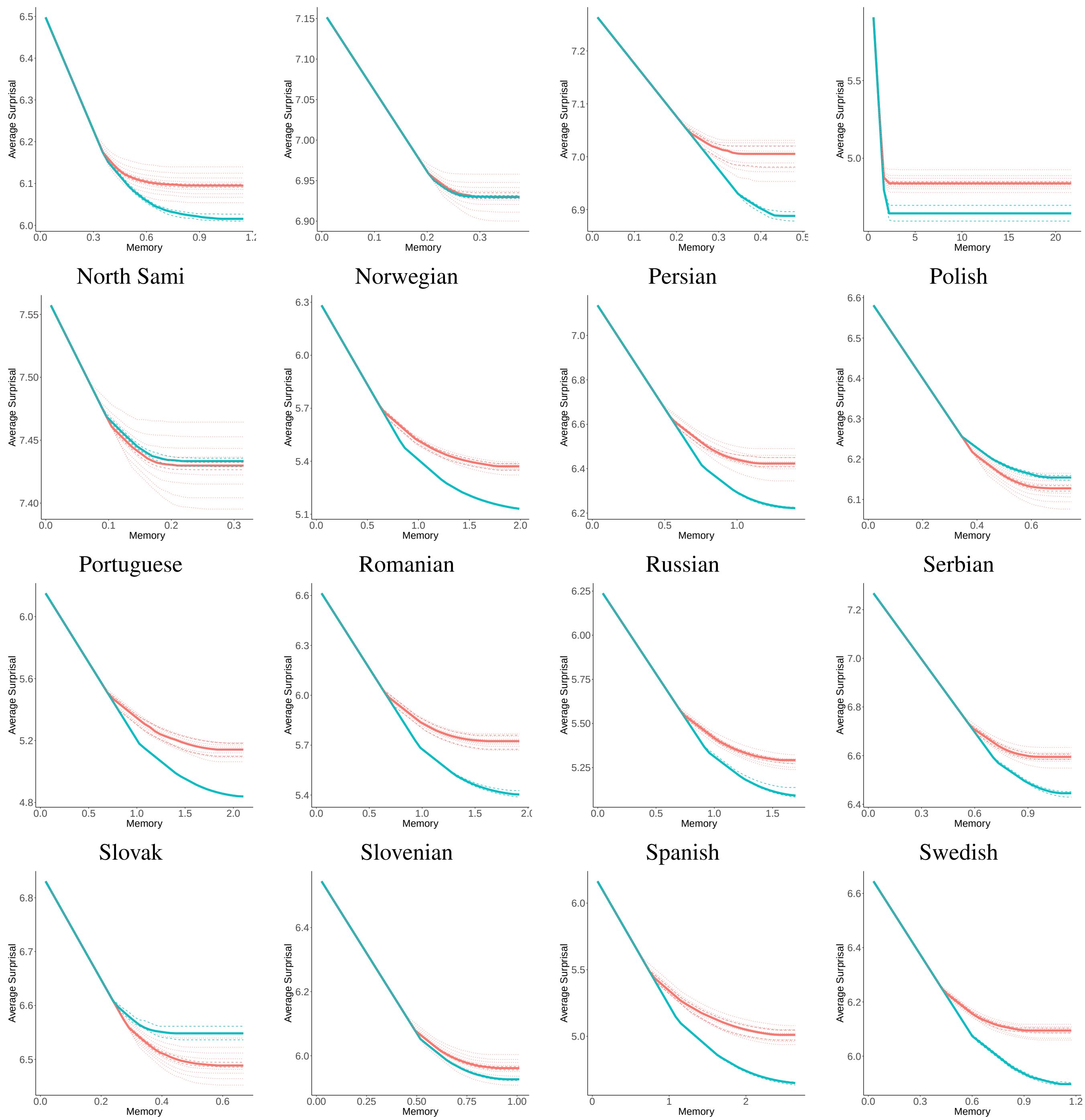


Figure 6: Medians (cont.)

Kurmanji Latvian Maltese Naija



North Sami | Norwegian | Persian | Polish

Portuguese | Romanian | Russian | Serbian

Slovak | Slovenian | Spanish | Swedish

Figure 7: Medians (cont.)

Thai | Turkish | Ukrainian | Urdu

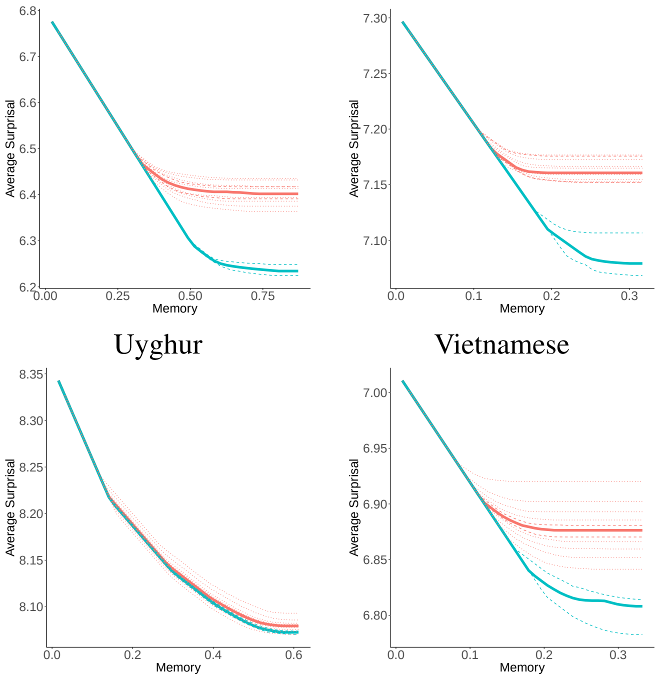
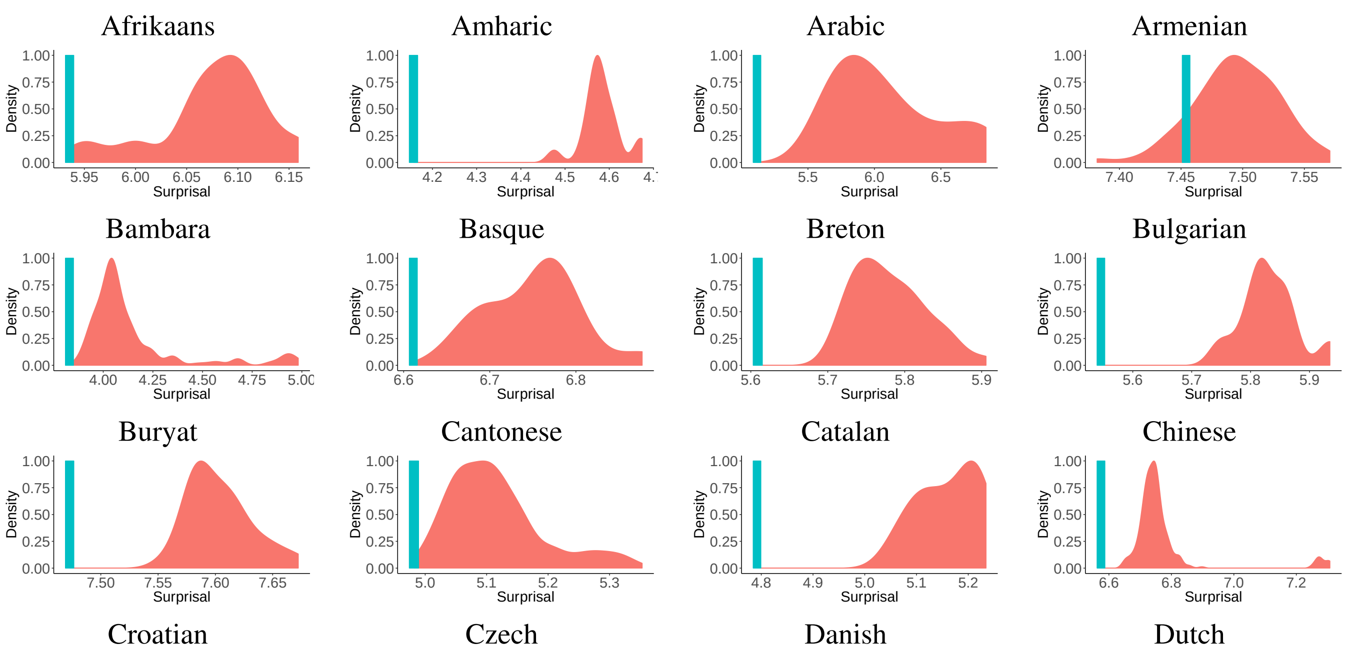
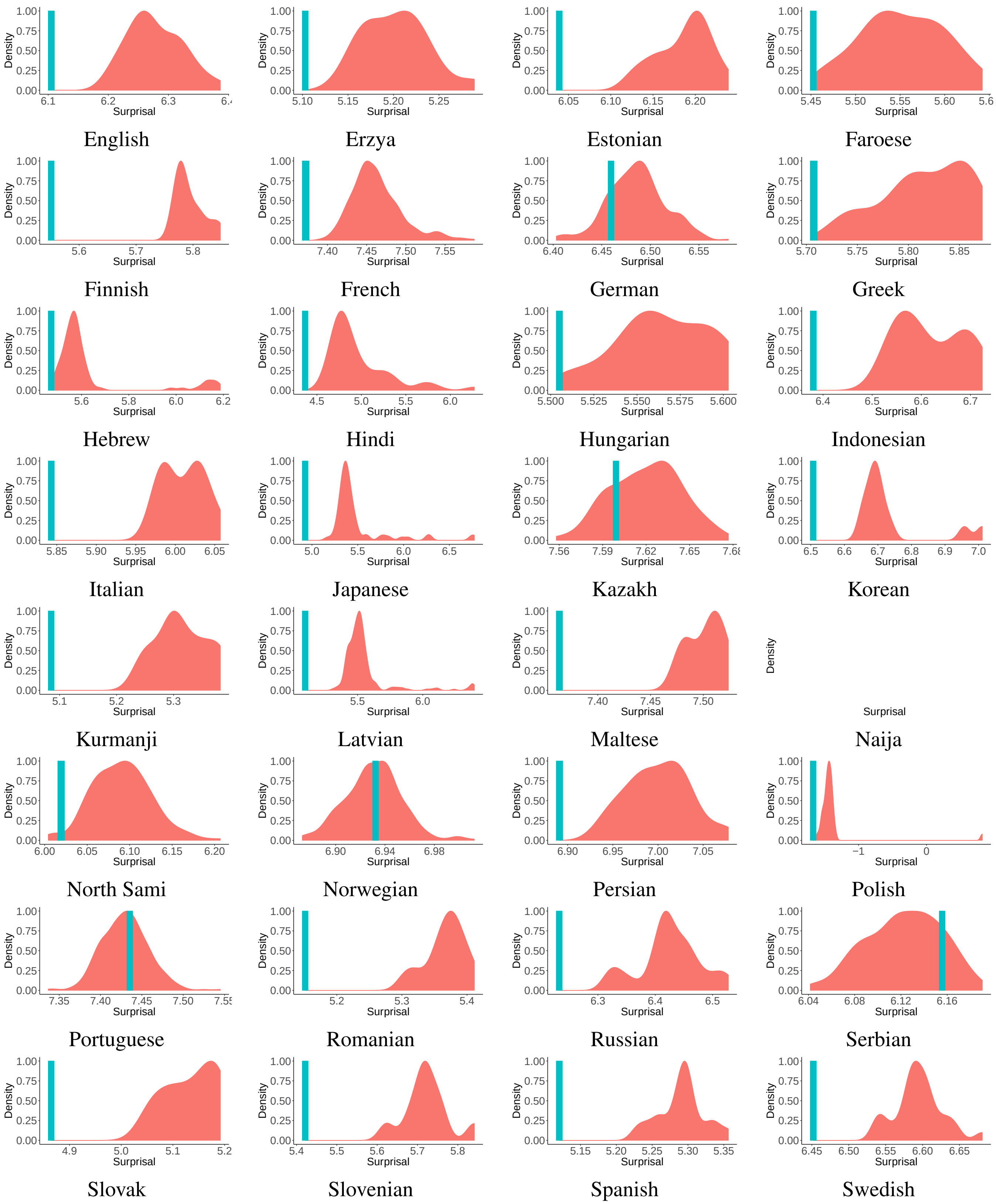


Figure 8: Medians (cont.)

5.2 Surprisal at Maximum Memory




English
Erzya
Estonian
Faroese
Finnish
French
German
Greek
Hebrew
Hindi
Hungarian
Indonesian
Italian
Japanese
Kazakh
Korean
Kurmanji
Latvian
Maltese
Naija
North Sami
Norwegian
Persian
Polish
Portuguese
Romanian
Russian
Serbian
Slovak
Slovenian
Spanish
Swedish
Density
Surprisal

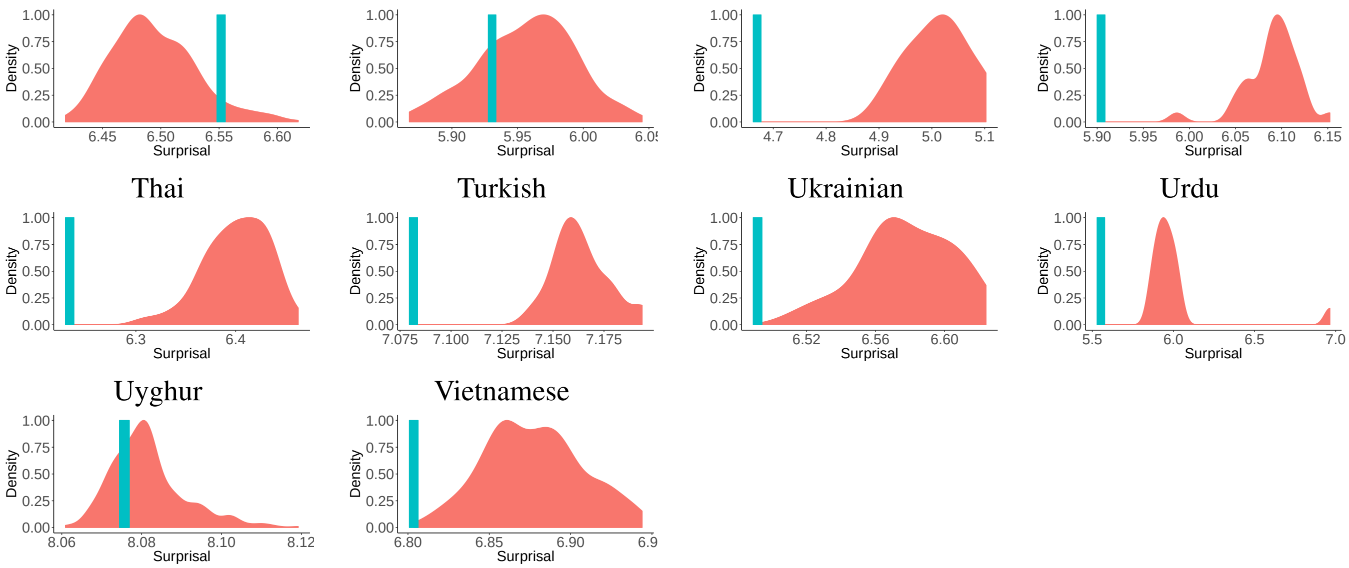


Figure 9: Histograms: Surprisal, at maximum memory.

{tab:slice-h

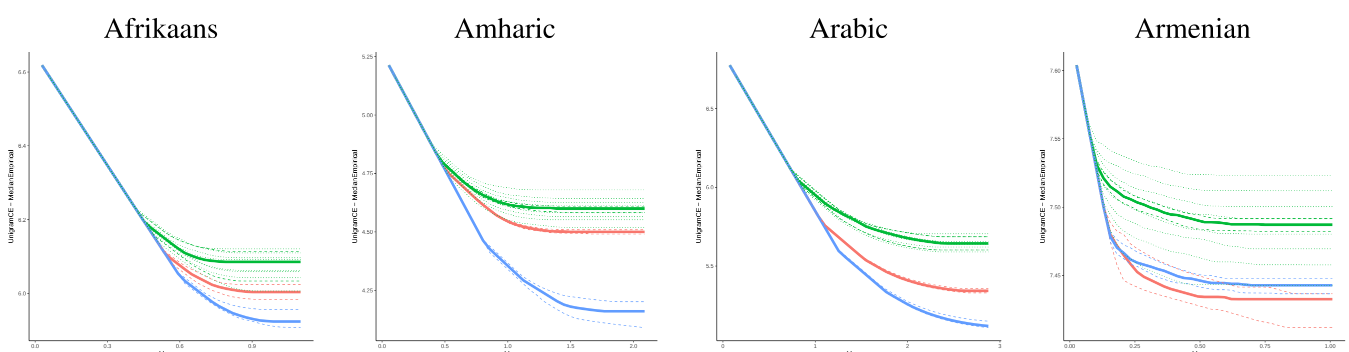
5.3 Samples Drawn (Experiment 3)

Language	Base.	MLE	Language	Base.	MLE
Afrikaans	13	10	Indonesian	11	10
Amharic	137	71	Italian	10	10
Arabic	11	10	Japanese	25	10
Armenian	140	17	Kazakh	11	10
Bambara	25	10	Korean	11	10
Basque	15	10	Kurmanji	338	101
Breton	35	10	Latvian	308	132
Bulgarian	14	10	Maltese	30	10
Buryat	26	10	Naija	214	93
Cantonese	306	135	North Sami	335	101
Catalan	11	10	Norwegian	12	10
Chinese	21	10	Persian	25	10
Croatian	30	10	Polish	309	131
Czech	18	12	Portuguese	15	99
Danish	33	10	Romanian	10	10
Dutch	27	10	Russian	20	13
English	13	10	Serbian	26	11
Erzya	846	101	Slovak	303	138
Estonian	347	10	Slovenian	297	12
Faroese	27	10	Spanish	14	10
Finnish	83	54	Swedish	31	10
French	14	12	Thai	45	10
German	19	10	Turkish	13	10
Greek	16	10	Ukrainian	28	10
Hebrew	11	10	Urdu	17	10
Hindi	11	10	Uyghur	326	132
Hungarian	220	35	Vietnamese	303	132

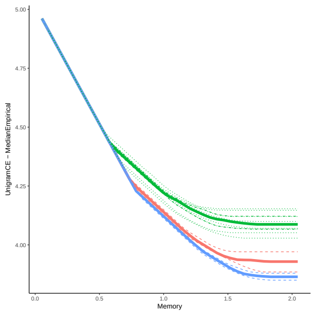
Figure 10: Experiment 3: Samples drawn per language according to the precision-dependent stopping criterion.

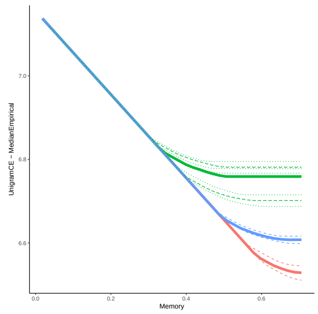
{tab:samples

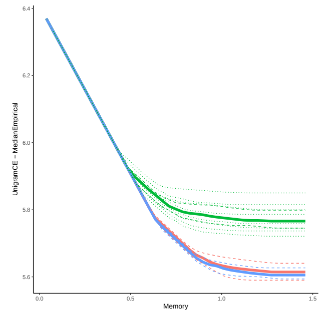
5.4 Medians (Experiment 3)

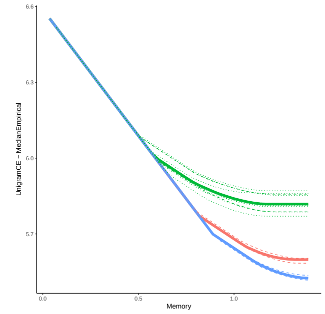


Bambara Basque Breton Bulgarian

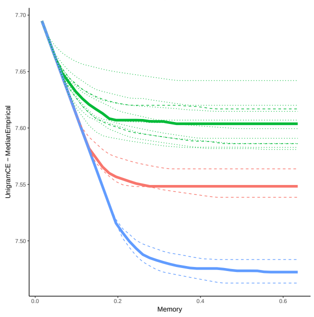


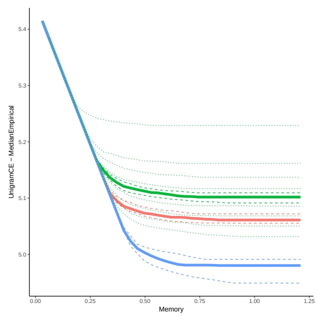


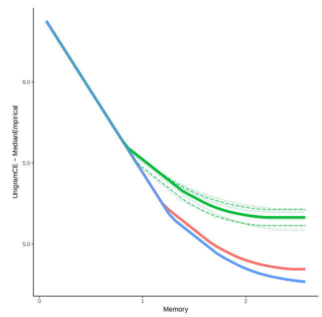


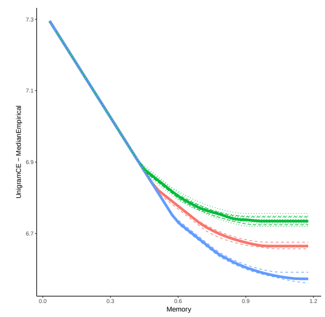


Buryat Cantonese Catalan Chinese

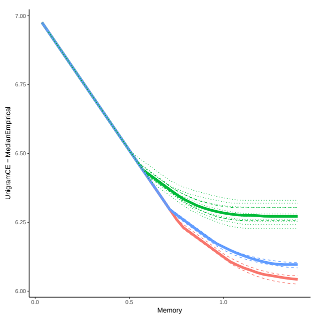


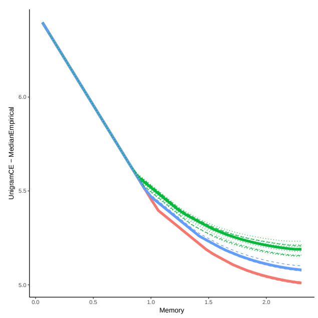


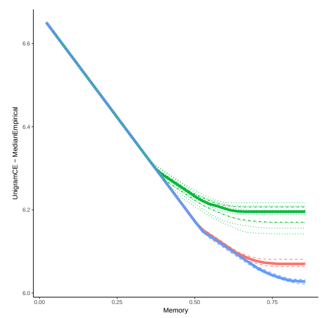


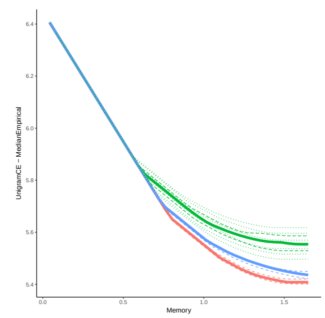


Croatian Czech Danish Dutch

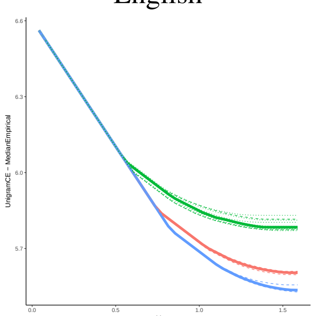


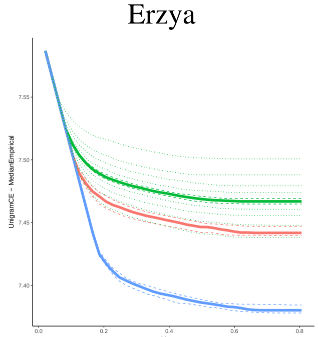


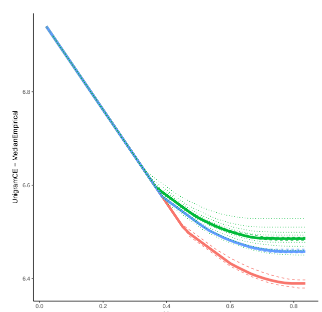


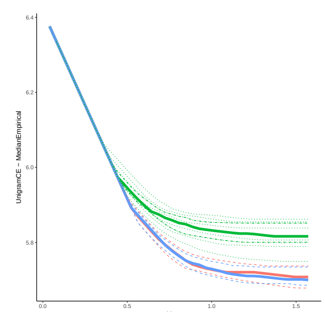


English Erzya Estonian Faroese

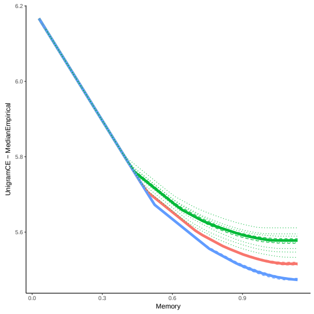




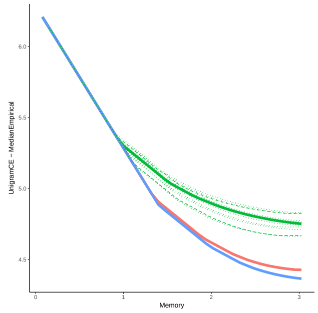




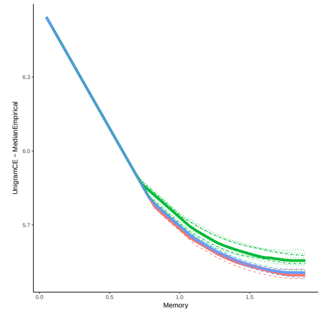
Finnish French German Greek



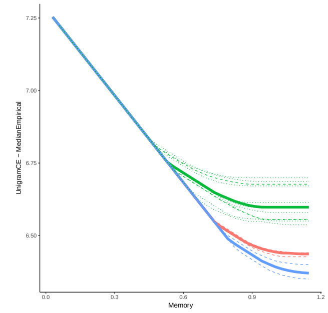
Hebrew



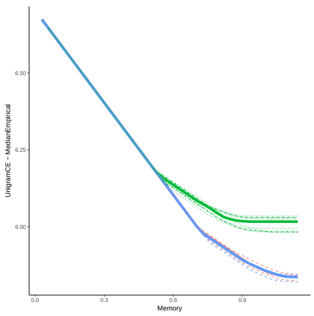
Hindi



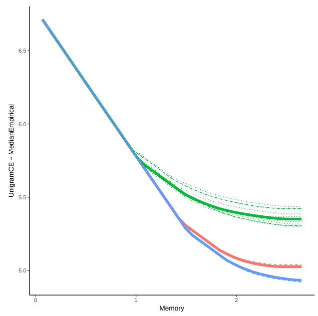
Hungarian



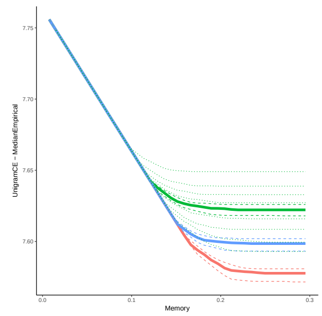
Indonesian



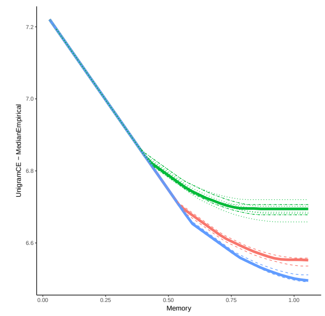
Italian



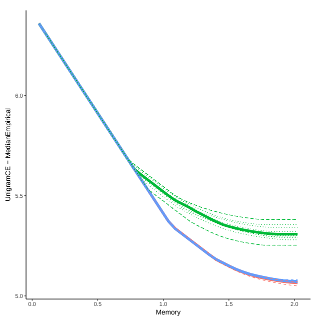
Japanese



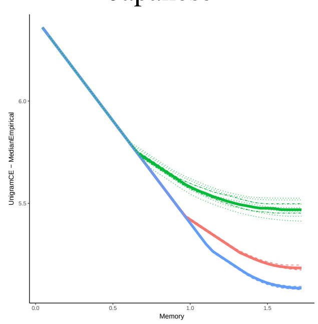
Kazakh



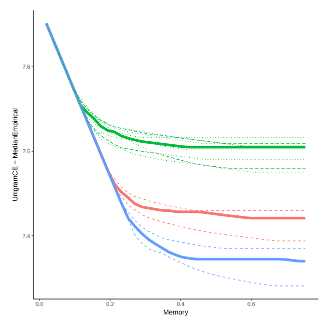
Korean



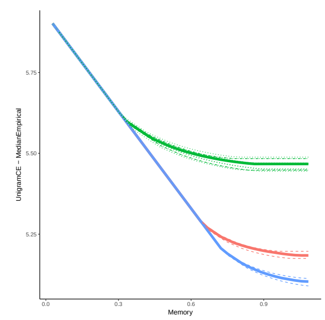
Kurmanji



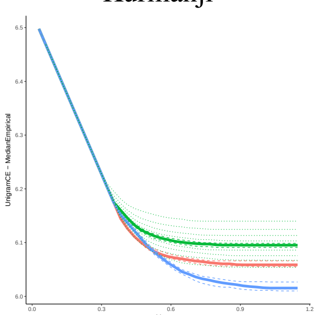
Latvian



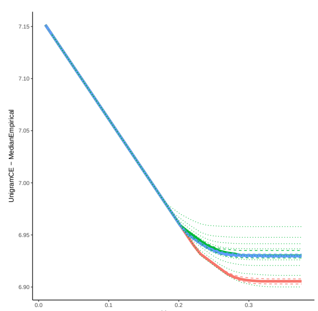
Maltese



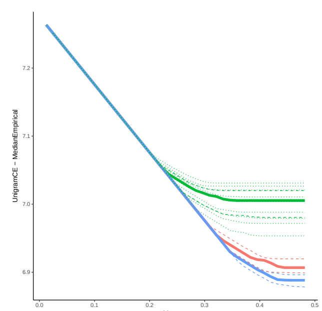
Naija



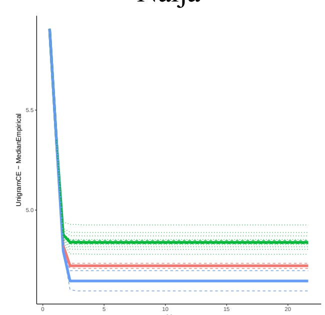
North Sami



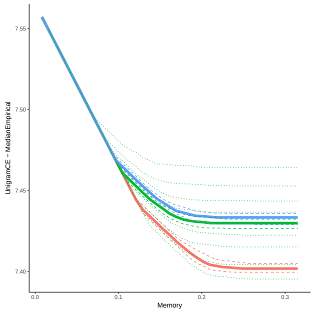
Norwegian



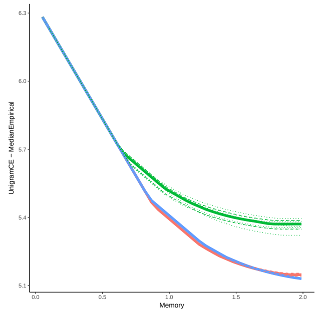
Persian



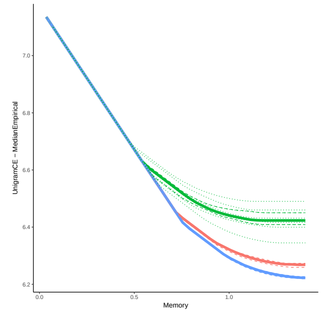
Polish



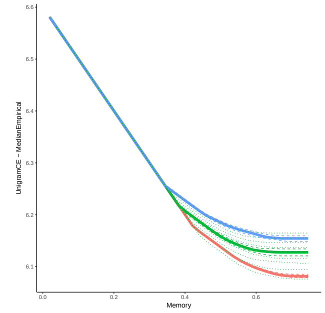
Portuguese



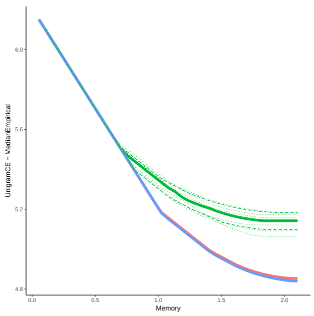
Romanian



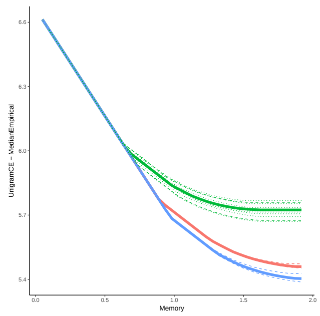
Russian



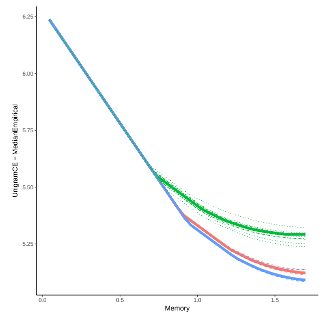
Serbian



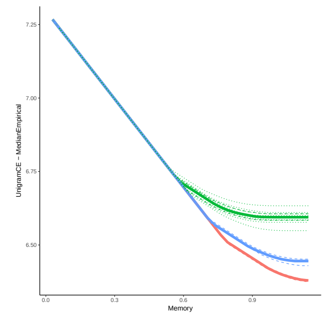
Slovak



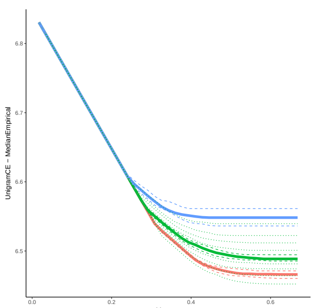
Slovenian



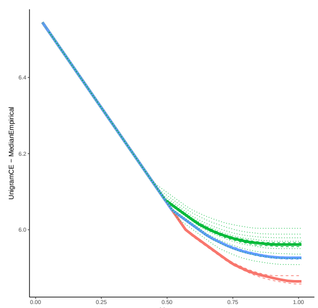
Spanish



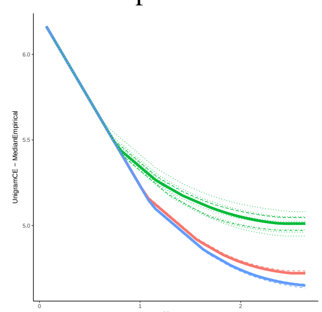
Swedish



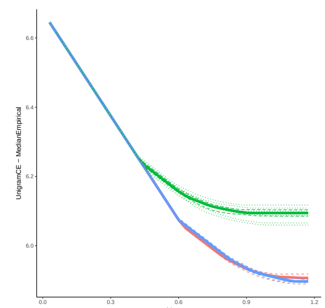
Thai



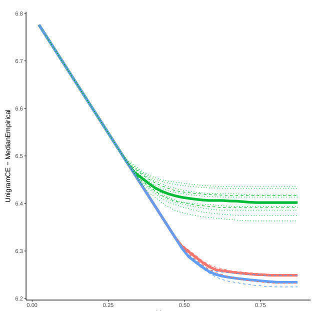
Turkish



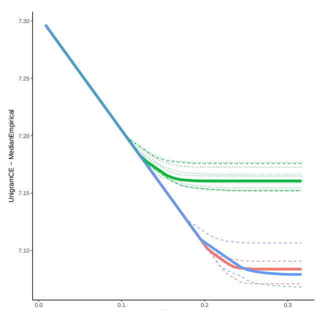
Ukrainian



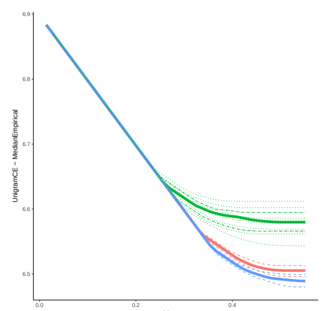
Urdu

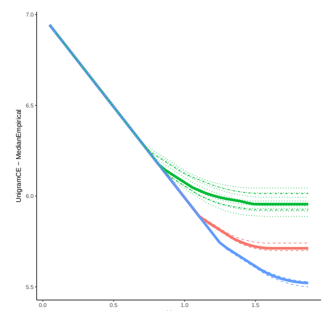


Uyghur



Vietnamese





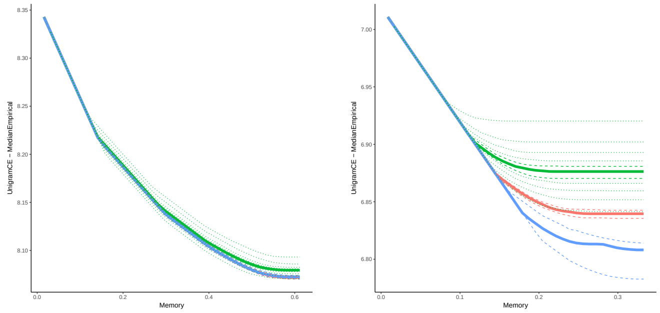
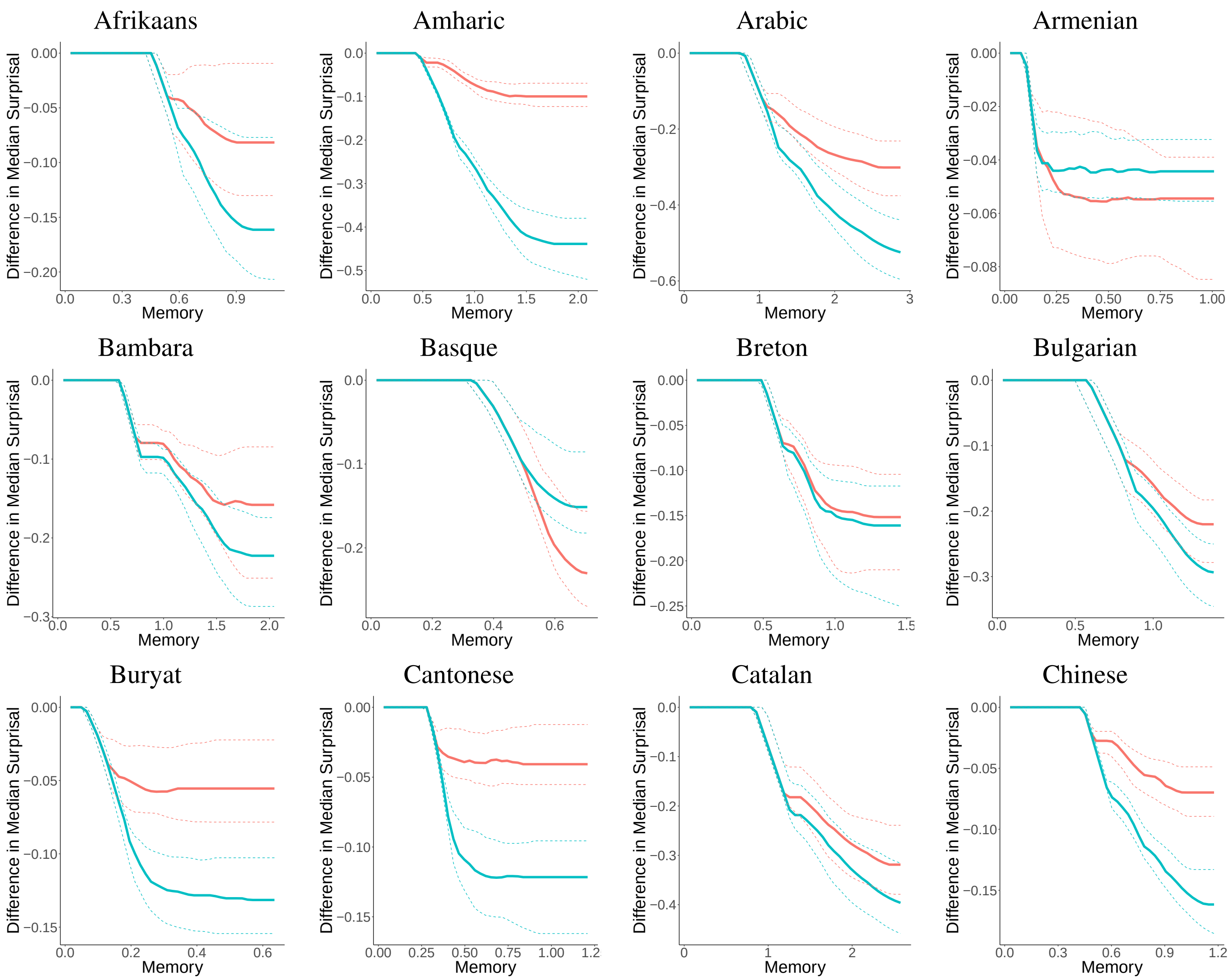
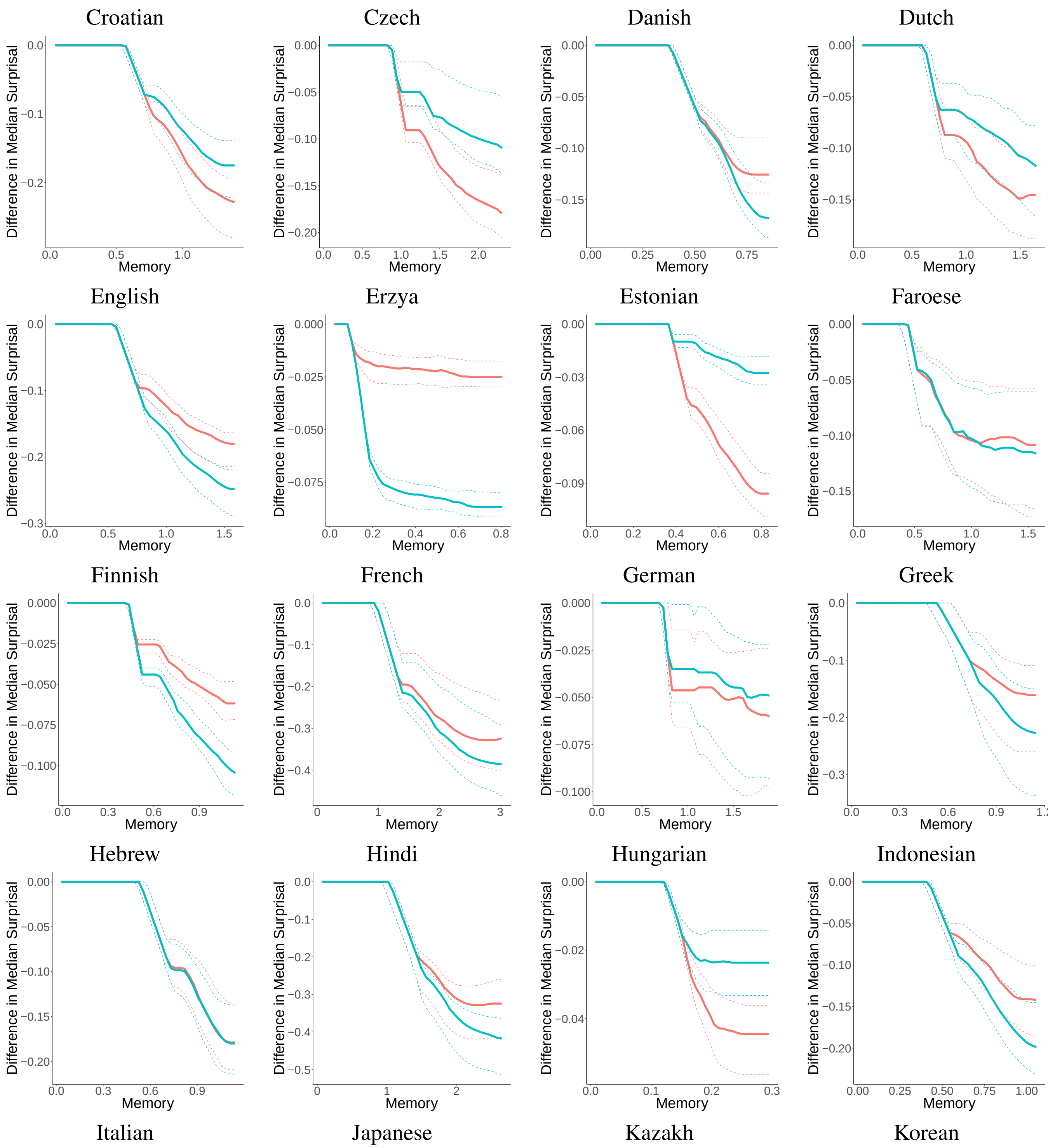
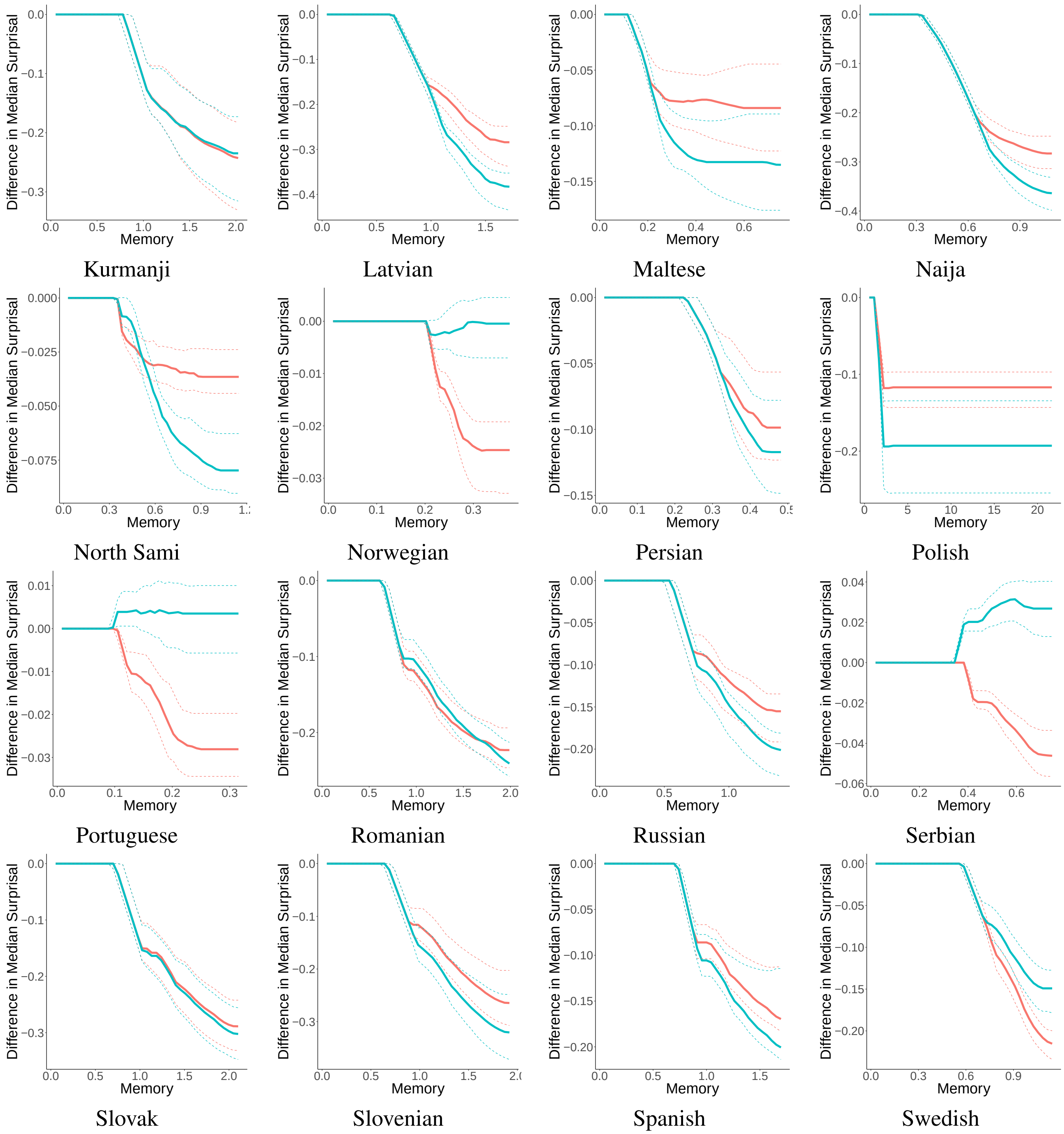


Figure 11: Experiment 3. Medians: For each memory budget, we provide the median surprisal for real and random languages. Solid lines indicate sample medians, dashed lines indicate 95 % confidence intervals for the population median. Green: Random baselines; blue: real language; red: maximum-likelihood grammars fit to real orderings.

{tab:medians




Croatian
Czech
Danish
Dutch
English
Erzya
Estonian
Faroese
Finnish
French
German
Greek
Hebrew
Hindi
Hungarian
Indonesian
Italian
Japanese
Kazakh
Korean
Difference in Median Surprisal
Memory


Difference in Median Surprisal
Memory
Kurmanji
Latvian
Maltese
Naija
North Sami
Norwegian
Persian
Polish
Portuguese
Romanian
Russian
Serbian
Slovak
Slovenian
Spanish
Swedish

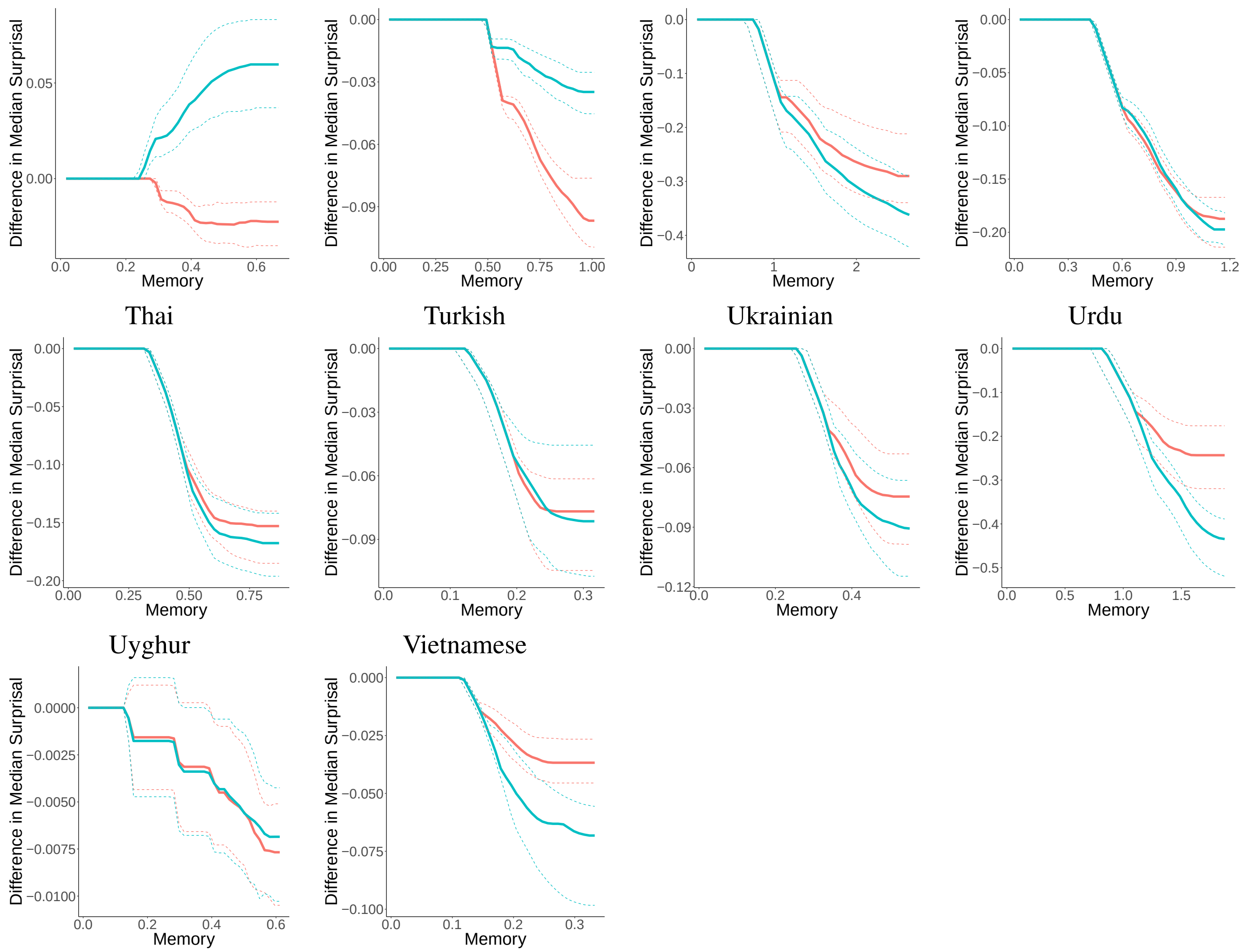
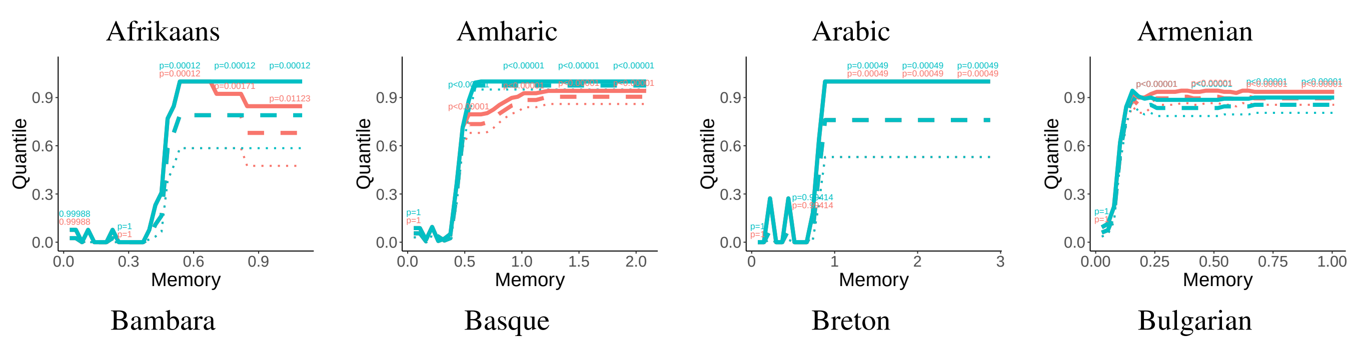
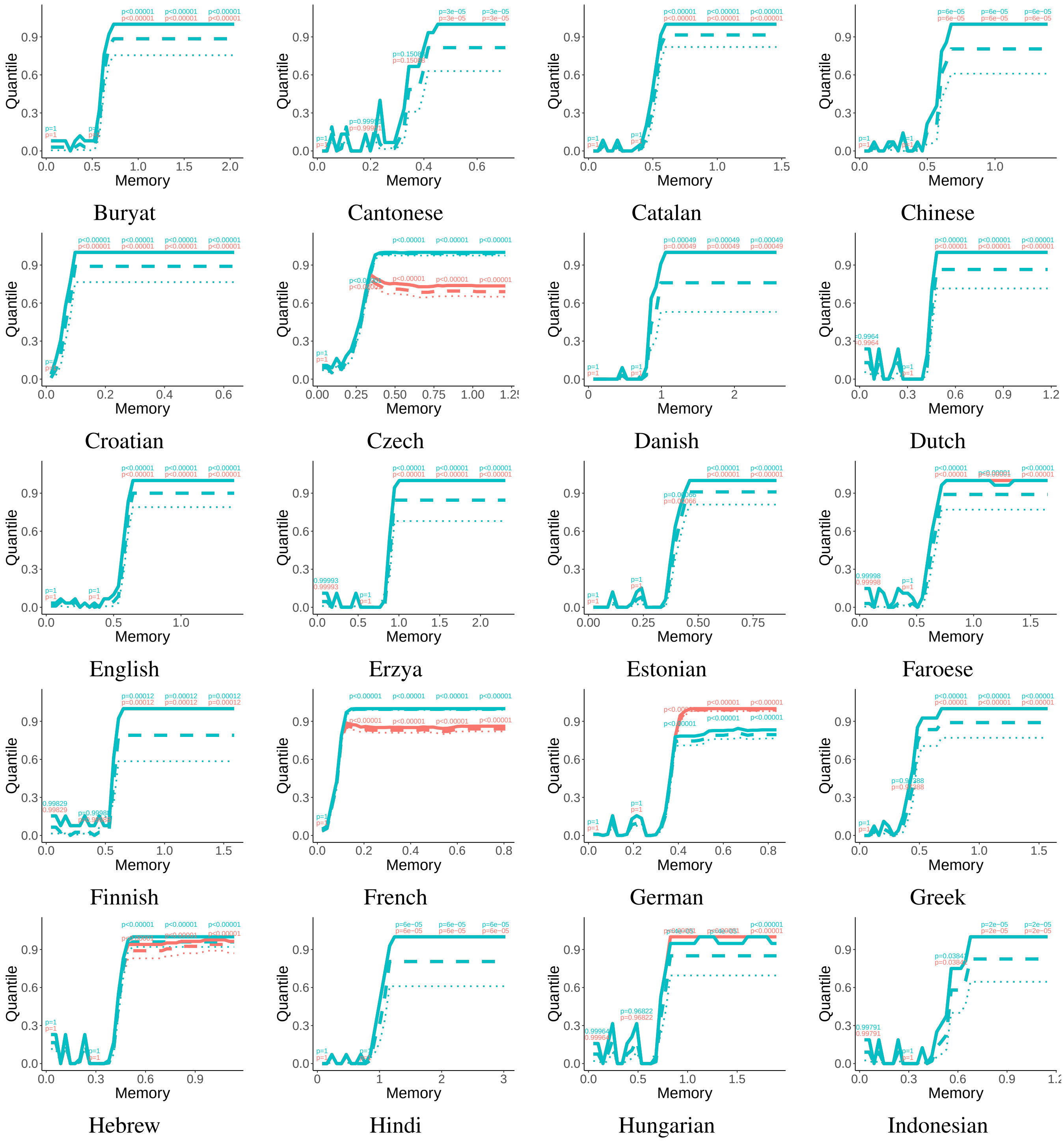
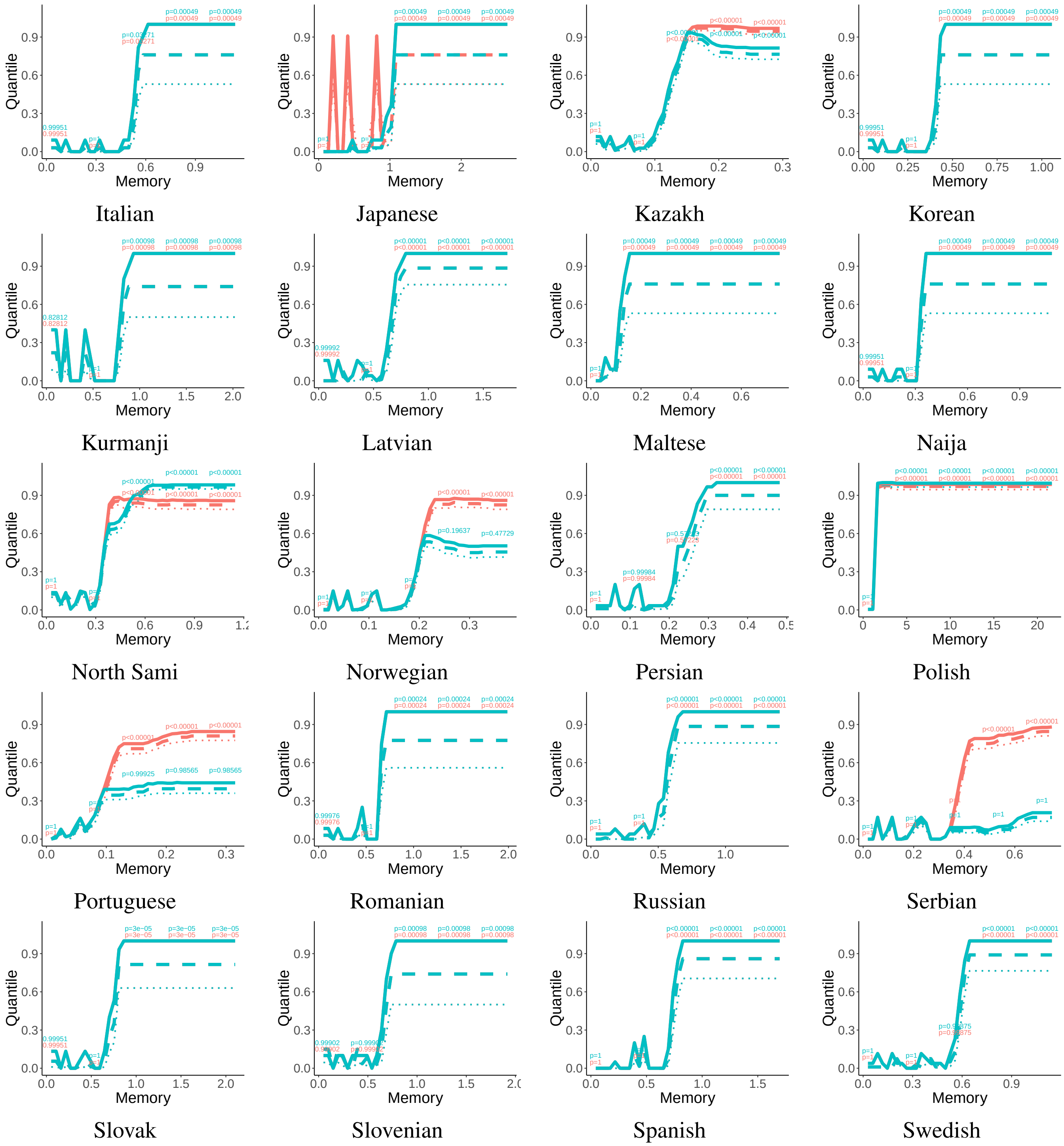


Figure 12: Median Differences between Real and Baseline: For each memory budget, we provide the difference in median surprisal between real languages and random baselines; for real orders (blue) and maximum likelihood grammars (red). Lower values indicate lower surprisal compared to baselines. Solid lines indicate sample means. Dashed lines indicate 95 % confidence intervals. {tab:median_




Quantile
Memory
Buryat
Cantonese
Catalan
Chinese
Croatian
Czech
Danish
Dutch
English
Erzya
Estonian
Faroese
Finnish
French
German
Greek
Hebrew
Hindi
Hungarian
Indonesian


Italian
Japanese
Kazakh
Korean
Kurmanji
Latvian
Maltese
Naija
North Sami
Norwegian
Persian
Polish
Portuguese
Romanian
Russian
Serbian
Slovak
Slovenian
Spanish
Swedish

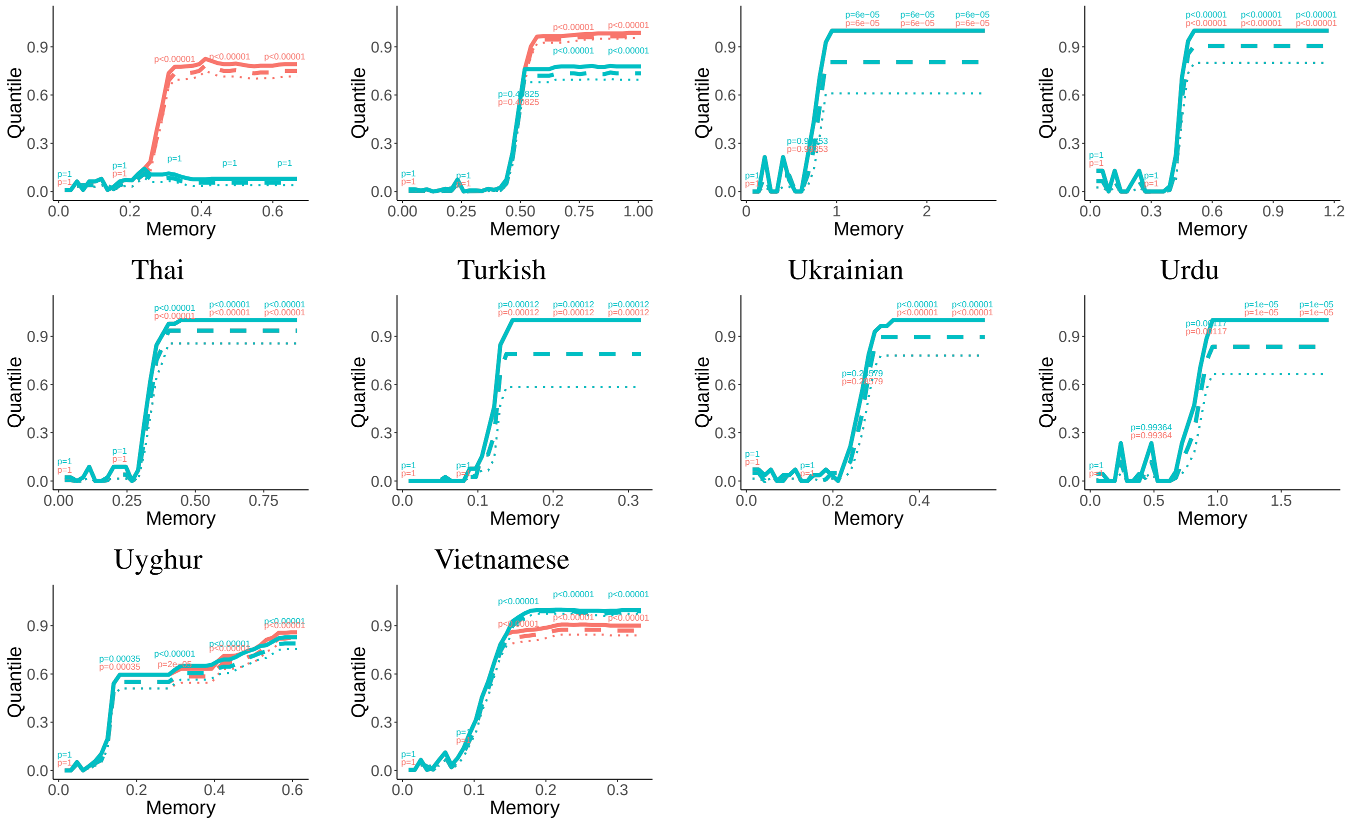


Figure 13: Quantiles: At a given memory budget, what percentage of the baselines results in higher listener surprisal than the real language? Solid curves represent sample means, dashed lines represent 95 % confidence bounds; dotted lines represent 99.9 % confidence bounds. At five evenly spaced memory levels, we provide a p-value for the null hypothesis that the actual population mean is 0.5 or less. Confidence bounds and p-values are obtained using an exact nonparametric method (see text). {tab:quantil

6 Details for Neural Network Models

7 N-Gram Models

7.1 Method

We use a version of Kneser-Ney Smoothing. For a sequence $w_1 \dots w_k$, let $N(w_{1\dots k})$ be the number of times $w_{1\dots k}$ occurs in the training set. The unigram probabilities are estimated as

$$p_1(w_t) := \frac{N(w_t) + \delta}{|Train| + |V| \cdot \delta} \tag{32}$$

where $\delta \in \mathbb{R}_+$ is a hyperparameter. Here $|Train|$ is the number of tokens in the training set, $|V|$ is the number of types occurring in train or held-out data. Higher-order probabilities $p_t(w_t|w_{0\dots t-1})$ are estimated recursively as follows. Let $\gamma > 0$ be a hyperparameter. If $N(w_{0\ldots t-1}) < \gamma$, set

$$p_t(w_t|w_{0\ldots t-1}) := p_{t-1}(w_t|w_{1\ldots t-1}) \tag{33}$$

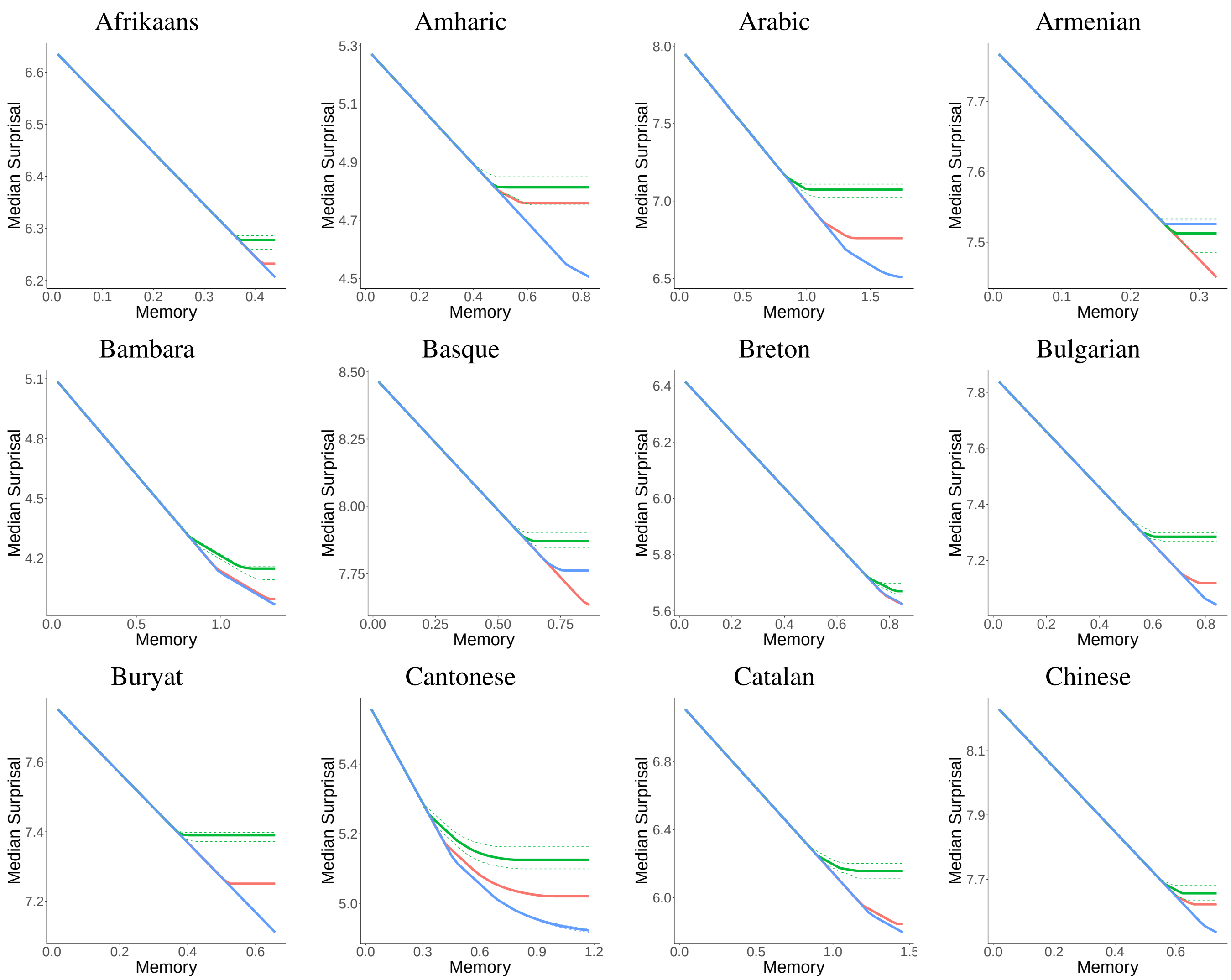
Otherwise, we interpolate between t-th order and lower-order estimates:

$$p_t(w_t|w_{0\ldots t-1}) := \frac{\max(N(w_{0\ldots t}) - \alpha, 0.0) + \alpha \cdot \#\{w : N(w_{0\ldots t-1}w) > 0\} \cdot p_{t-1}(w_t|w_{1\ldots t-1})}{N(w_{0\ldots t-1})} \tag{34}$$

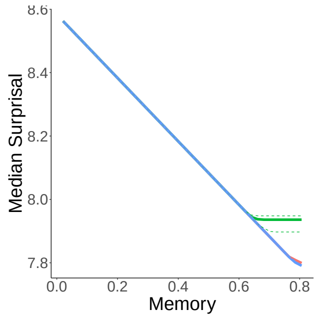
where $\alpha \in [0, 1]$ is also a hyperparameter. (CITE) show that this definition results in a well-defined probability distribution, i.e., $\sum_{w \in V} p_t(w|w_{0\ldots t-1}) = 1$.

Hyperparameters α, γ, δ are tuned with the same strategy as for the neural network models.

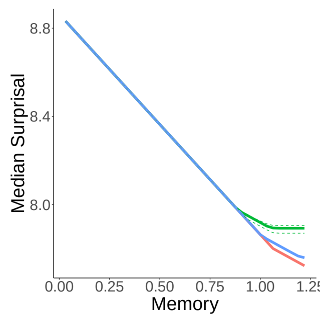
7.2 Results



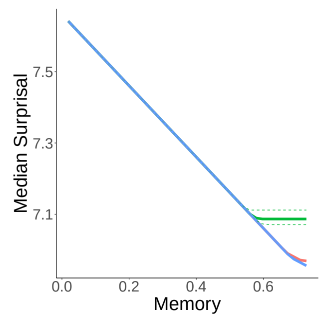
Croatian Czech Danish Dutch



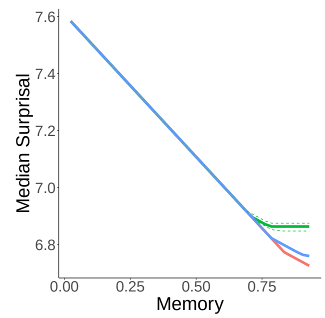
English



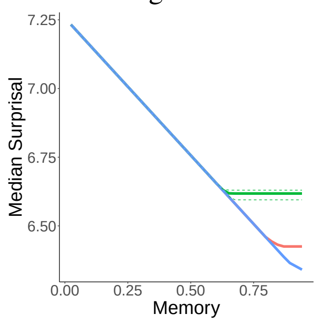
Erzya



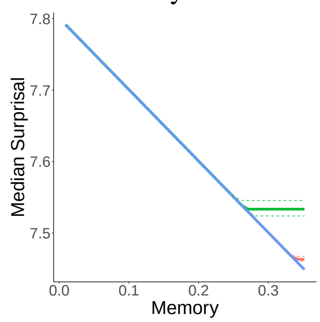
Estonian



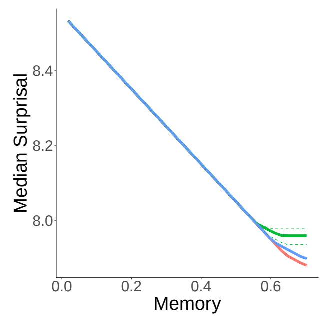
Faroese



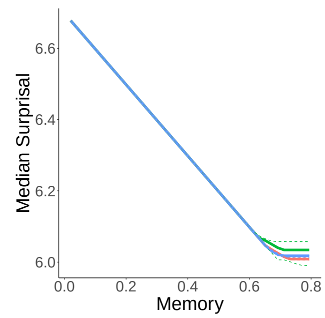
Finnish



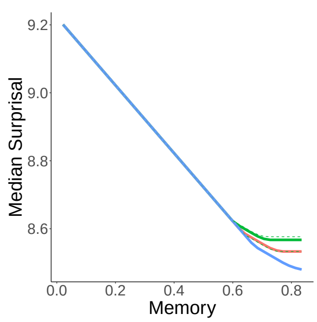
French



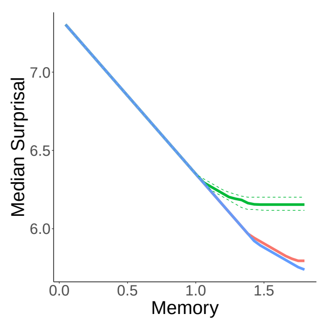
German



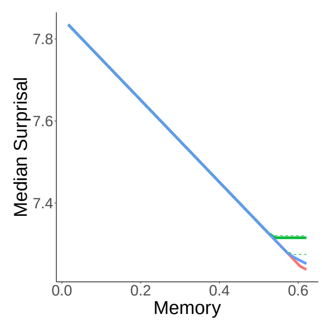
Greek



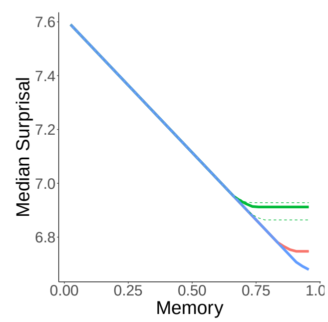
Hebrew



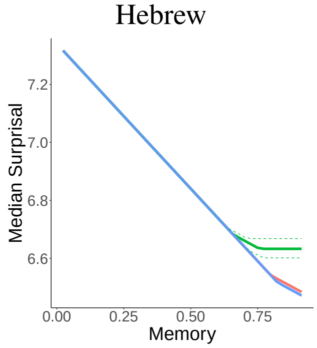
Hindi



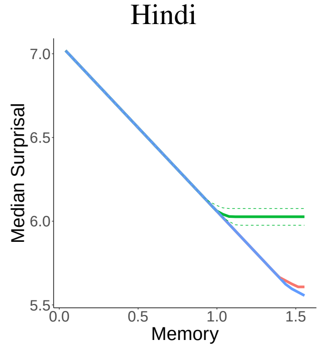
Hungarian



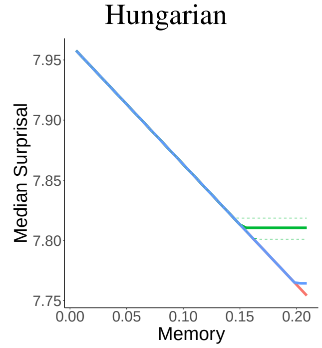
Indonesian



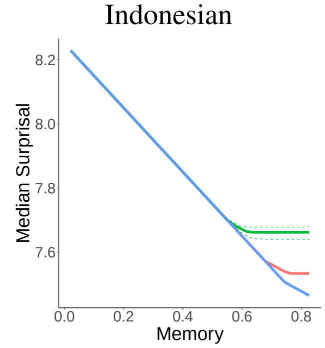
Italian



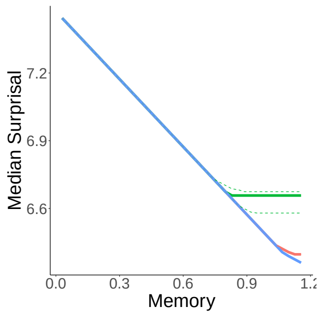
Japanese



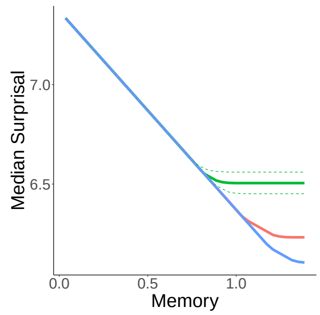
Kazakh



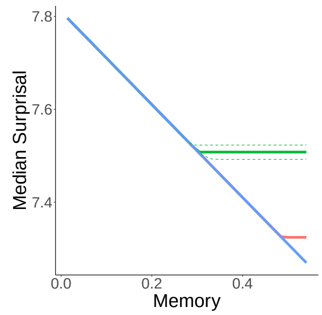
Korean



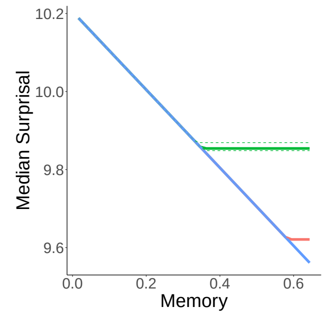
Kurmanji



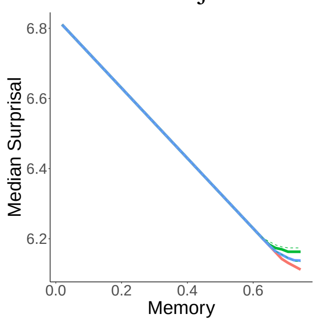
Latvian



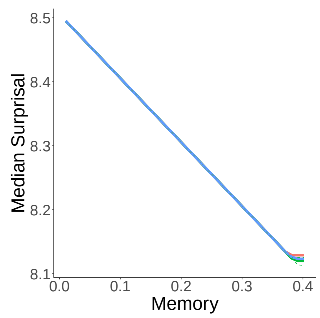
Maltese



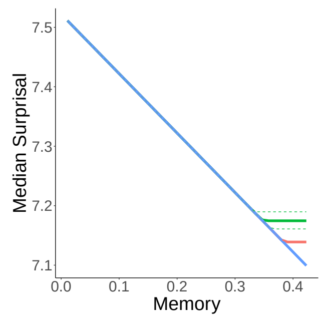
Naija



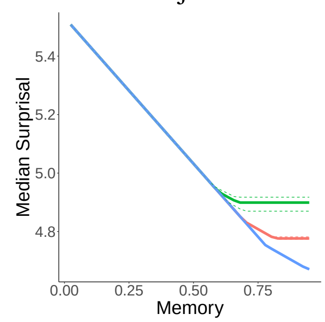
North Sami



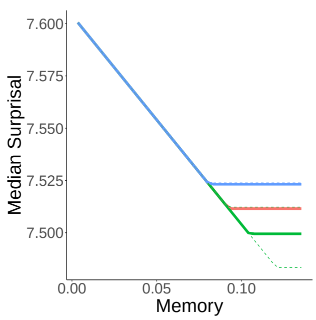
Norwegian



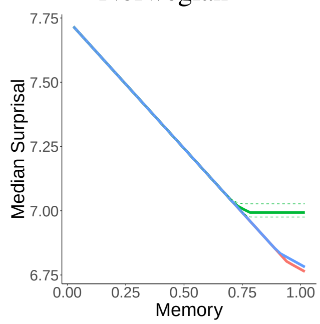
Persian



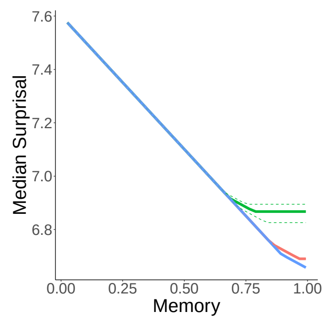
Polish



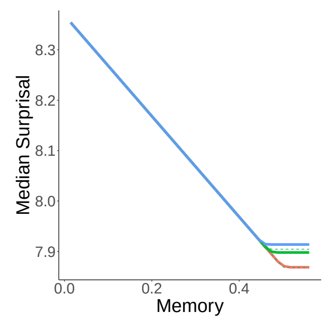
Portuguese



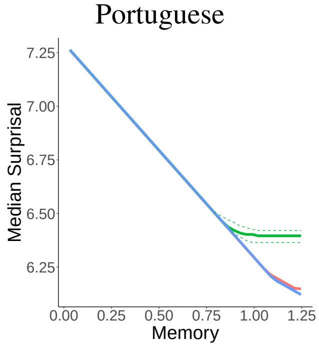
Romanian



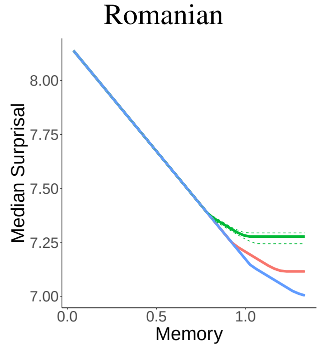
Russian



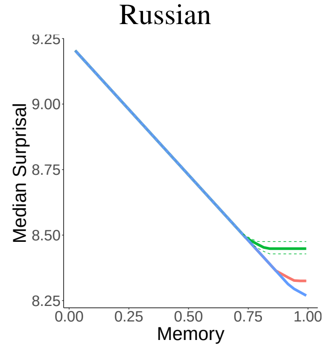
Serbian



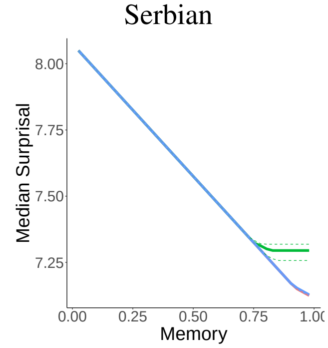
Slovak



Slovenian



Spanish



Swedish

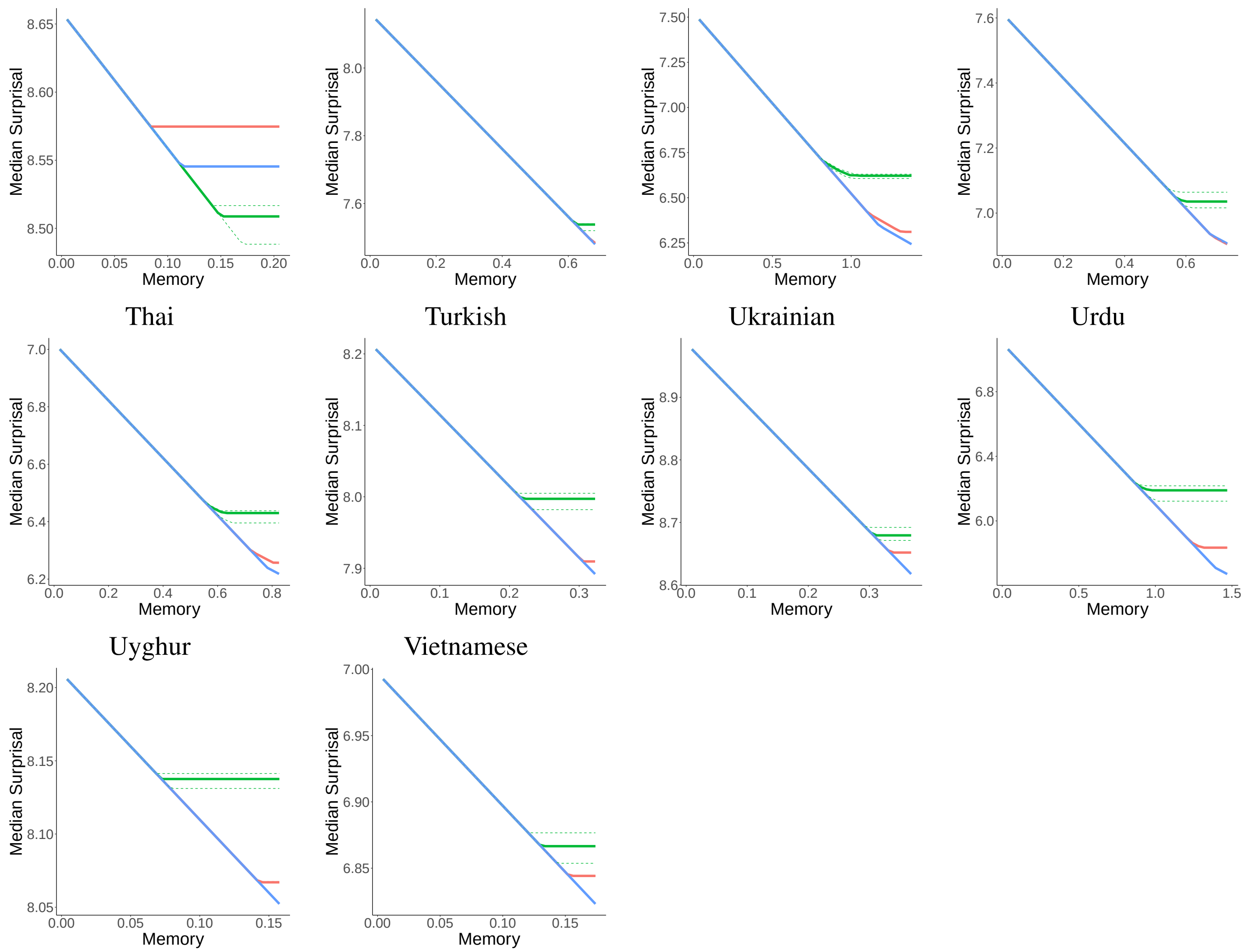
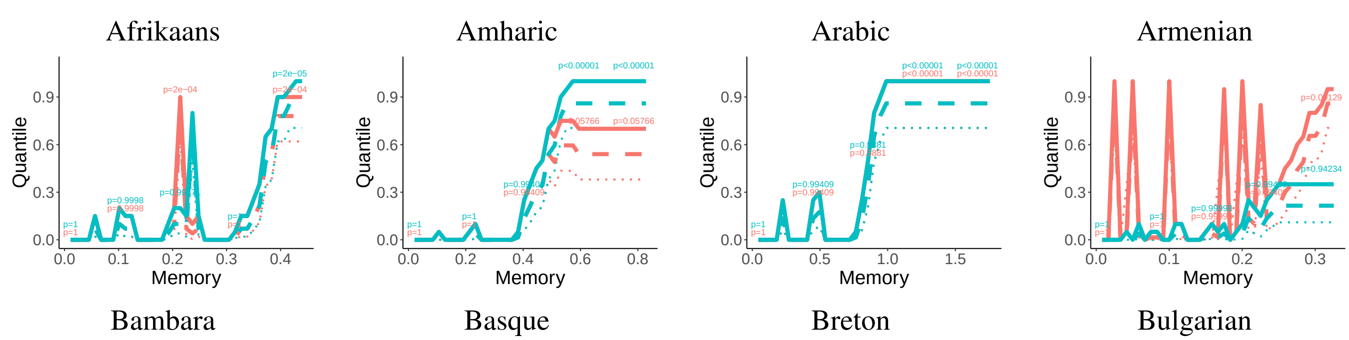
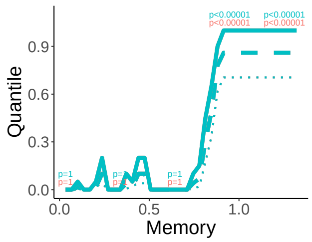


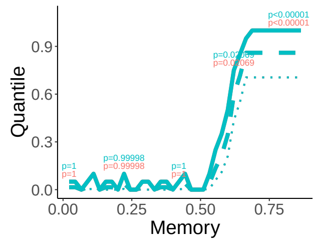
Figure 14: Medians (estimated using n-gram models): For each memory budget, we provide the median surprisal for real and random languages. Solid lines indicate sample medians for ngrams, dashed lines indicate 95 % confidence intervals for the population median. Green: Random baselines; blue: real language; red: maximum-likelihood grammars fit to real orderings.

{tab:medians_

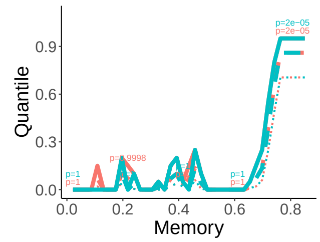




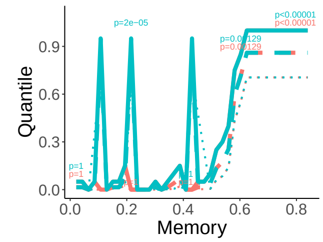
Buryat



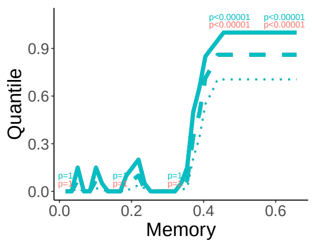
Cantonese



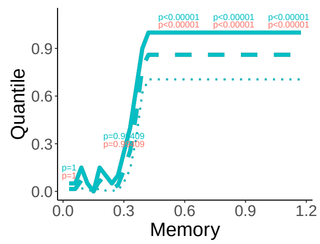
Catalan



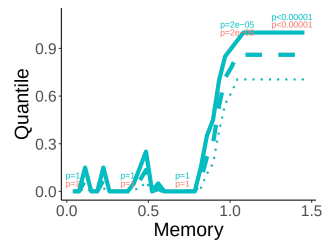
Chinese



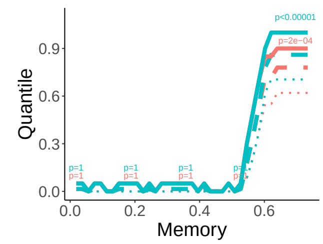
Croatian



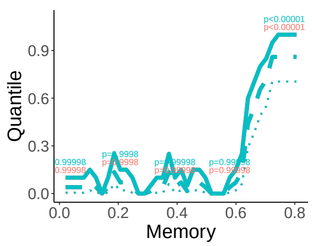
Czech



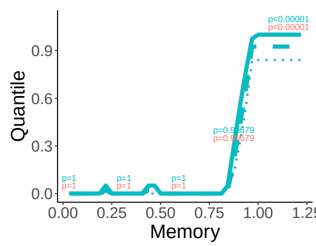
Danish



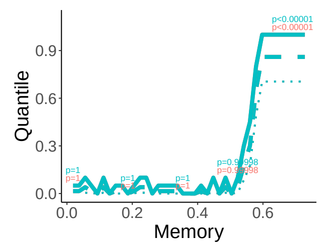
Dutch



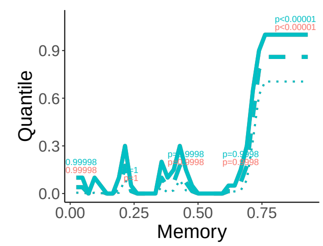
English



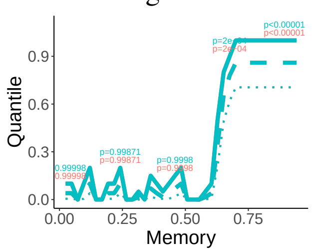
Erzya



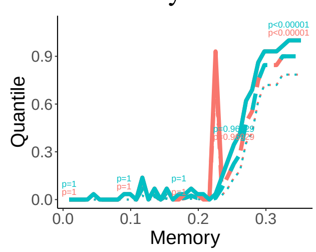
Estonian



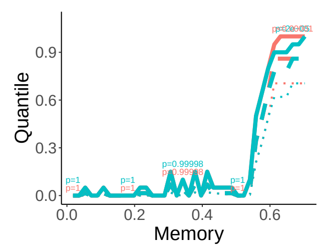
Faroese



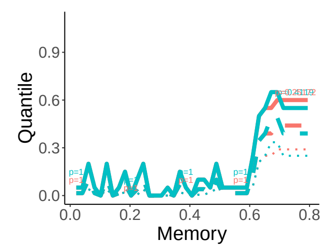
Finnish



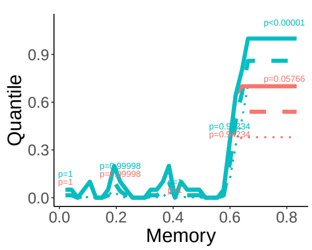
French



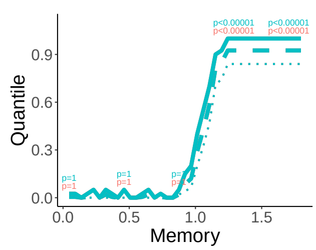
German



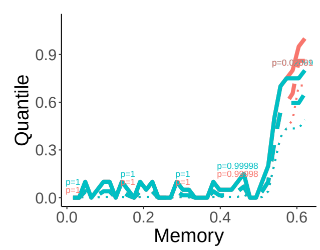
Greek



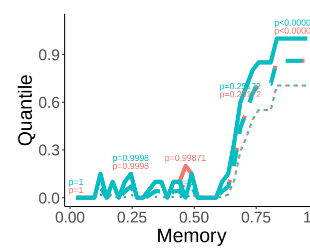
Hebrew



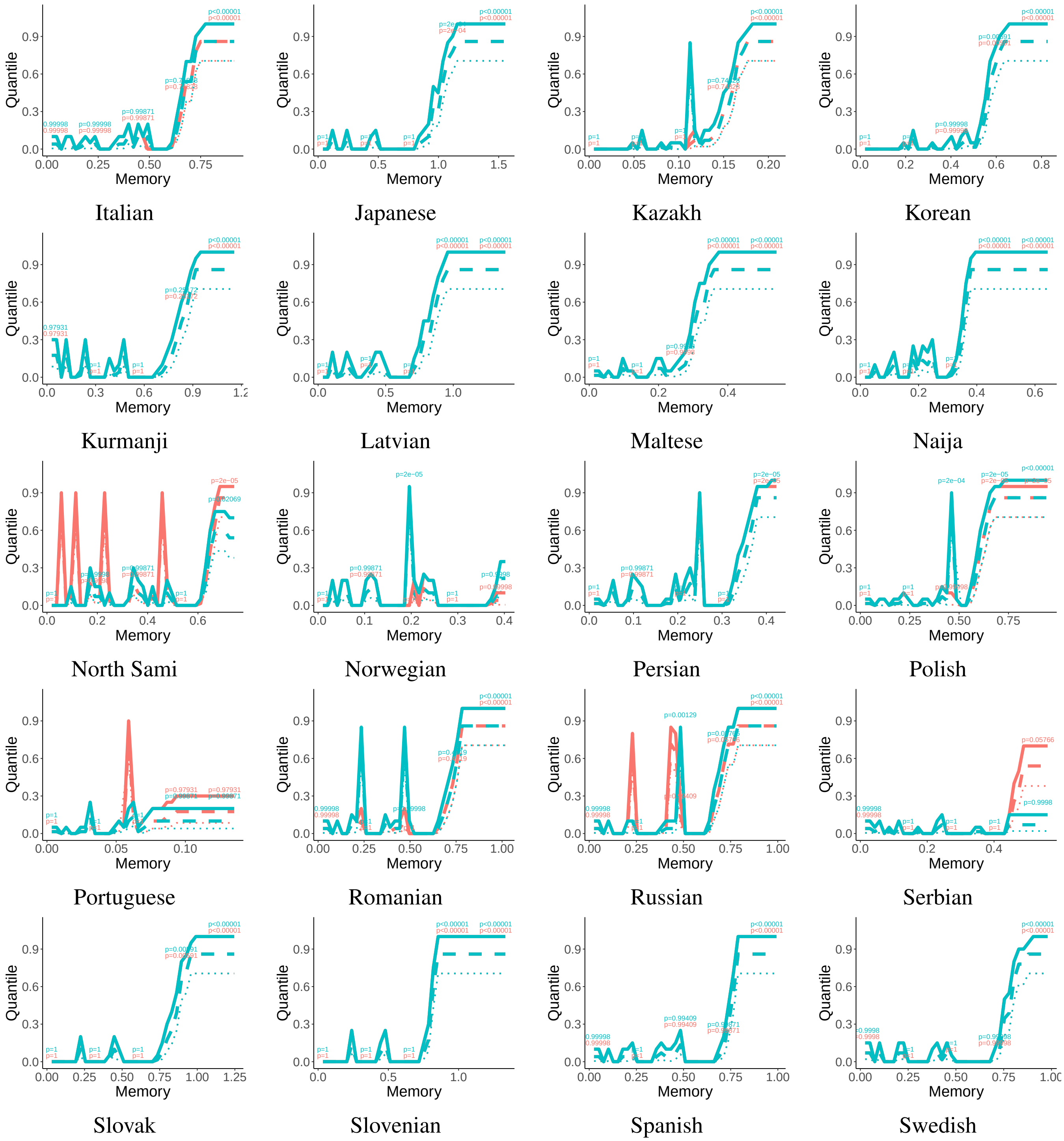
Hindi



Hungarian



Indonesian


Italian
Japanese
Kazakh
Korean
Kurmanji
Latvian
Maltese
Naija
North Sami
Norwegian
Persian
Polish
Portuguese
Romanian
Russian
Serbian
Slovak
Slovenian
Spanish
Swedish
Quantile
Memory

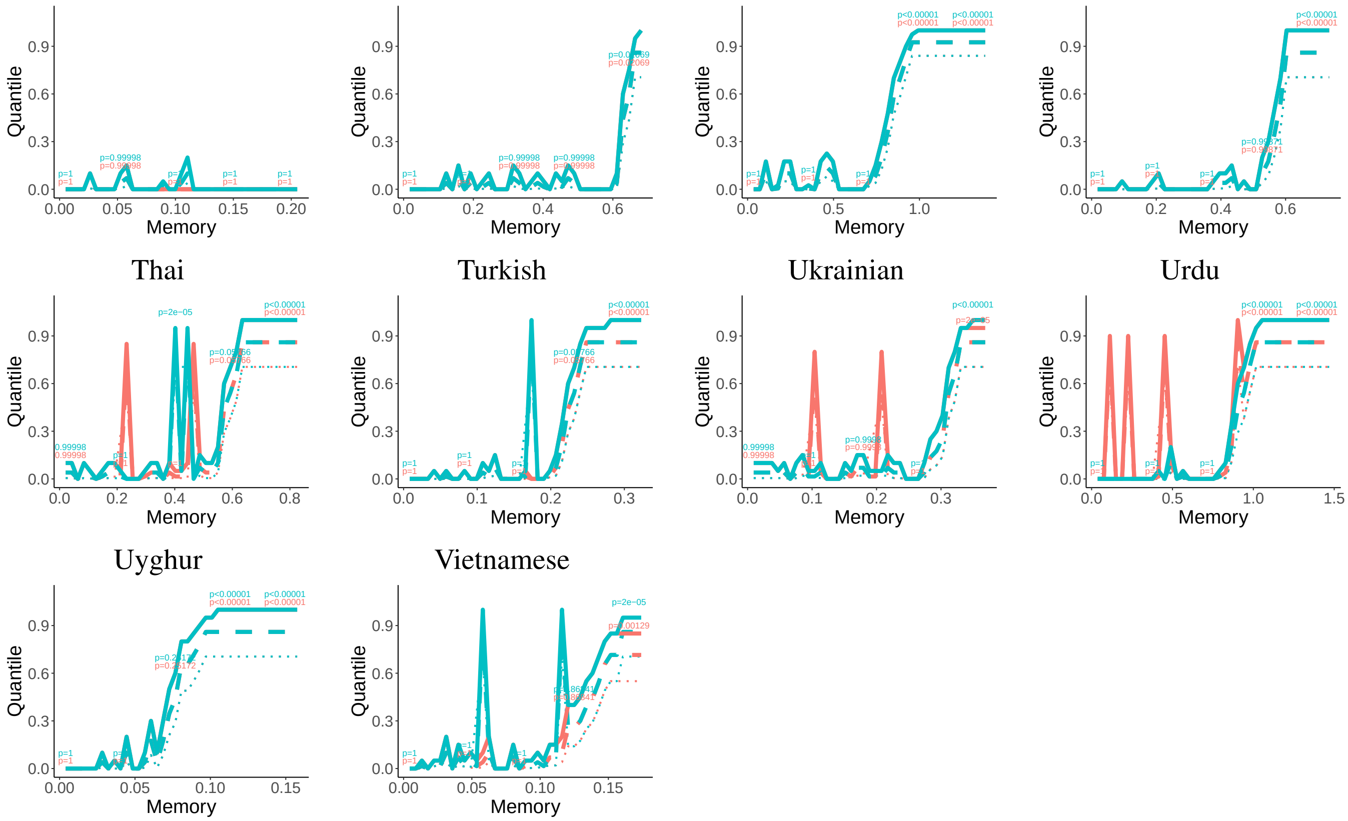


Figure 15: Quantiles: At a given memory budget, what percentage of the baselines results in higher listener surprisal than the real language? Solid curves represent sample means, dashed lines represent 95 % confidence bounds; dotted lines represent 99.9 % confidence bounds. At five evenly spaced memory levels, we provide a p-value for the null hypothesis that the actual population mean is 0.5 or less. Confidence bounds and p-values are obtained using an exact nonparametric method (see text). {tab:quantil

References

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Lewis, R. L. and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of psycholinguistic research*, 29(2):111–123.

Nicenboim, B. and Vasishth, S. (2018). Models of retrieval in sentence comprehension: A computational evaluation using bayesian hierarchical modeling. *Journal of Memory and Language*, 99:1–34.

Vasishth, S., Nicenboim, B., Engelmann, F., and Burchert, F. (2019). Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences*.