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

Michael Hahn, Judith Degen, Richard Futrell

2018

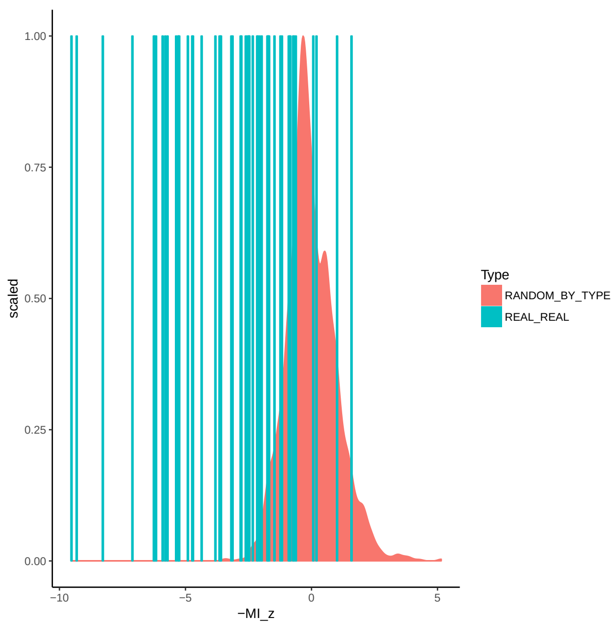


Figure 1: Histogram

{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.

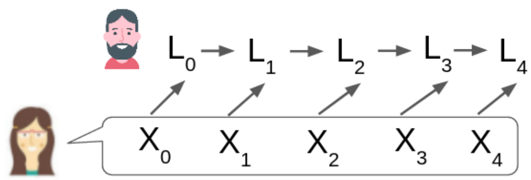


Figure 2: Illustration of (3). As the utterance unfolds, the listener maintains a memory state. After receiving word w_t, the listener computes their new memory state m_t based on the previous memory state m_{t-1} and the new word w_t.

{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 We now analyze memory from the perspective of the listener, who needs to maintain information about the past to predict the future. As the speaker's utterance unfolds, the listener maintains a memory state m_t.

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}. 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}. This is formalized as a statement about conditional probabilities:

$$p(m_1|(w_t)_{t\in\mathbb{Z}}, m_0) = p(m_1|m_0, w_1) \tag{1}$$

This says that m_1 contains no information about the utterances beyond what is contained in m_0 and w_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 ($T \in \{1,2,3,...\}$), and consider a listener using at most*

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

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

$$p(m_{t+1}|(w_{t'})_{t'\in\mathbb{Z}}, m_t) = p(m_{t+1}|m_t, w_t) \tag{3}$$

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

$$\mathrm{H}[w_t|m_t] - \mathrm{H}[w_t|w_{<t}] = \frac{1}{T}\sum_{t'=1}^{T}\left(\mathrm{H}[w_{t'}|m_{t'}] - \mathrm{H}[w_{t'}|w_{<t'}]\right) \tag{4}$$

We first show a lemma:

Lemma 2. *For any positive integer t, the following inequality holds:*

$$H[w_t|m_t] \geq H[w_t|w_{1\dots t-1}, m_1] \tag{5}$$

Proof of the Lemma. By Bayes' Theorem

$$p(w_t|m_0, m_1, w_{0\dots t-1}) = \frac{p(m_1|m_0, w_{0\dots t})}{p(m_1|m_0, w_{0\dots t-1})} \cdot p(w_t|m_0, w_{0\dots t-1})$$

By Equation 3, the quotient on the RHS is equal to 1, so

$$p(w_t|m_0, m_1, w_{0\dots t-1}) = p(w_t|m_0, w_{0\dots t-1})$$

So we have a Markov chain

$$(w_t) \to (m_0, w_{0\dots t-1}) \to (m_1, w_{1\dots t-1}) \tag{6}$$

Thus, by the Data Processing Inequality,

$$H[w_t|w_{1\dots t-1}, m_1] \geq H[w_t|w_{0\dots t-1}, m_0] \tag{7}$$

Finally, iteratively applying this inequality, we get:

$$H[w_t|m_t] \geq H[w_t|w_{t-1}, m_{t-1}] \geq H[w_t|w_{t-2,t-1}, m_{t-2}] \geq \dots \geq H[w_t|w_{1\dots t-1}, m_1]$$

$\square$

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

Plugging this inequality into Equation 4 above:

$$\begin{aligned}
\mathrm{H}[w_t|m_t] - \mathrm{H}[w_t|w_{<t}] &\geq \frac{1}{T}\sum_{t=1}^{T}(\mathrm{H}[w_t|w_{1\ldots t-1}, m_1] - \mathrm{H}[w_t|w_{1\ldots t-1}, w_{\leq 0}]) \\
&= \frac{1}{T}\left(\mathrm{H}[w_{1\ldots T}|m_1] - \mathrm{H}[w_{1\ldots T}|w_{\leq 0}]\right) \\
&= \frac{1}{T}\left(I[w_{1\ldots T}|w_{\leq 0}] - I[w_{1\ldots T}|m_1]\right)
\end{aligned}$$

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

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

Therefore

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

$I[w_{1\ldots T}|m_1]$ 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. We describe an information-theoretic formalization of a model that has both a limited WM and an unlimited STM.

We consider a model that maintains both a small working memory m_t and an unlimited short-term memory s_t.

Predictions are made based on working memory m_t, incurring surprisal $H[w_t|m_t]$. In each time step, 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.

This is instantiated by ACT-R. We can explicitly explain how this covers the McElree ideas and the Lewis and Vasishth ACT-R model.

This model has two bottlenecks: The working memory capacity, which we model as $H[m_t]$, and the amount of information that is added through retrieval, modeled as $H[r_t|m_t]$.

Theorem 3. *Assume the average working memory $H[m_t]$ is bounded as $H[L_t] \leq \sum_{t=1}^{T} tI_t$, and the average amount of retrieval is bounded as $H[R_t] \leq \sum_{t=T+1}^{S} I_t$ (per word). Then $H[w_t|m_t] \geq H[w_t|x_{<t}] + \sum_{t>S} I_t$.*

Proof. The negative surprisal gap is $\leq \frac{1}{T}(I[X_1 \dots X_T | (M_0, R_1, ..., R_T)] - \sum_{t=1}^{T} tI_t - T\sum_{t>T} I_t) \leq \frac{1}{T}(T \cdot H[R_t] - T\sum_{t>T} I_t) = H[R_t] - \sum_{t>T} I_t$

□

We can also think of this as a multi-objective optimization, aiming to minimize $\lambda_1 WM + \lambda_2 Retrieval$ to achieve a given surprisal level.

If $\frac{\lambda_2}{\lambda_1} \rightarrow \infty$ (retrievals get more expensive), recover previous model.

If $\frac{\lambda_2}{\lambda_1} \rightarrow 0$ (retrievals get cheaper), locality effect gets weaker, and disappears in the limit[3]

Objective

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

has solution $T \approx \frac{\lambda_2}{\lambda_1}$.[4]

Consequence: As long as retrievals are more expensive than keeping the same amount in WM, locality is predicted.

2 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
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

[3](Of course, even in this limit, there might be additional factors that may still favor locality in a specific implementation of memory – e.g., in ACT-R, decay and interference are less problematic if there is locality.)

[4]Can do simple proof using the continuous-T-version.

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

3 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}

4 Detailed Results per Language

4.1 Median Surprisal per Memory Budget

Afrikaans Amharic Arabic Armenian

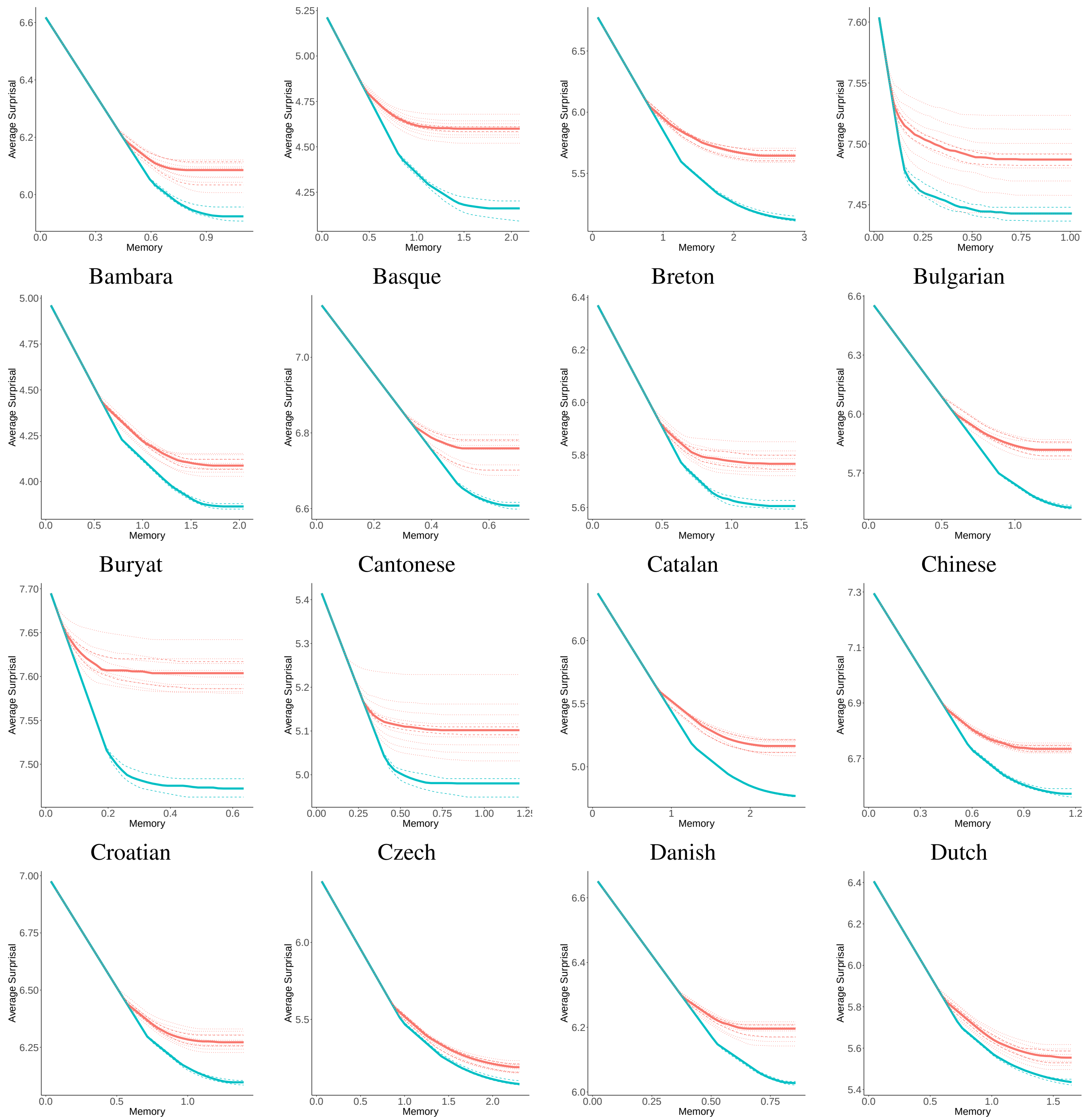


Figure 5: Medians: For each memory budget, we provide the median surprisal for real and random languages. Solid lines indicate sample medians, dashed lines indicate 95 % confidence intervals for the population median, dotted lines indicate empirical quantiles (10%, 20%, . . . , 80%, 90%). Green: Random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

{tab:medians

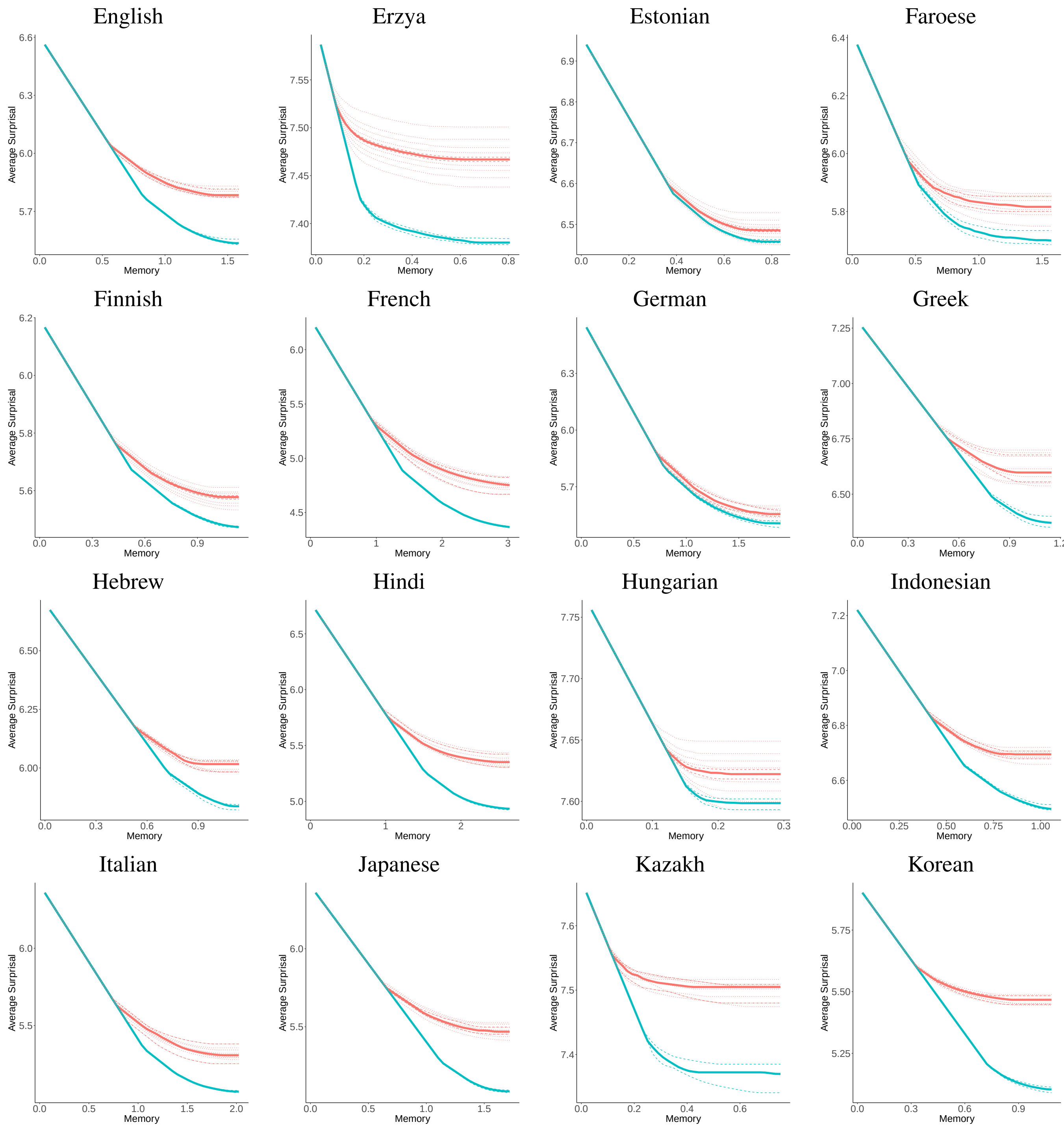


Figure 6: Medians (cont.)

Kurmanji Latvian Maltese Naija

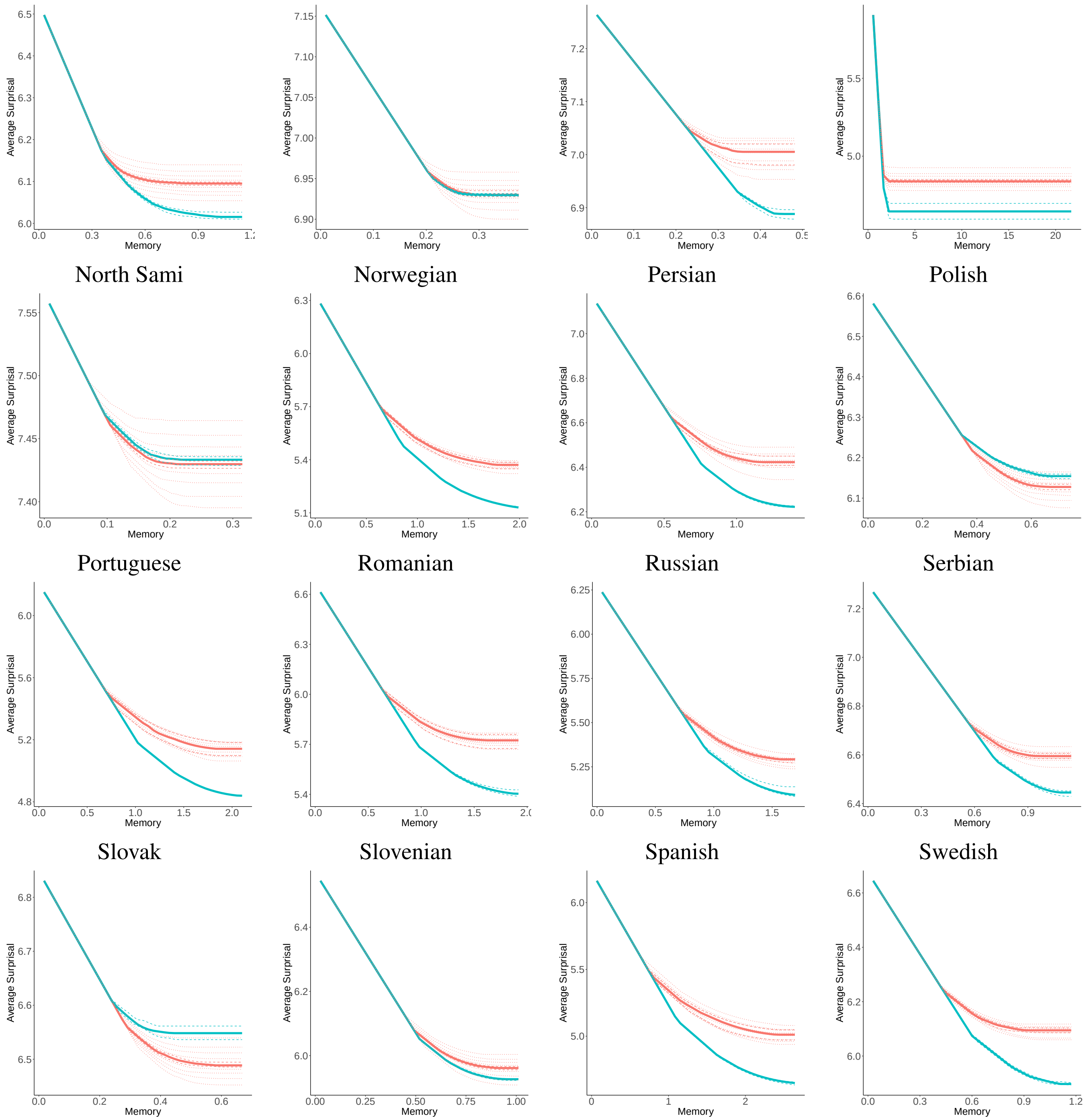
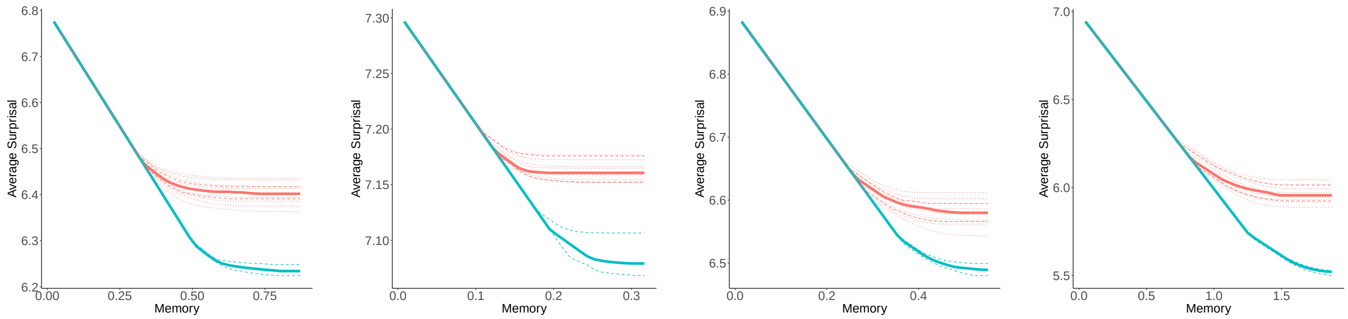
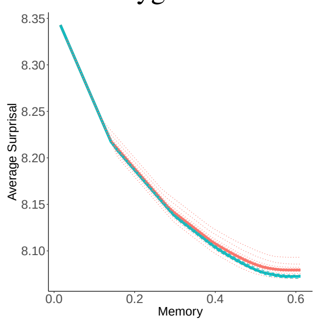


Figure 7: Medians (cont.)

Thai Turkish Ukrainian Urdu



Uyghur



Vietnamese

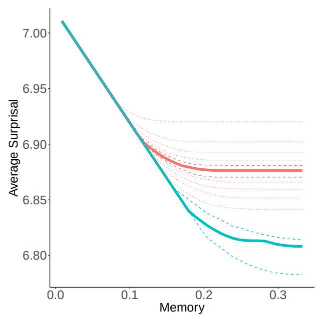


Figure 8: Medians (cont.)

4.2 Surprisal at Maximum Memory

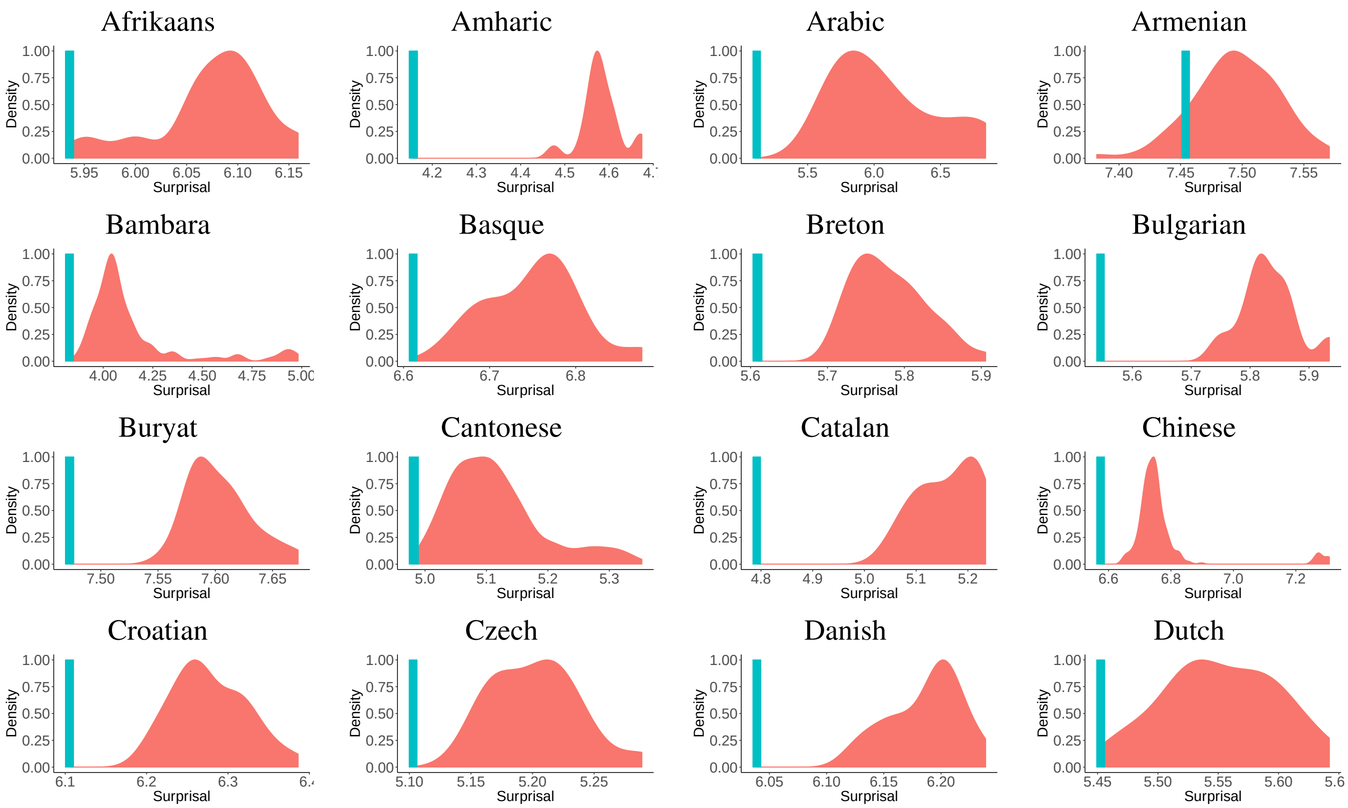


Figure 9: Histograms: Surprisal, at maximum memory.

{tab:slice-h

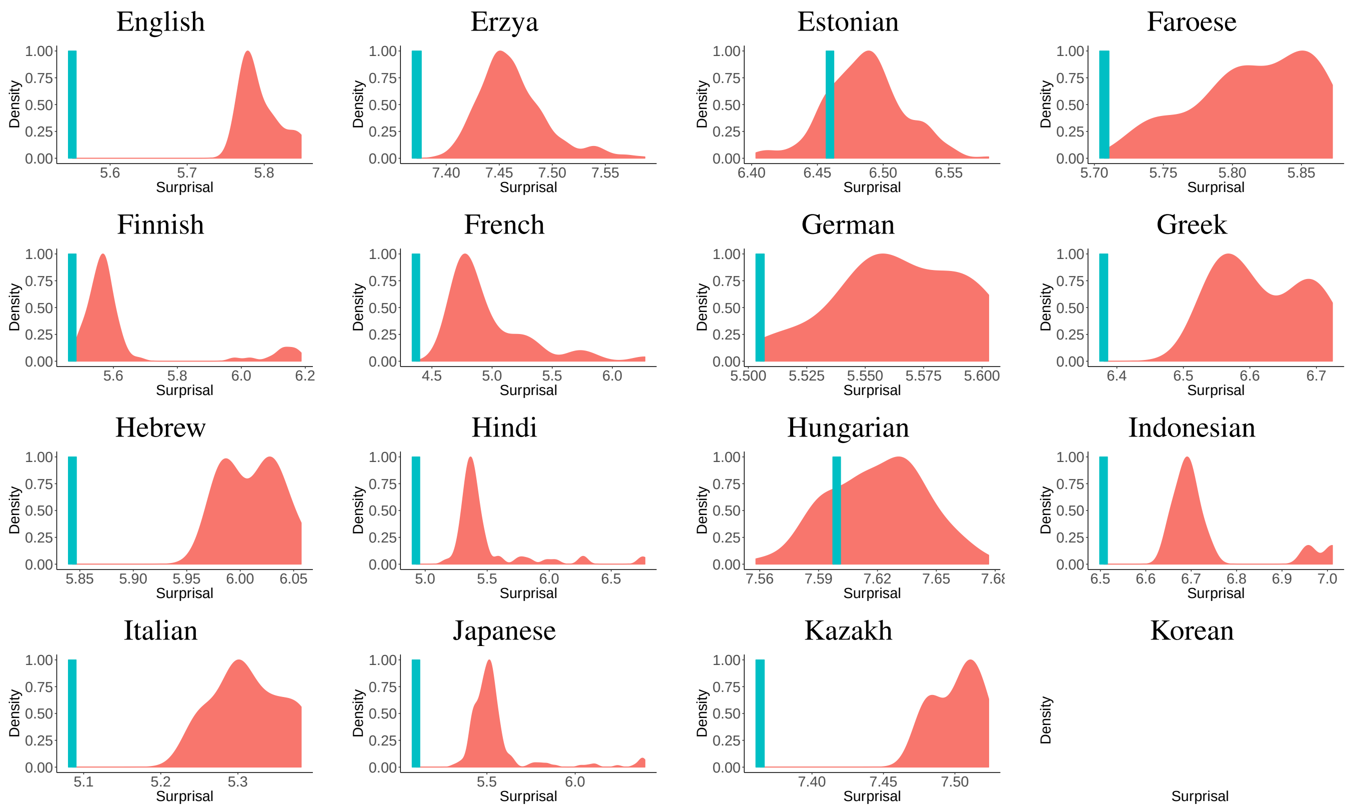


Figure 10: Medians (cont.)

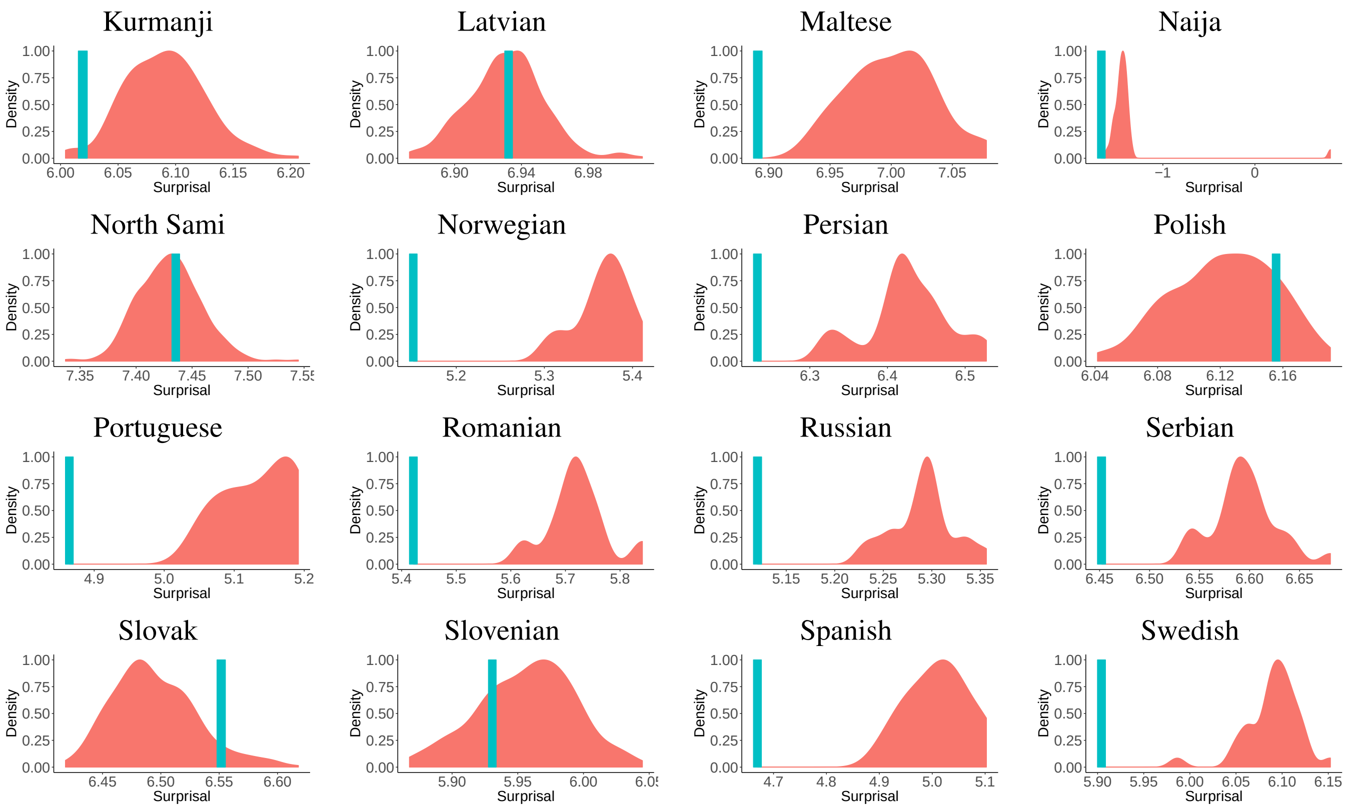


Figure 11: Medians (cont.)

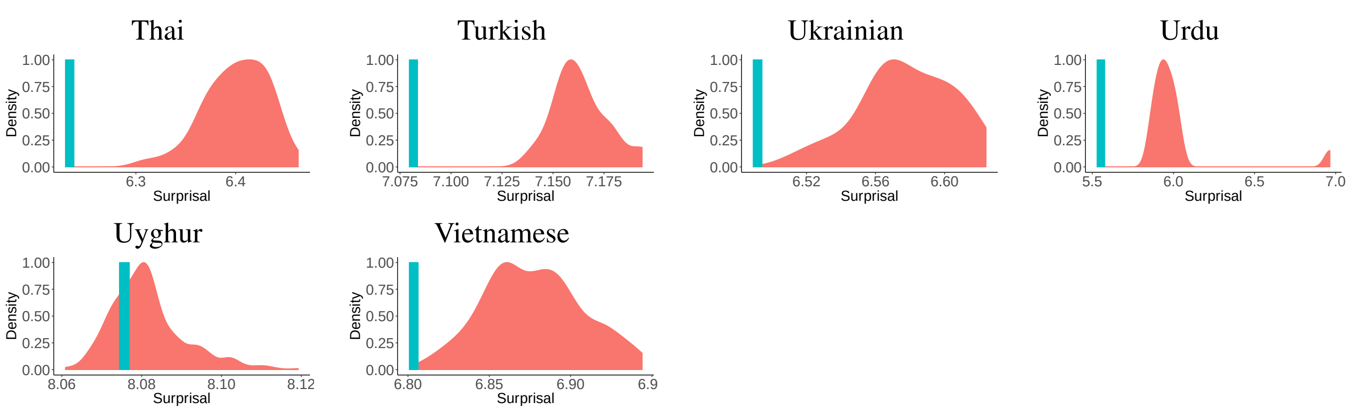


Figure 12: Medians (cont.)

4.3 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 13: Experiment 3: Samples drawn per language according to the precision-dependent stopping criterion. {tab:samples

4.4 Medians (Experiment 3)

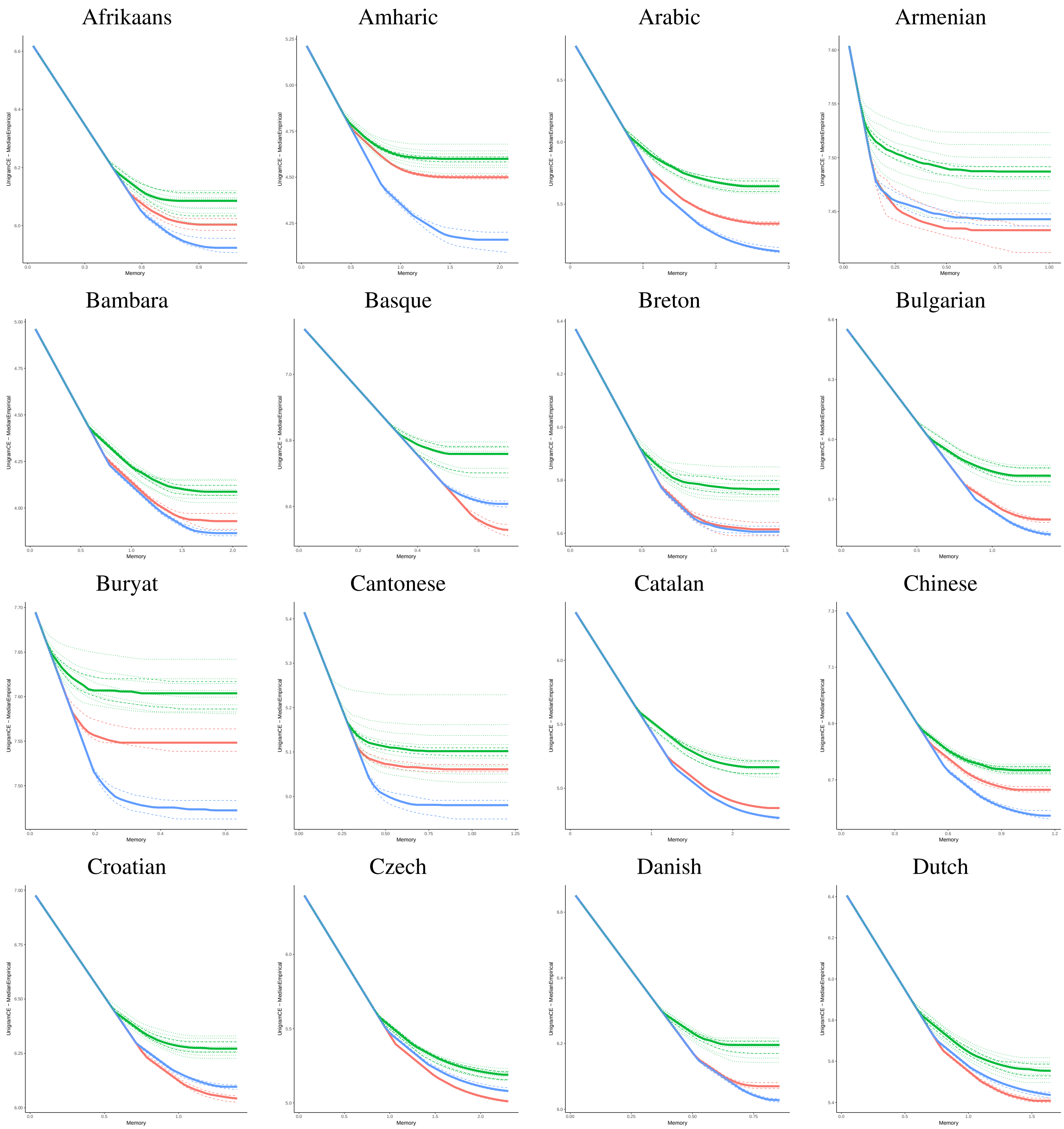


Figure 14: Experiment 3. Medians: For each memory budget, we provide the median surprisal for real and random languages. Solid lines indicate sample medians, dashed lines indicate 95 % confidence intervals for the population median. Green: Random baselines; blue: real language; red: maximum-likelihood grammars fit to real orderings.

{tab:medians

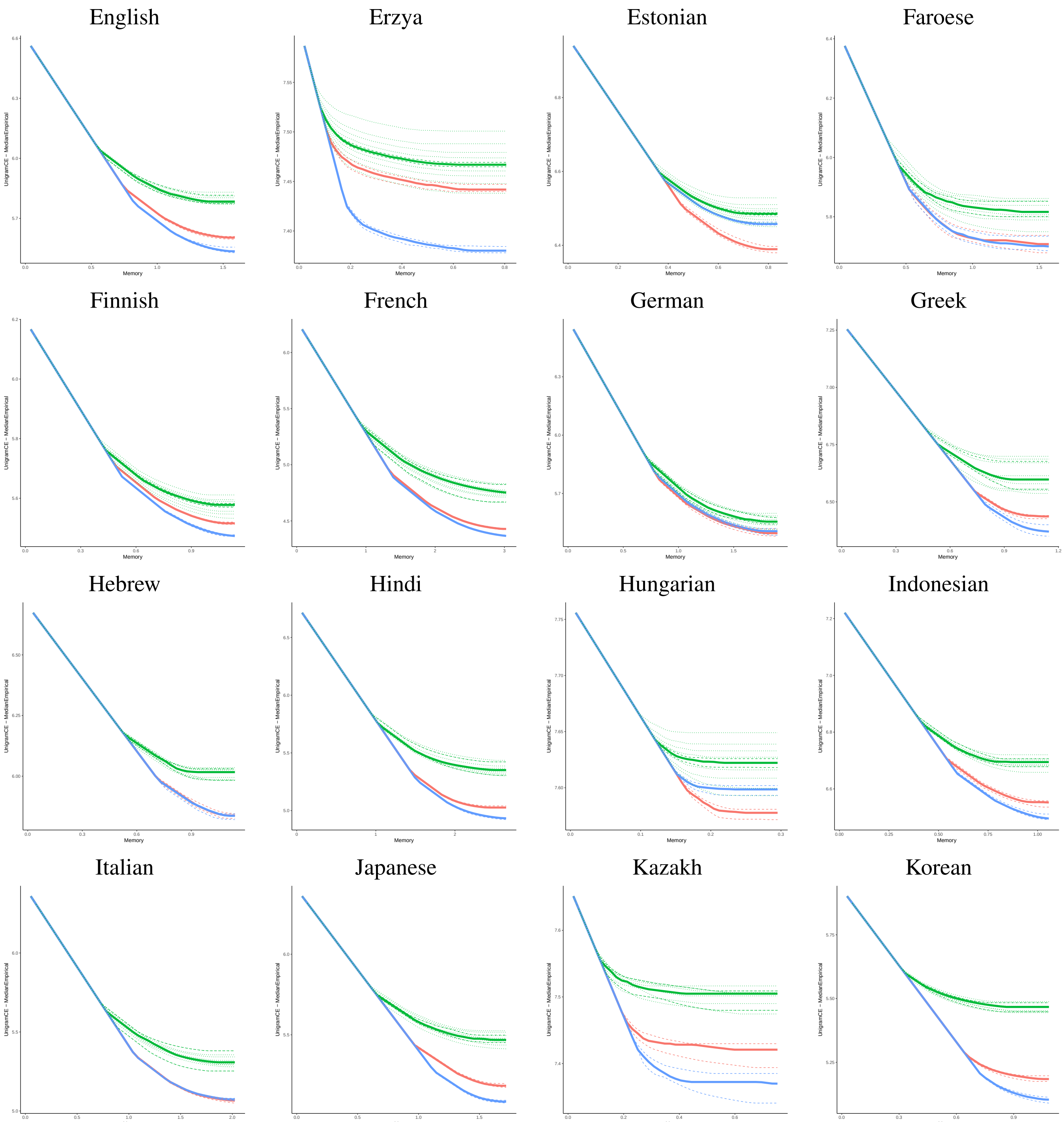


Figure 15: Medians (cont.)

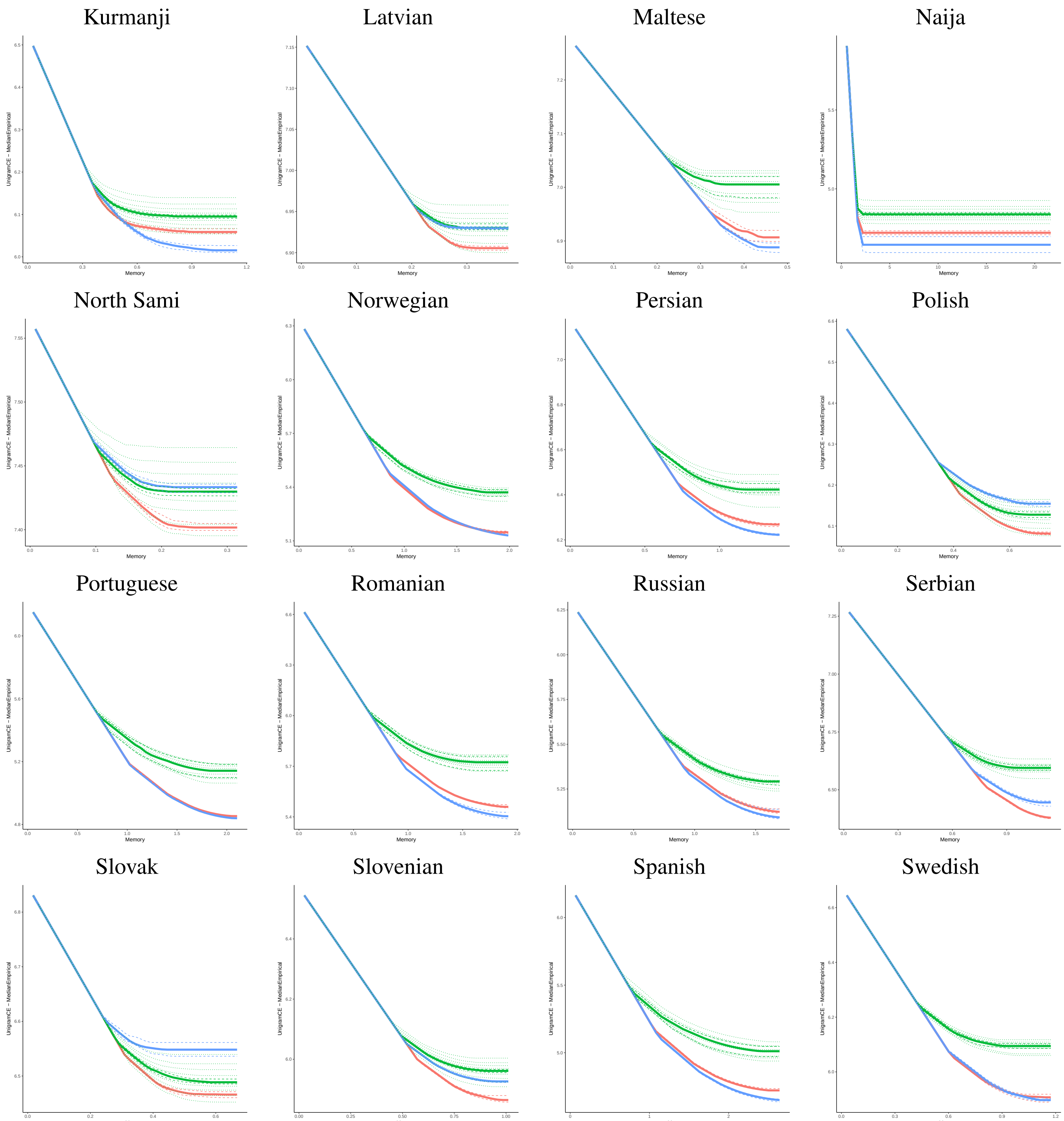


Figure 16: Medians (cont.)

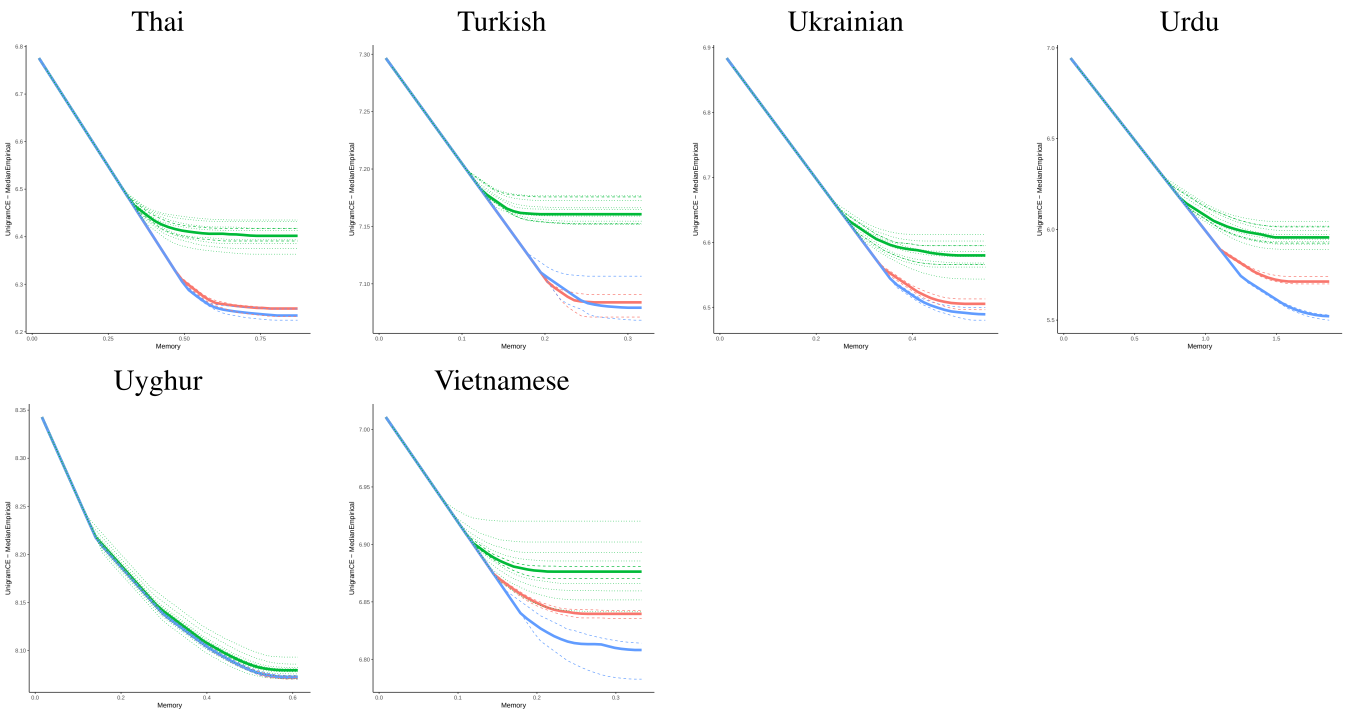


Figure 17: Medians (cont.)

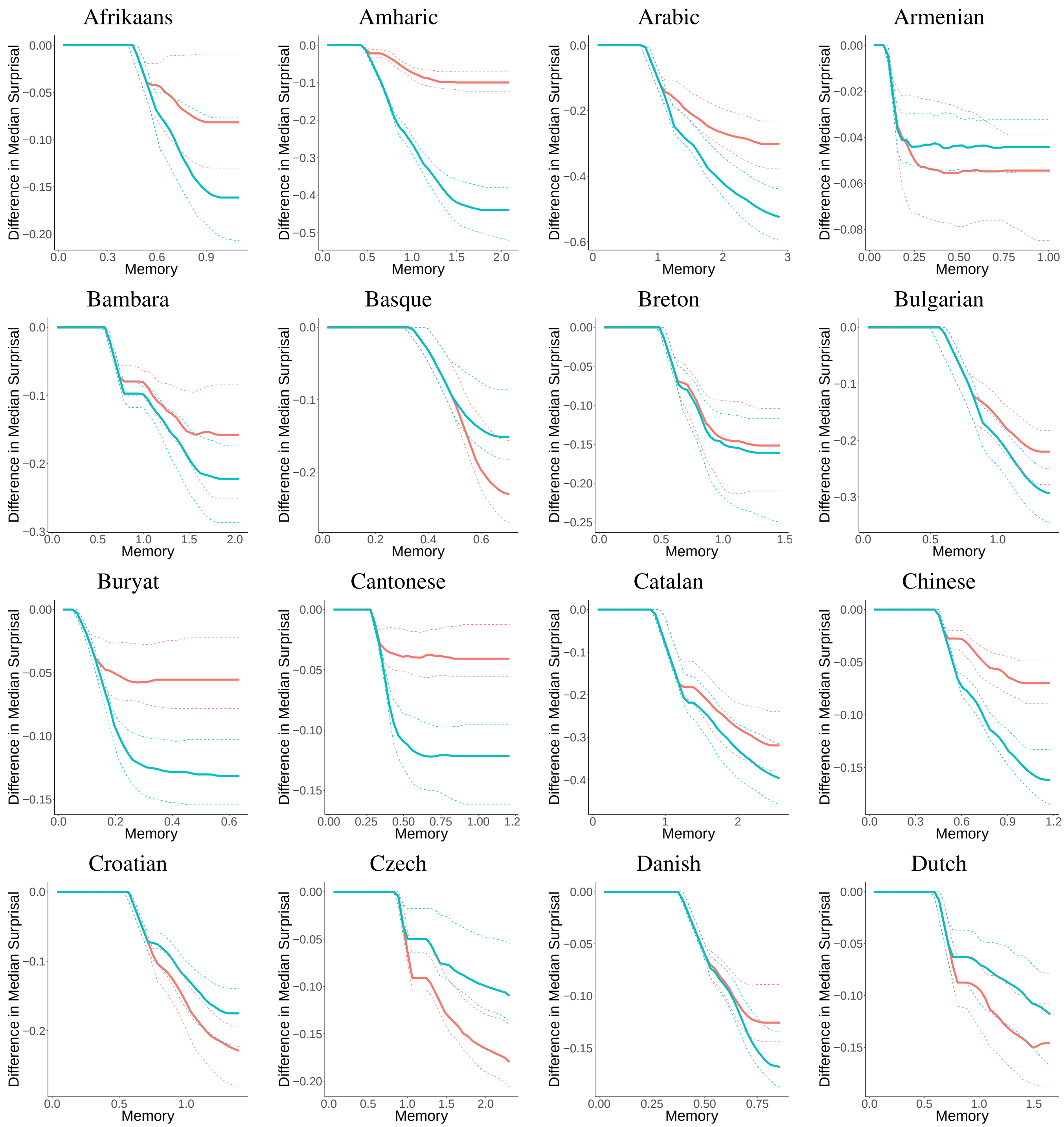


Figure 18: Median Differences between Real and Baseline: For each memory budget, we provide the difference in median surprisal between real languages and random baselines; for real orders (blue) and maximum likelihood grammars (red). Lower values indicate lower surprisal compared to baselines. Solid lines indicate sample means. Dashed lines indicate 95 % confidence intervals. {tab:median_

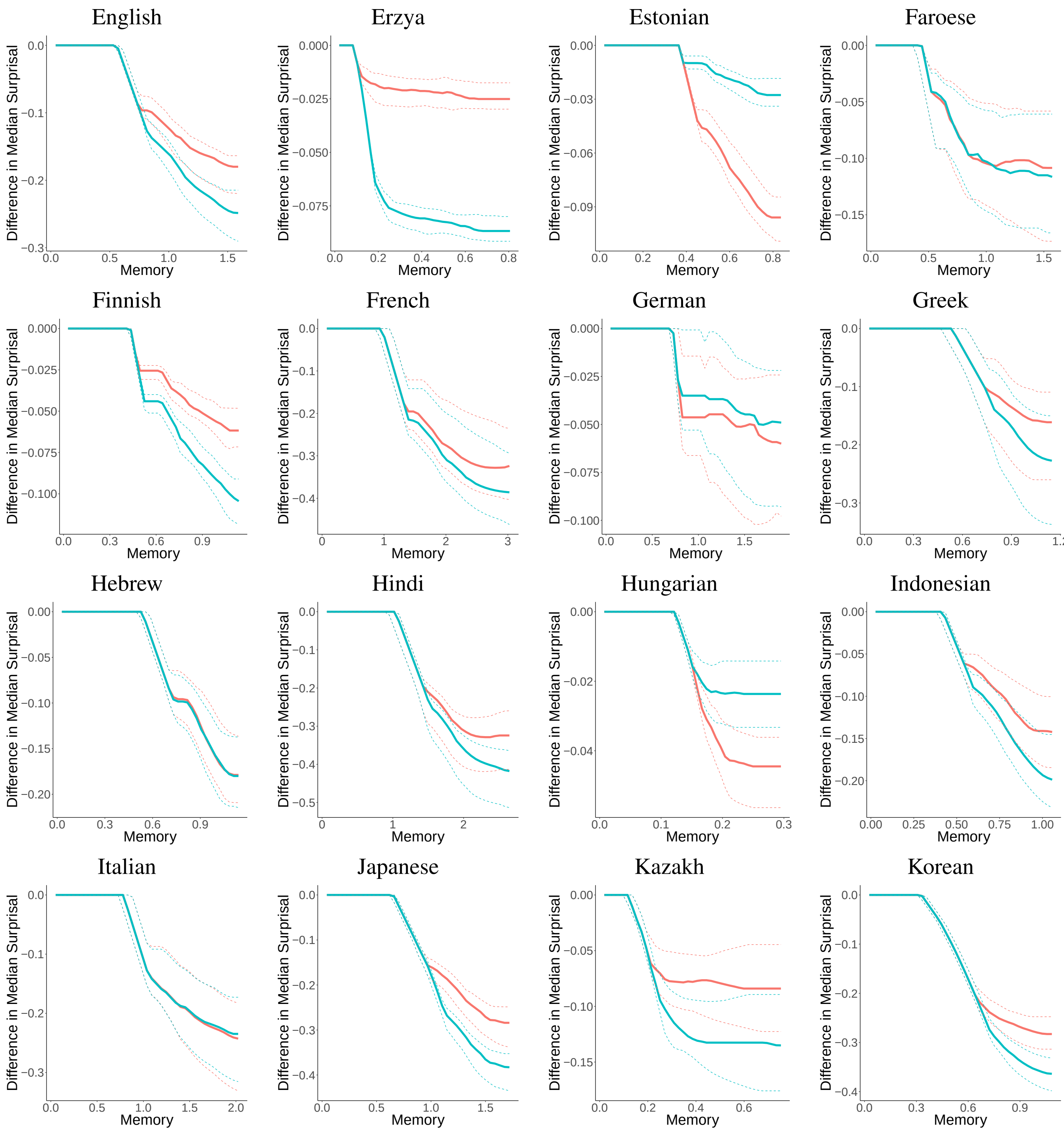


Figure 19: Median Differences (Part 2)

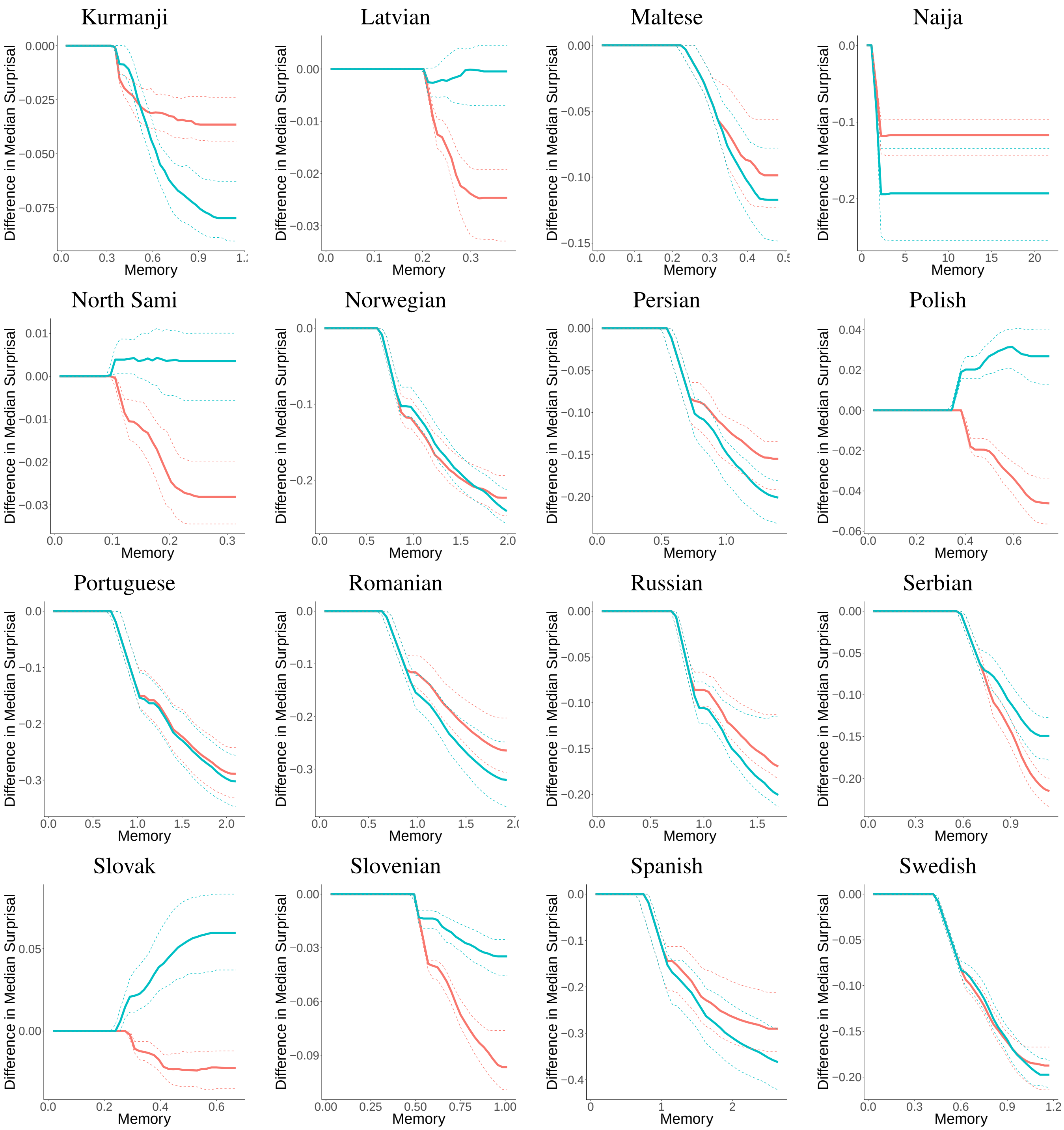


Figure 20: Median Differences (Part 3)

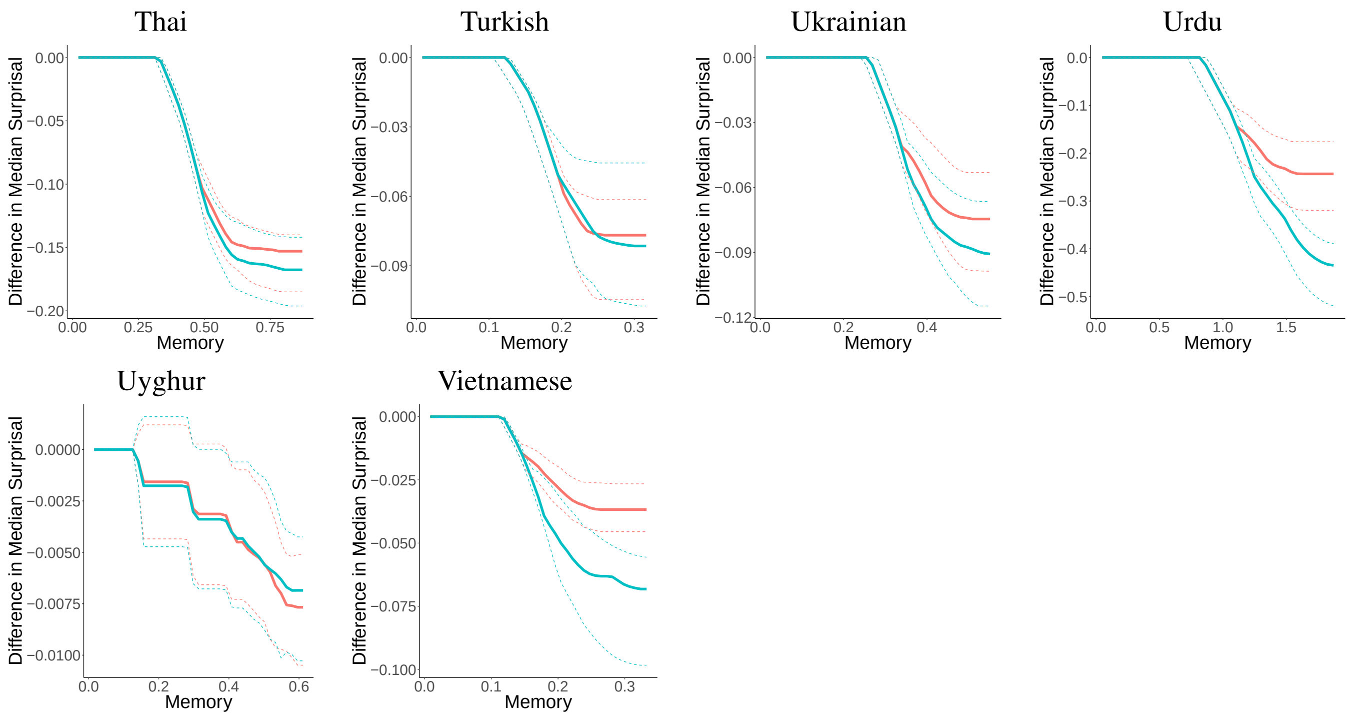


Figure 21: Median Differences (Part 4)

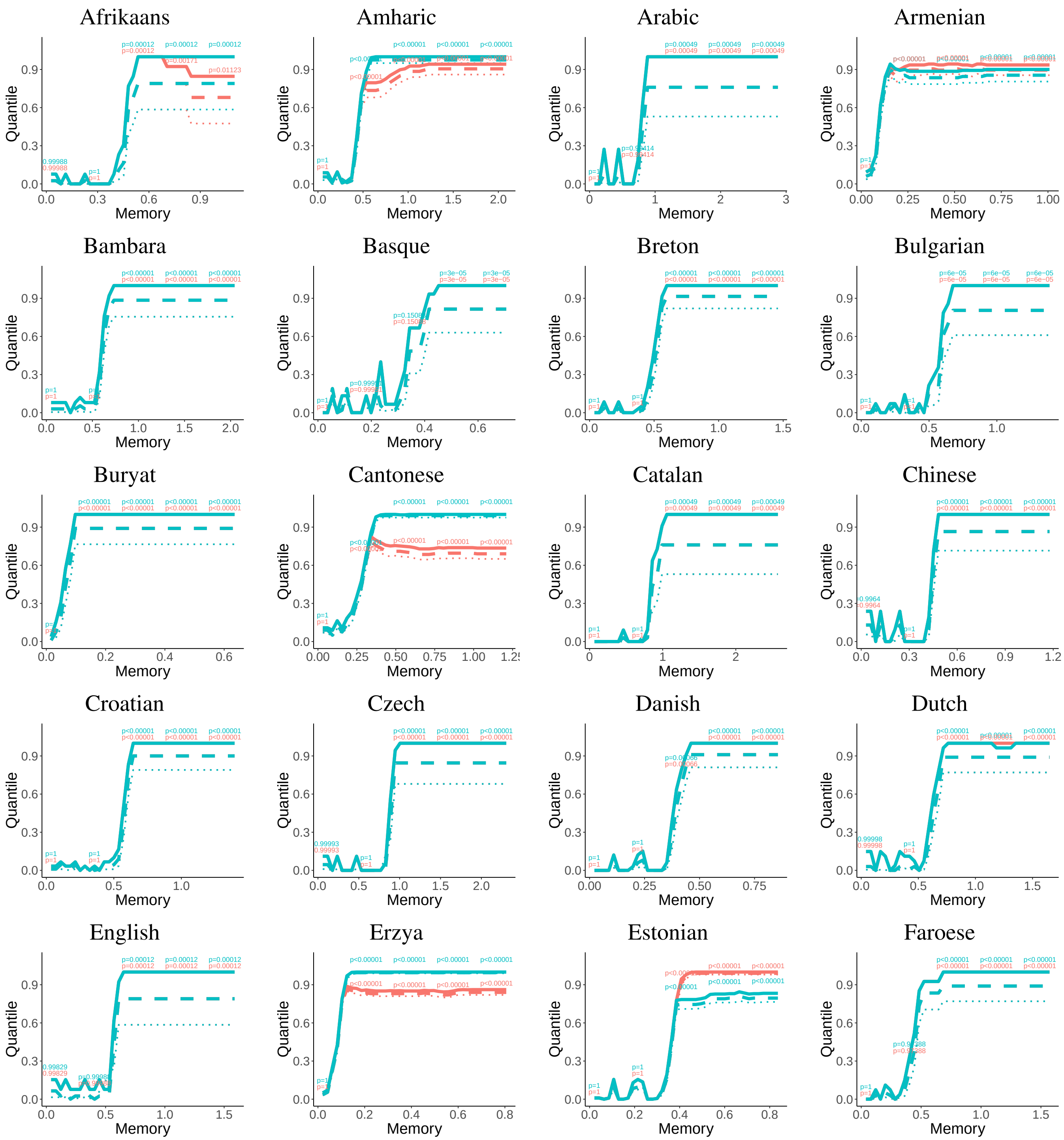


Figure 22: Quantiles: At a given memory budget, what percentage of the baselines results in higher listener surprisal than the real language? Solid curves represent sample means, dashed lines represent 95 % confidence bounds; dotted lines represent 99.9 % confidence bounds. At five evenly spaced memory 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

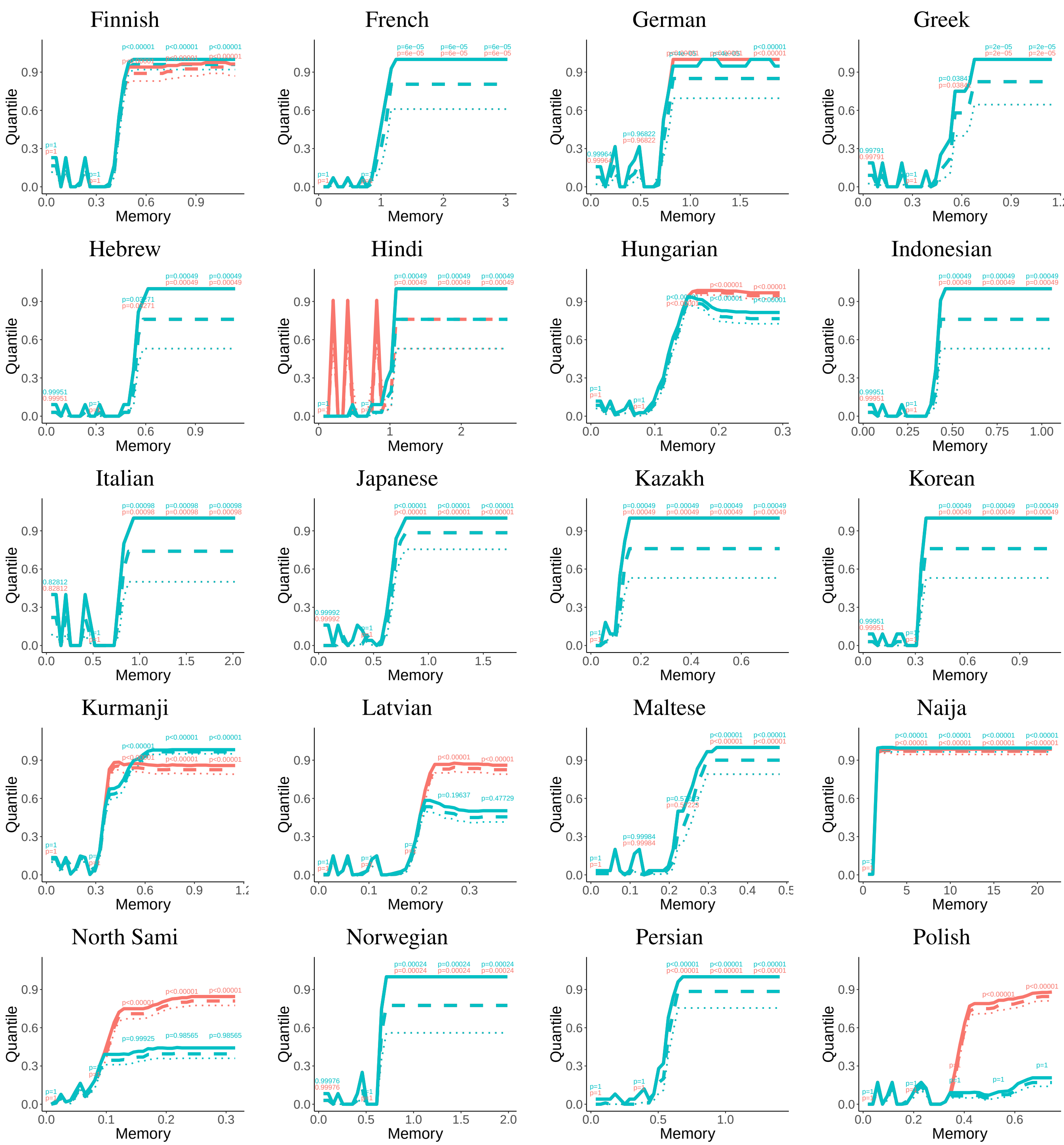


Figure 23: Quantiles (part 2)

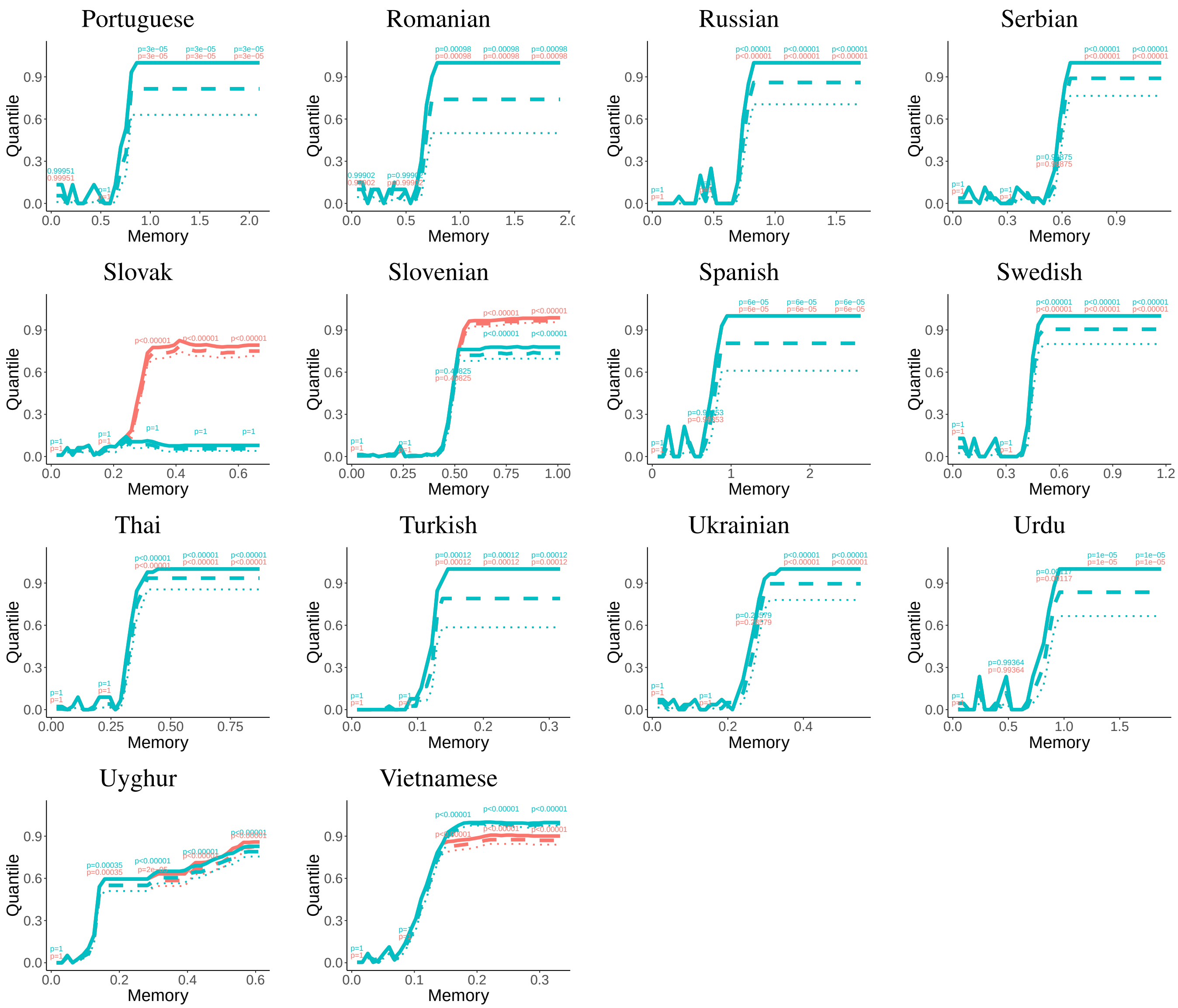


Figure 24: Quantiles (part 3)

5 Details for Neural Network Models

6 N-Gram Models

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

$$p_1(w_t) := \frac{N(w_t) + \delta}{|Train| + |V| \cdot \delta} \tag{8}$$

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 recur-

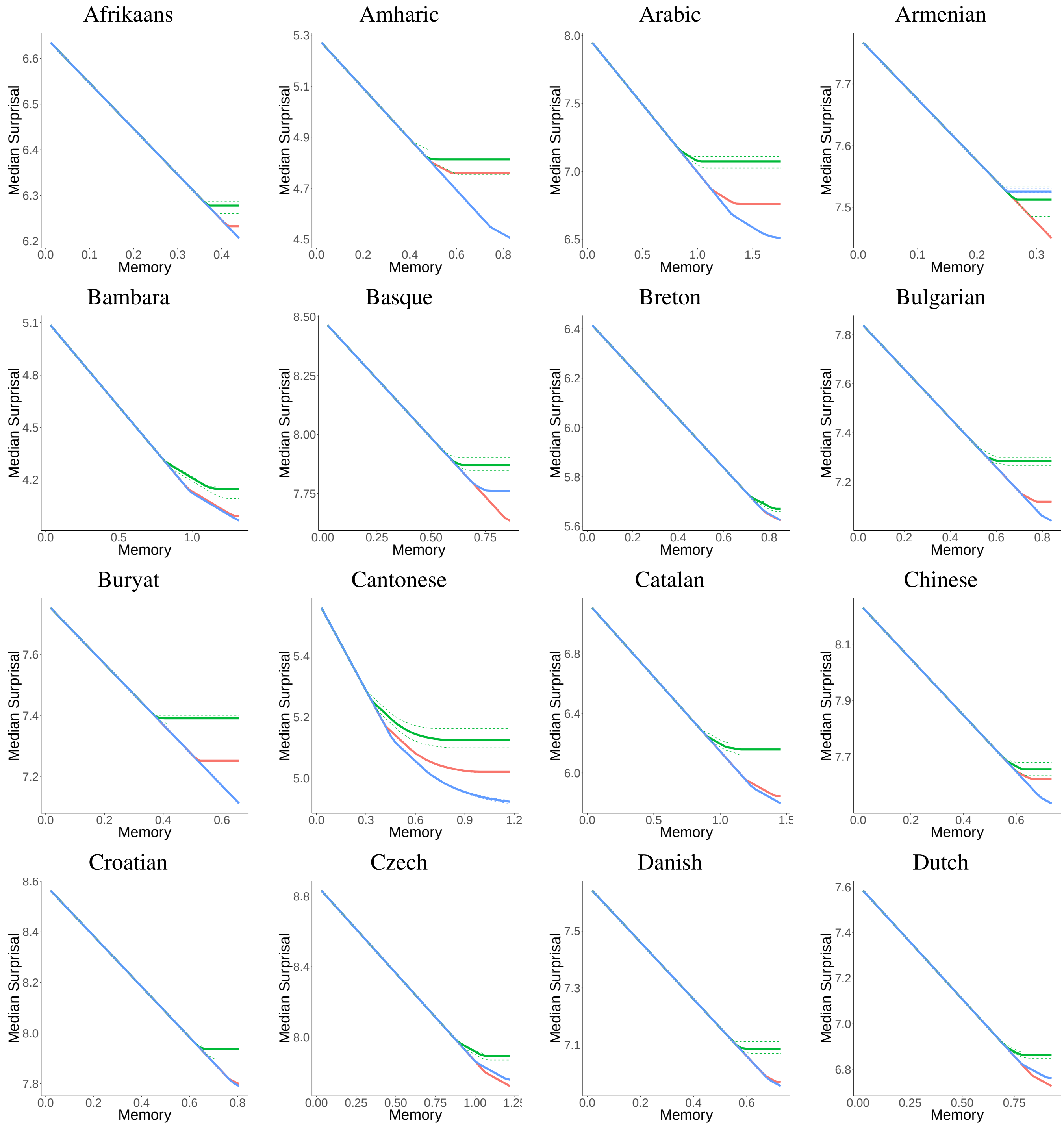


Table 9: Medians (estimated using n-gram models): For each memory budget, we provide the median surprisal for real and random languages. Solid lines indicate sample medians for ngrams, dashed lines indicate 95 % confidence intervals for the population median. Green: Random baselines; blue: real language; red: maximum-likelihood grammars fit to real orderings.

{tab:medians

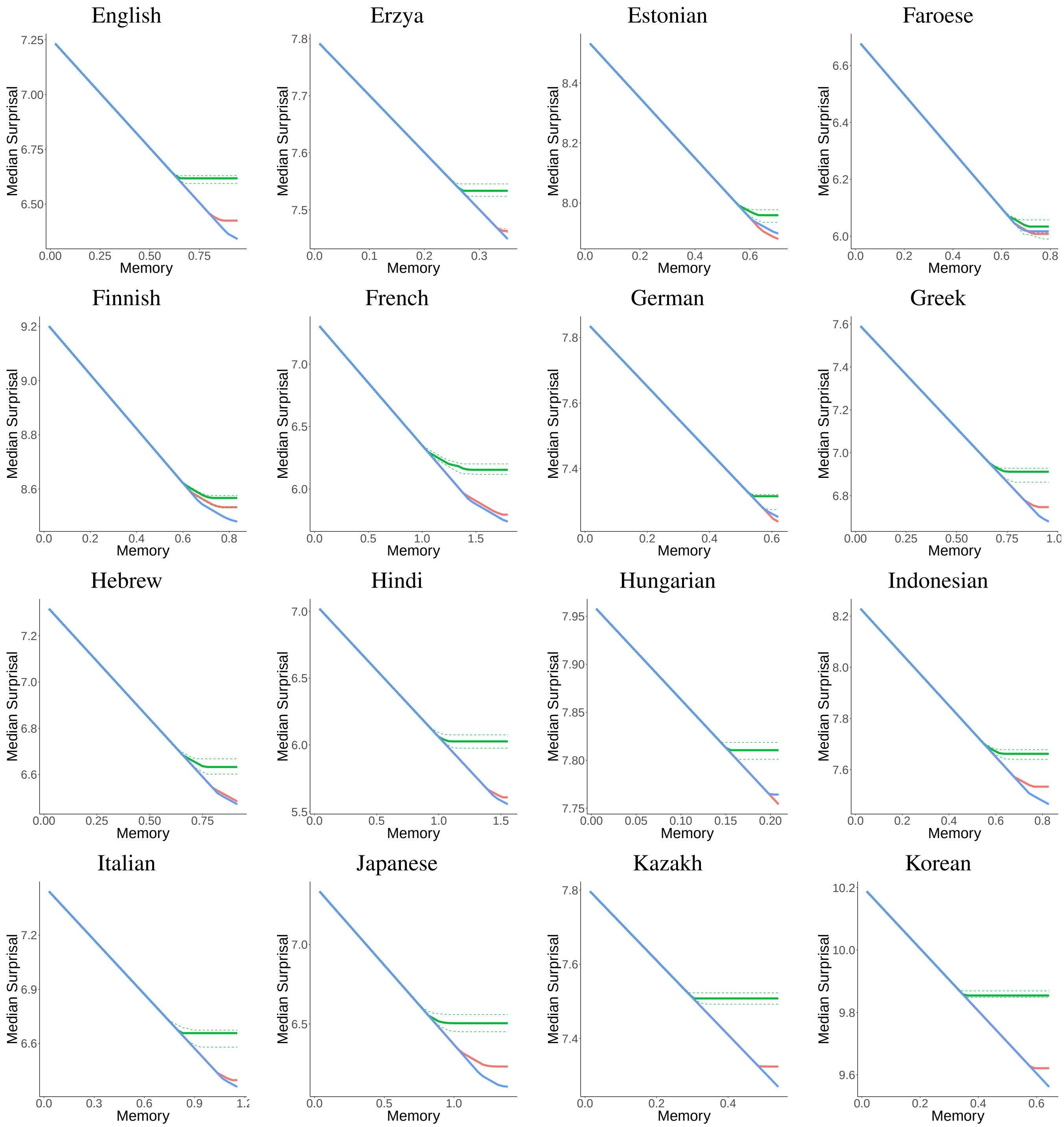


Table 10: Medians (cont.)

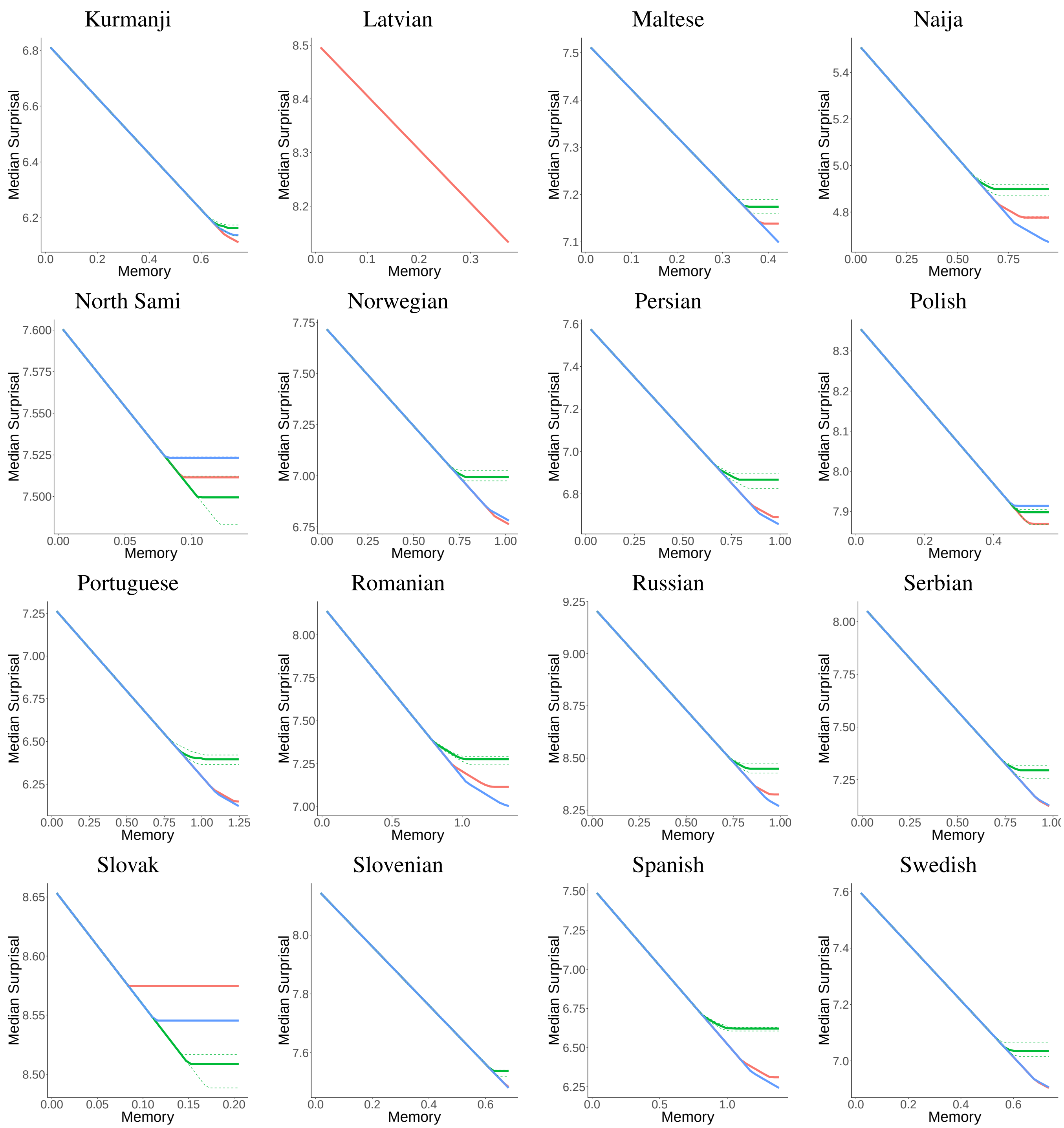


Table 11: Medians (cont.)

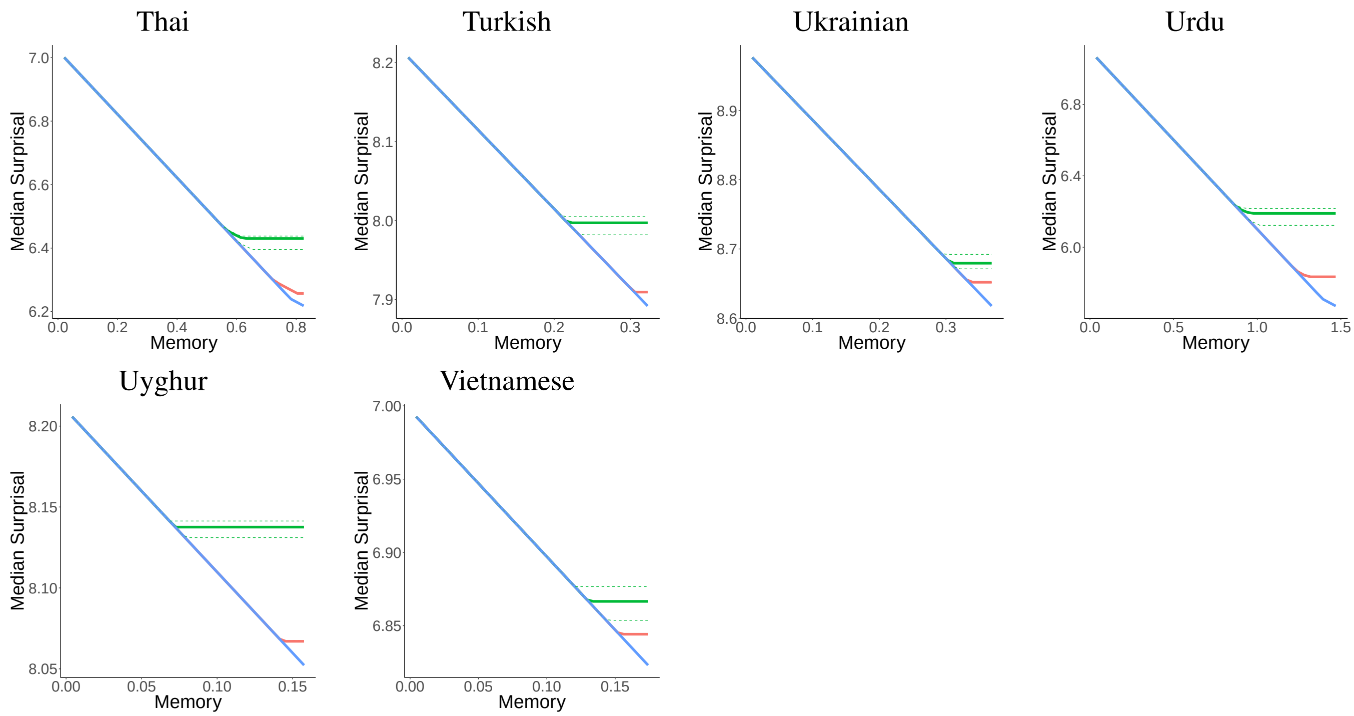


Table 12: Medians (cont.)

sively as follows. Let $\gamma > 0$ be a hyperparameter. If $N(w_{0...t-1}) < \gamma$, set $p_t(w_t|w_{0...t-1}) := p_{t-1}(w_t|w_{1...t-1})$. Otherwise, we interpolate between n-th order and lower-order estimates:

$$p_t(w_t|w_{0...t-1}) := \frac{\max(N(w_{0...t}) - \alpha, 0.0) + \alpha \cdot \#\{w : N(w_{0...t-1}w) > 0\} \cdot p_{t-1}(w_t|w_{1...t-1})}{N(w_{0...t-1})} \tag{9}$$

where $\alpha \in [0, 1]$ is also a hyperparameter.

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

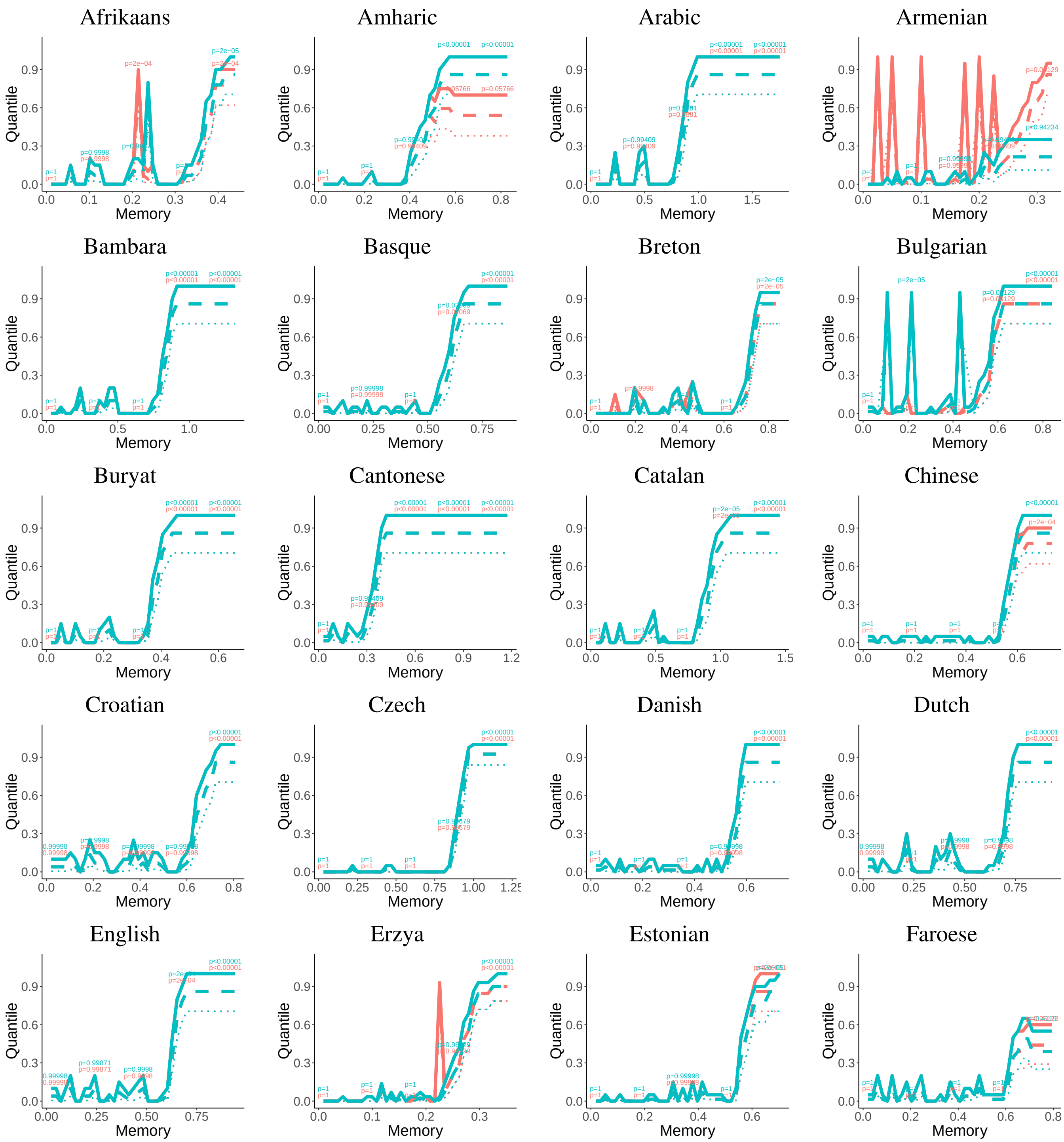


Table 13: Quantiles: At a given memory budget, what percentage of the baselines results in higher listener surprisal than the real language? Solid curves represent sample means, dashed lines represent 95 % confidence bounds; dotted lines represent 99.9 % confidence bounds. At five evenly spaced memory 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

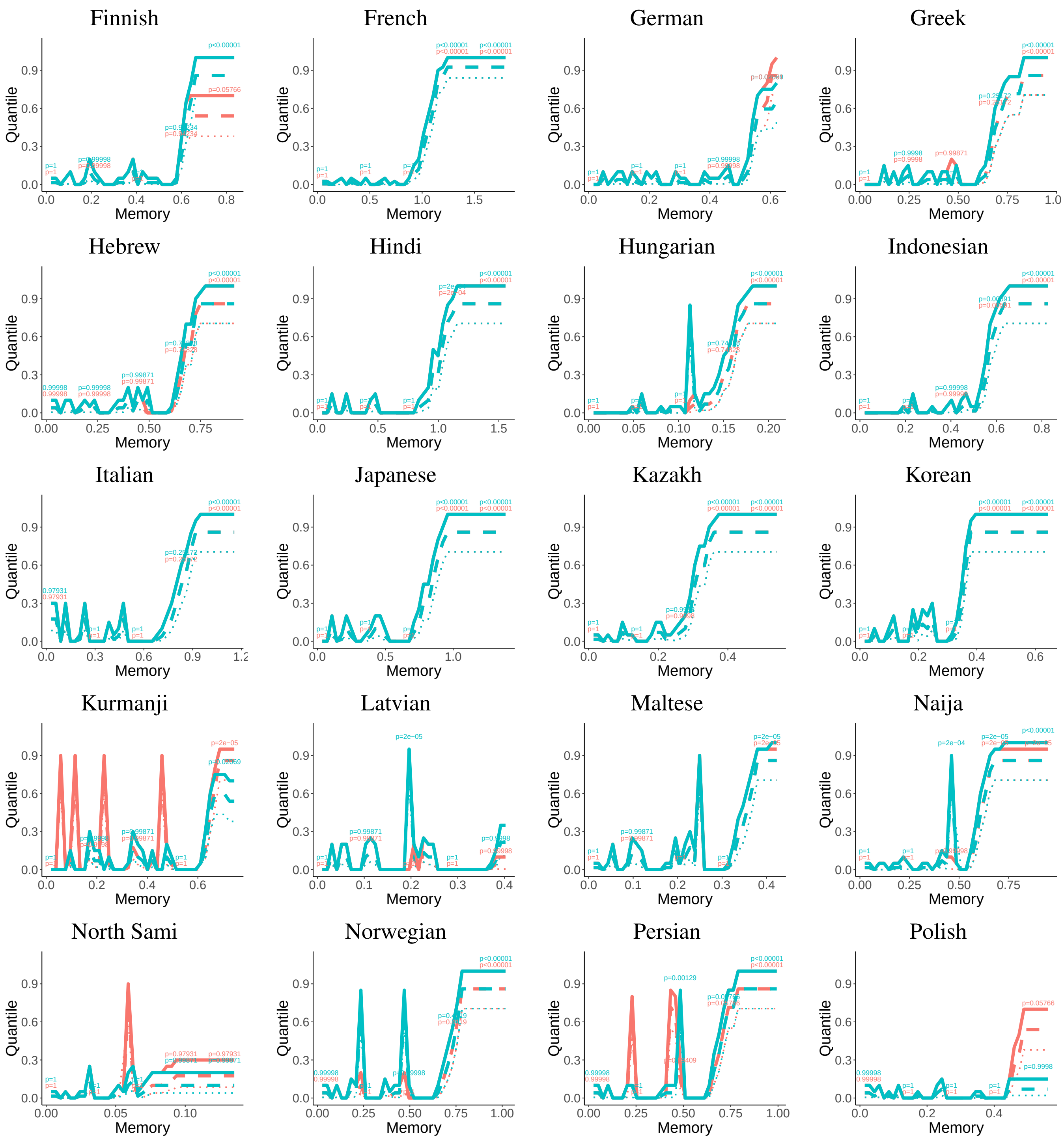


Table 14: Quantiles (part 2)

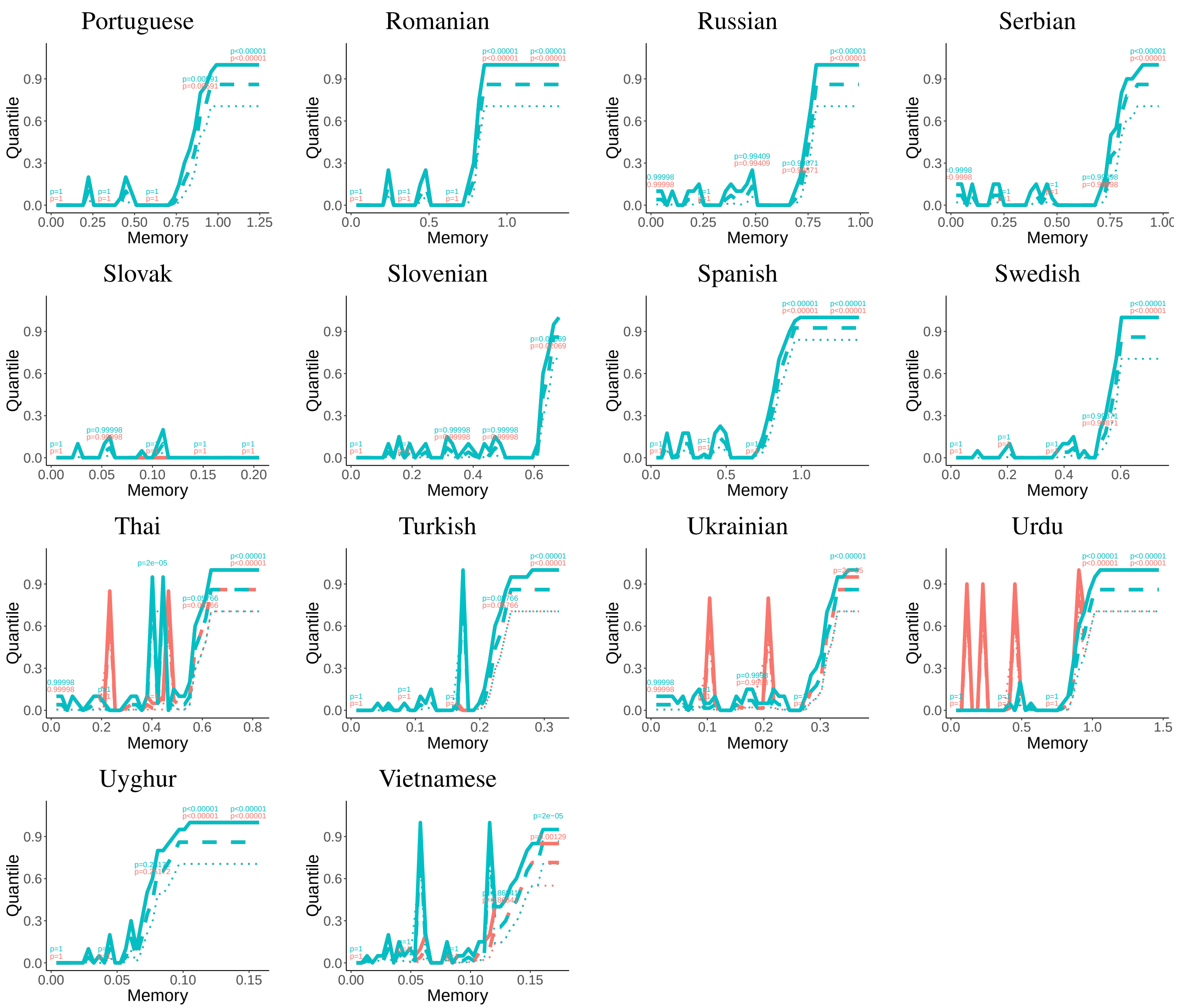


Table 15: Quantiles (part 3)