

Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>

Contents lists available at [ScienceDirect](http://www.sciencedirect.com)

Vision Research

journal homepage: www.elsevier.com/locate/visres

Human efficiency for classifying natural versus random text

Peter Neri ^{a,*}, Alicia Liu ^b, Dennis M. Levi ^b^a *Institute of Medical Sciences, University of Aberdeen, Foresterhill, Aberdeen AB25 2ZD, United Kingdom*^b *School of Optometry, 488 Minor Hall, UC Berkeley, Berkeley, CA 94720-2020, United States*

ARTICLE INFO

Article history:

Received 6 August 2009

Received in revised form 24 November 2009

Keywords:

Language processing
 Psycholinguistics
 Natural statistics
 Psychometric threshold
 Bayesian model

ABSTRACT

Humans are remarkably efficient at processing natural text. We quantified efficiency for discriminating a sample of meaningful text from a sample of random text by disrupting the meaningful sample, and measuring how much disruption human readers can tolerate before the two samples become indistinguishable. We performed these measurements for a wide range of conditions, involving samples of different lengths and containing letters, words or Chinese characters. We then compared human performance to the best possible performance achieved by a Bayesian estimator under the conditions in which we tested our participants, and in so doing we determined their absolute efficiency. Values were mostly in the range 5–40%, in agreement with reported efficiencies for many visual tasks. Although not intended as a veridical model of human processing, we found that the Bayesian model captured some (but not all) aspects of how humans classified text in our tasks and conditions.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

Natural text is highly regular in its statistical structure (Chater & Manning, 2006; Gigerenzer & Todd, 1999). Words in a sentence are not independently assembled, but rather follow certain rules regardless of the message conveyed by the sentence. These rules specify constraints on the statistics of text, as demonstrated by Shannon's simple experiment of picking two consecutive words at random from a book, then looking for the same two words at some other random point in the book, then choosing the word that immediately follows, and repeating this procedure recursively (Shannon, 1948). An example of a sentence that can be obtained in this way is "the head and in frontal attack on an English writer that the character of this point is therefore another method", which is remarkably close to natural text. The same logic applies to strings of letters (Damashek, 1995), as letters can only be strung up in certain ways. String regularity is exploited by Dasher, a software application that facilitates typing by prioritizing the availability of upcoming characters depending on their probability of occurrence at the present position in the string (Ward & MacKay, 2002).

Clearly humans can learn these statistical regularities as they acquire language. Our goal was to quantify this ability in units of absolute efficiency. By efficiency we refer to a specific concept from signal detection theory (Burgess, Wagner, Jennings, & Barlow, 1981), the ratio between human and ideal d' (an unbiased measure of sensitivity, see Tanner and Birdsall (1958)). Efficiency can also

be rewritten in terms of threshold signal-to-noise ratios (SNR) which is the formulation we adopt here (Green & Swets, 1966). In order for these quantities to be measurable it is necessary that stimuli, tasks and behavioral outputs are well-defined and that an ideal strategy can be formulated. Previous research has demonstrated the human ability to exploit the statistical regularities of text (Gigerenzer & Todd, 1999; Jurafsky, 2003; Lee & Corlett, 2003), but the chosen behavioral metrics could not be explicitly translated into SNR's. For example, the pattern of reading eye-movements is affected by the statistical predictability of text (Ehrlich & Rainer, 1981) and so are Cloze completion studies (Taylor, 1953), but it is not obvious how the resulting behavioral measurements would translate into SNR. This makes it difficult to incorporate them within the context of signal detection theory for the purpose of computing absolute efficiency and comparing this metric across different perceptual domains (e.g. with vision, Barlow (1980)), which was the main goal of our study.

For these reasons we chose a behavioral task of extreme simplicity: we asked human participants to select, between two strings, the one conforming more closely to their natural language. This task proved to be surprisingly robust: it was easy to learn and generated a large dataset of stable and reliable threshold measurements. By adopting a simple task and well-defined SNR characteristics we could formulate an ideal strategy for performing the task. Because all aspects of the stimulus and behavioral output are specified, the ideal strategy is based on likelihood ratios and is essentially a problem of Bayesian estimation (Geisler, 2003). We implemented the estimator via Monte Carlo simulations, and obtained threshold predictions corresponding to all measurements from human participants. A direct comparison between the two

* Corresponding author.

E-mail addresses: pn@white.stanford.edu, peter.neri@abdn.ac.uk (P. Neri).

sets of threshold SNR's yielded efficiency estimates for a variety of conditions, allowing us to provide a detailed and quantitative characterization of the human data.

2. Methods

2.1. Construction of databases

We tested three main conditions: 'word string', 'letter string' and Chinese 'character string'. For the 'word string' condition we used large amounts of digitized English text, which we assembled into one long sequence of text. We then eliminated all characters apart from the 26 letters of the alphabet and set all of them to lower case (Fig. 1A). We retained blank spaces for segmenting the text into words. The entire dataset consisted of ~200 K words (~15 K distinct words). When used in relation to this database the term 'element' refers to one word. For the 'letter string' condition we also removed all blank spaces to obtain one long sequence of uninterrupted text which we truncated at ~200 K letters. When used in relation to this database the term 'element' refers to one letter. For the Chinese 'character string' condition we created the database in a very similar way using Chinese characters. We identified ~1.2 K distinct characters within an overall database of ~5.5 K characters. When used in relation to this database the term 'element' refers to 1 character.

2.2. Stimuli and tasks

Stimuli consisted of text strings (Arial font) displayed on a CRT monitor. Our goal was to ensure that the only bottleneck for performance was language-related. We therefore adjusted the visual and temporal parameters of our stimuli for each subject to ensure comfortable and relaxed reading. Viewing distance was typically between 1 and 2 m, each character subtending $\sim\frac{1}{2}$ – 1° . Each string was typically presented for a total time corresponding to a reading speed of ~ 3 word/s for 'word' and 'character' conditions, and 10–12 characters/s for the 'letter' condition (i.e. stimulus duration scaled with string length).

Each trial consisted of two intervals, one containing a 'target' stimulus, the other one containing a 'non-target' stimulus, separated by a 1-s gap. Target and non-target were generated from the database as described in Fig. 1. We outline the procedure for the 'word string' condition, but the logic was identical for the other two conditions. The target string consisted of a segment from the database, preserving the original sequence of the elements and thus providing a small sample of the statistics that governed the structure of the database (Fig. 1B, left-most example). The non-target string consisted of a sequence of elements that were randomly selected from the database, indicated by black numbers in Fig. 1B. These numbers refer to the original order of the elements in the database and are shown here for clarity, but no such numbers were displayed to the participants during stimulus presentation.

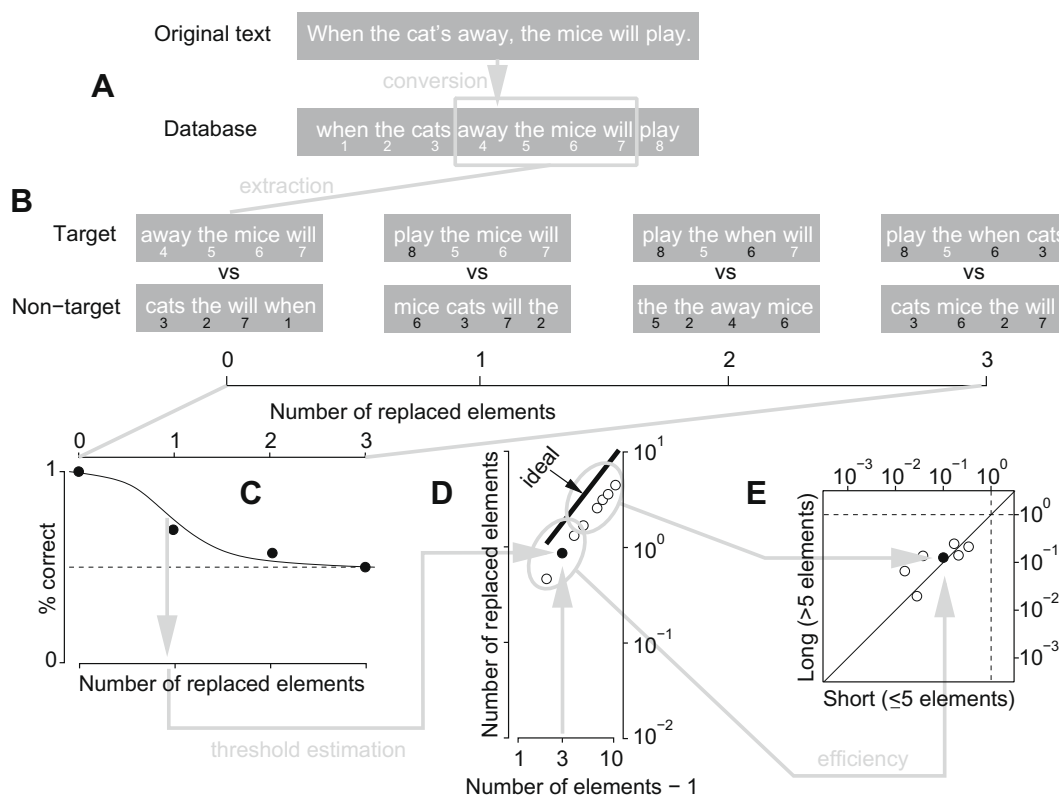


Fig. 1. Stimuli, task and metrics used in the experiments. (A) We generated a large database of text converted from written samples such as books and newspaper articles. The conversion only preserved the 26 letters of the alphabet and spaces between words. (B) Participants saw two strings on each trial and were asked to select the 'target' string (two alternative forced choice). The 'target' string was extracted from the database, while the 'non-target' string was generated by randomly selecting elements from the database one at a time. Participants were typically able to identify the target correctly on all trials (100% correct) for this condition (left-most point in C). We then replaced one or more elements in the 'target' string with randomly selected elements from the database. Randomly selected elements are indicated by black numbers. As the number of replaced elements increases (going from left to right in B and C), the percentage of correct responses decreases until it reaches chance (50% correct), shown by the psychometric curve in C. The number of replaced elements corresponding to 75% was taken as the basic threshold measurement (see Section 2). We repeated this measurement for different string lengths (the example in the figure is for length = 4), and plotted threshold number of replacements versus string length minus 1. (D). For all the conditions tested, points fell on a straight line in log–log units. We computed corresponding thresholds for an ideal estimator (solid line in D, see Section 2) and measured efficiency by taking the ratio between human and ideal thresholds. We computed average efficiency for thresholds corresponding to strings containing >5 elements ('long'), and plotted it against efficiency for 'short' strings (≤ 5) for each subject (shown in E, where different points refer to different participants). Fig. 2 is based on panel D and Fig. 3 is based on panel E.

Our participants were instructed to select the interval that most closely resembled their native language by pressing one of two keys. We tested 13 participants aged between 20–35 years old with normal vision and no known history of reading impairments. Participants S6, S8–S13 are native speakers of Chinese and participated in the ‘character string’ condition. One participant (S6) is proficient in English and was tested for the other two conditions as well. S1–S5 are native speakers of English, and were tested in the ‘word string’ and ‘letter string’ conditions. S7 is a native speaker of Italian, and was tested in the Italian equivalent of these two conditions, i.e. we created an Italian database for this subject in the same way that the English database was created (total of ~39 K words and ~191 K letters). Two authors participated in the experiments, S6 (AL) and S7 (PN). The remaining participants were naïve and knew nothing about the nature of the replacement process or the goal of the experiments.

We refer to the task described above as the ‘recognition’ task. We ran a set of additional experiments where the ‘target’ was generated by repeating one element that was randomly selected (with equal probability for all elements) from the database (in Fig. 1 potential targets would be ‘away away away away’ or ‘mice mice mice mice’) and participants were explicitly instructed to select this target. We refer to this task as the ‘repetition’ task. The purpose of this additional task was to estimate the extent to which the visual appearance of our stimuli may have impacted the results, independently of their language-related component. For example, the visual appearance of Chinese sentences (‘character’ condition) is markedly different from that of English sentences (‘word’ condition). When comparing results from the two conditions with relation to language processing, we must be sure that the observed differences may not be simply attributable to differences in the visual appearance of the two classes of stimuli.

2.3. Disruption by replacement and threshold measurements

Target and non-target were easily discriminated for trials like the one depicted in the left-most example of Fig. 1B. We refer to this condition as containing 0 replacements, as indicated by the

numeric label at the bottom of the two strings in Fig. 1B. Fig. 1C shows that this condition corresponds to 100% correct choices of the target interval (left-most point on the smooth curve). We then randomly replaced some of the elements in the target string according to the same rule that was used to generate the non-target string. This is shown for 1–3 replacements in the examples that follow the 0 replacement condition in Fig. 1B. As the number of replacements is increased, the target string is increasingly disrupted until it is statistically indistinguishable from a non-target string (right-most example in Fig. 1B). When the number of replacements equals the number of elements in the string minus 1, performance must drop to chance level (dashed horizontal line in Fig. 1C) because target and non-target are statistically indistinguishable.

For intermediate numbers of replacements, performance decreases smoothly from 100% to 50% correct following a trend well fit by a cumulative Gaussian (least-squares). The mean parameter associated with the best fit (corresponding to 75% correct performance) is the threshold number of replacements. For the example in Fig. 1B,C this is close to 1 (indicated by the arrow in C), meaning that a target string of four words can be discriminated from a random non-target string of four words (Fig. 1B) provided no more than ~1 word is randomly replaced in the target string. The number of replacements on a given trial was controlled by a 2-up 1-down staircase procedure and was constrained to be smaller than string length minus 1. We repeated this threshold measurement for different string lengths mixed within the same block (multiple parallel staircases). After running the staircase, on some occasions we increased the statistical reliability of the psychometric curve by adding trials at specific numbers of replacements (method of constant stimuli). Each threshold (one data point in Fig. 2) was estimated by combining all trials for that condition from all blocks in which they appeared (we collected 113 ± 55 trials (average \pm standard deviation across participants and conditions) per threshold for ‘recognition’ experiments, and 22 ± 4 for ‘repetition’ experiments). These trials were then used to compute a single psychometric curve, and the cumulative Gaussian was fitted to this curve (Probit analysis) to yield the threshold estimate.

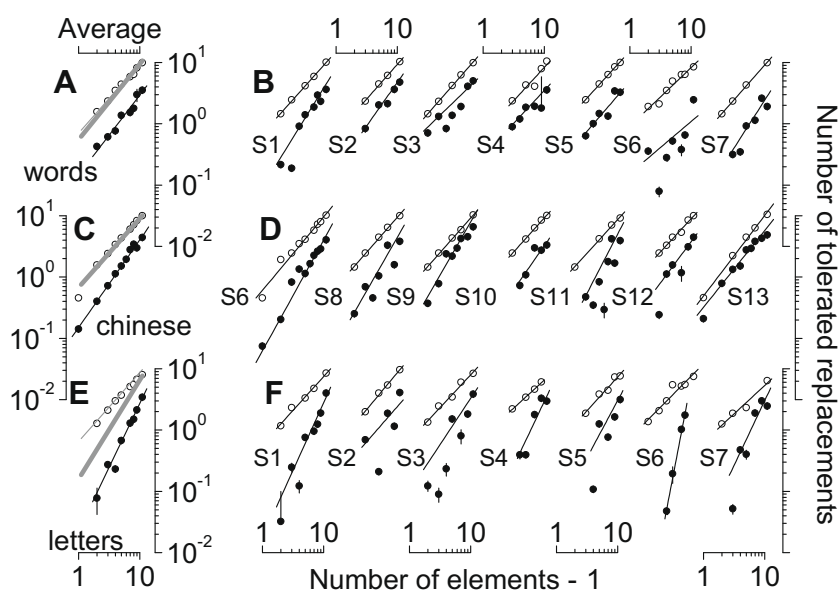


Fig. 2. Replacement thresholds for the ‘word string’ (top row), Chinese ‘character string’ (middle row) and ‘letter string’ (bottom row) conditions. Panels A, C and E show data averaged across participants (corresponding data for individual participants S1–13 is shown in B, D and F). Thresholds are plotted on the y axis against number of string elements – 1 on the x axis (refer to Fig. 1D). Solid symbols for recognition task, open symbols for repetition task. Grey lines show ideal predictions for recognition (thick) and repetition (thin), not separately visible in A and C because overlapping. Error bars show ± 2 s.e.m. (smaller than symbol when not visible).

2.4. Bayesian estimator

The ideal detector implements Bayes' rule by computing the likelihood ratio associated with a given string for its being a target as opposed to a non-target:

$$l(s_i) = p(s_i|T)/p(s_i|NT)$$

where s_i is the string presented on interval i , T means 'target' and NT means 'non-target'. The denominator is easily computed. If $s_i = [e_1, e_2, \dots, e_n]$, where n is the total length of the string s in units of number of ordered elements e , then $p(s_i|NT) = \prod_{j=1}^n p(e_j)$, where $p(e_j)$ is the frequency of element e_j in the database (which is easily computed from the database itself). The numerator can be expressed as follows:

$$p(s_i|T) = \sum_{j=0}^{n-2} p(s_i|T, r=j)p(r=j) = \frac{1}{n-1} \sum_{j=0}^{n-2} p(s_i|T, r=j)$$

where r is the number of replacements, i.e. the probability that s_i is a target is computed for each possible and equally probable (hence the $1/(n-1)$ factor) number r of replacements (we are assuming here that the ideal observer does not take into account the length dependency introduced by the staircase because a significant portion of the dataset was collected using the method of constant stimuli). The final probability of s_i being a target equals the sum of these probabilities (only computed for r up to $n-2$ because $r < n-1$ in the experiments: we never actually presented $r = n-1$ in the experiments because we know by design that performance must be at chance for this condition as target and non-target are statistically identical). For $r = 0$, the above probability is simply computed by identifying the frequency of occurrence of s_i across the entire dataset. This operation was performed by straightforward template matching for all positions within the dataset. For $r > 0$, the probability for each value of r is further broken down into:

$$p(s_i|T, r) = \frac{1}{n C_r} \sum_{n C_r} p(s_i|T, \{\rho_1, \rho_2, \dots, \rho_r\})$$

where the summation is taken over all (equiprobable) combinations of n elements taken r at a time, denoted by $n C_r$. The positions within s_i of the elements that are selected as putative replacements are listed within the $\{\}$ brackets and denoted by ρ_r . For example, if $n = 8$ and $r = 2$ (i.e. two replacements on a string composed of eight elements), there are 28 different ways in which the replacement can happen. One of these involves replacing elements at positions 2 and 5, which we express as $\{\rho_1 = 2, \rho_2 = 5\}$. Clearly we must assume here that not only does the ideal detector not know how many replacements were applied to a given string, but it also does not know which elements were replaced when it computes the probability for the possibility that there were r replacements. Consequently, the probability must be computed for all possible positions of the replacements within the string. The probability associated with each specific choice $\rho_1, \rho_2, \dots, \rho_r$ of replacement positions (argument of the summation above) can be written as:

$$p(s_i|T, \{\rho_1, \rho_2, \dots, \rho_r\}) = p(s_i^*|T, r=0) \times \prod_{j=1}^r p(e_{\rho_j})$$

where $s_i^* = s_i - \{\rho_1, \rho_2, \dots, \rho_r\}$ indicates a string where the elements at the positions that are being subtracted (the replaced ones) become irrelevant for the purpose of matching the string to the database (as if $r = 0$). Going back to the example above, $s_i - \{\rho_1 = 2, \rho_2 = 5\} = [e_1, *, e_3, e_4, *, e_6, e_7, e_8]$. The probability $p([e_1, *, e_3, e_4, *, e_6, e_7, e_8]|T)$ is then computed by simple template matching with the database, where a positive match is obtained regardless of the elements encountered across the database at the positions indicated by the $*$.

We challenged this ideal estimator with the same stimuli used for the psychophysical experiments. On each trial the estimator selected the interval associated with largest l as being the 'target' (if $l(s_1) = l(s_2)$ the estimator generated a random choice). We obtained full psychometric curves like that in Fig. 1C (using the method of constant stimuli – no staircase was used for the model), and determined the ideal psychometric threshold using the same fitting procedure used for human participants. Similar to the psychophysical data, the function describing how threshold varies with n conformed very closely to a line in log-log coordinates (grey lines in Fig. 2). For the 'letter string' condition we also studied the effect of decreasing the % of database available to the Bayesian estimator for computing likelihood ratios. We estimated thresholds for $n = 4$ and $n = 8$ when the estimator only had access to 1%, 2% and 10% of the database (as opposed to 100% for ideal estimation); we chose these values because pilot simulations showed that they provided adequate sampling of the underlying trend. Thresholds decrease very similarly for the two values of n , keeping their ratio unchanged at ~ 3.6 . We also performed a similar (though not identical) set of simulations to determine whether the size of the dataset for the 'word string' condition was sufficient to generate asymptotic threshold behavior of the Bayesian estimator. We estimated ideal thresholds for $n = 4$ and $n = 6$ when the word database was limited to 1%, 2% and 10% of its original size. There was no appreciable change in thresholds for any of these different dataset sizes, demonstrating that 1% of the database was already sufficient to allow asymptotic estimation of ideal thresholds for the 'word string' condition. More generally, we found that even gross assumptions about database structure (for example setting $p(e_j)$ to be the same for all elements) had little or no impact on ideal thresholds for all conditions within the range of resolution that is relevant to the experimental measurements presented here.

2.5. Efficiency measurement

Efficiency is defined as the squared ratio of experimental to ideal d' . Because d' is related to the signal-to-noise ratio (SNR) in the stimulus, efficiency depends on the ratio between the ideal and the experimental SNR at threshold (Green & Swets, 1966). A natural definition of threshold $1/\text{SNR}$ for a target string in our experiments is the threshold % of disrupted string (the ratio between the threshold number of replacements and the length of the string), so that efficiency is transparently expressed as the ratio between human and ideal thresholds. In visual psychophysics thresholds are often squared for the purpose of computing efficiency (e.g. Pelli, Farell, & Moore, 2003) because signal and noise are defined in terms of energy (or power). In the context of the experiments described here it makes little sense to square letter thresholds, so we simply took their ratio: efficiency = t_h/t_i where t_h is the human threshold and t_i the ideal threshold.

3. Results

3.1. Thresholds scale with string length

Our database was directly derived from natural text (Fig. 1A; see Section 2). Participants saw two intervals on each trial, a 'target' interval and a 'non-target' interval. Both were obtained by picking words from the dataset, but this was done sequentially in the 'target' interval (white digits in Fig. 1B) and randomly in the 'non-target' interval (black digits), so that only the 'target' interval conformed to the statistics of natural text. Participants were asked to identify the interval containing text that most closely resembled their natural language (see Section 2 for details). This task turned out to be unexpectedly easy to learn and perform,

generating stable and reproducible measurements. The task is increasingly more difficult when some of the words in the 'target' interval are replaced by words randomly selected from the database. We measured the threshold number of replacements that corresponded to a fixed performance level (Fig. 1C). A similar procedure was used for the letter and Chinese character experiments, except we replaced individual characters in these conditions (see Section 2).

All plots in Fig. 2 show the total number of string elements -1 on the abscissa versus number of tolerated replacements on the ordinate (see Fig. 1D for how these plots relate to individual threshold measurements). We plot number of elements -1 (rather than just number of elements) because signal = 0 occurs for string length = 1 (when target and non-target are statistically identical), and we wish to plot signal intensity on the x axis. Not surprisingly all plots show that participants can tolerate more and more replacements as the string gets longer (see Neri, Morrone, and Burr (1998) and Wong and Barlow (2000) for related effects in vision and audition). When plotted on log-log axes, this trend is very well described by a line. Fig. 2A shows data for the word experiment (solid symbols), averaged across participants. The error-weighted linear fit (indicated by black line in Fig. 2A) has a slope of 1.24 and a correlation coefficient of 0.98. The extraordinarily good quality of these experimental data is not simply a consequence of subject averaging, because high correlation coefficients were also found for linear fits to individual data (black lines in Fig. 2B; average correlation coefficient across participants 0.95 ± 0.04 sd). We conclude that a linear fit on log-log axes provides an adequate description of the data. We measured similar trends for the Chinese and letter experiments (Fig. 2C–F, correlation coefficients of 0.99 and 0.98 respectively), although the slope of the linear fit to the letter data was markedly steeper (2.16, see Fig. 2E).

3.2. Control experiment to exclude a role for visual bottlenecks

To ensure that our experiments were targeting language processing, and not visual processing, we devised a version of our tasks and stimuli that would correspond to a visual detection task. In this version of the experiment the 'target' string consists of a randomly selected word repeated n times (see Section 2). Data for this experiment is shown by the open symbols in Fig. 2 which invariably fall above the solid symbols. This demonstrates that when participants were at threshold for identifying meaningful strings they were always above thresholds for simply reading the words and being able to count how many times they repeated within the string, implying that the bottleneck for identifying meaningful strings was never visual (where the term 'visual' in-

cludes eye-movements). We also observed similar trends for identifying words and Chinese characters despite the different visual appearance of these two forms of writing. Finally, it is interesting that the slope of the linear fits to the detection data on log-log axes is always ~ 1 (word 1.07, character 1.24, letter 1.05), implying a linear relationship between string length and number of tolerated replacements. This relationship conforms to expectations from signal detection theory (Green & Swets, 1966; Neri et al., 1998) and is fully accounted for by our simulations (see below).

3.3. Efficiency for short versus long strings

We wrote a computer program that implemented a so-called 'ideal observer', a theoretical device which provides the best possible performance attainable in our tasks and databases (see Section 2) indicated by the grey lines in Fig. 2. As expected, these lines fall above the corresponding human data (thick line for recognition, thin line for repetition). We can combine human and ideal thresholds into one quantity called 'efficiency', the ratio between human performance and ideal performance (Fig. 1E). Fig. 3 plots efficiency values obtained from the data in Fig. 2, averaged for long strings (>5 elements) on the y axis versus short strings (≤ 5 elements) on the x axis. Each point refers to an individual subject. Points tend to lie above the unity line (higher efficiency for long strings as opposed to short strings). This effect is significant for Chinese characters ($p = 0.001$ paired t -test across participants for long $>$ short) and for letters ($p = 0.02$), but not for words ($p = 0.2$). No such effect was observed for the repetition task (open symbols), which showed an efficiency very close to 1 in all conditions (see magnifying plot in Fig. 3B).

4. Discussion

It is well known that humans can overcome disruption/impoverishment of meaningful text. For example, the following text can be read with quite amazing ease (Grainger & Whitney, 2004). In this study we quantified how much disruption can be tolerated by humans before a string of meaningful text becomes indistinguishable from random text. At first sight it may appear that the question we asked our participants was poorly defined. In our protocol, target strings are defined as belonging to the subject's natural language, with no description of specific string elements. Nevertheless, our sample of human participants learned this task very easily and generated well-behaved psychometric curves after 20–30 trials. The robustness of our measurements is demonstrated by the extremely high correlation coefficients (close to 1) for the

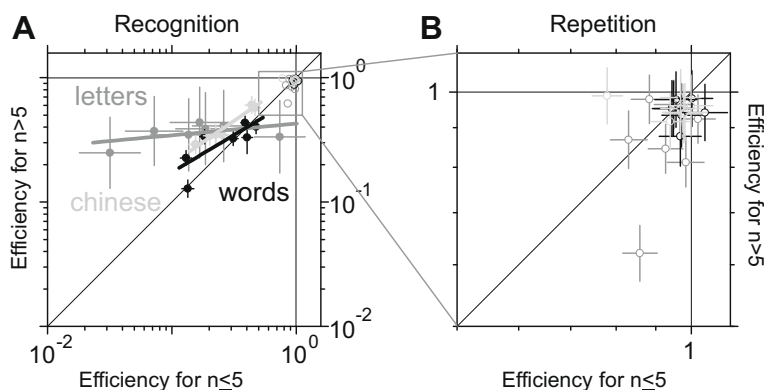


Fig. 3. Efficiency for strings containing >5 (y axis) versus ≤ 5 elements (x axis), where each element could be an English word (black), a Chinese character (light grey) or a Western letter (dark grey). Panel B magnifies a portion of A as indicated by the joining lines. Solid symbols for recognition task, open symbols for repetition task. Different data points refer to different participants (refer to Fig. 1E). Error bars show ± 1 s.e.m.

log–log fits to experimental data, indicating that participants were performing this task using a stable and consistent strategy (see Fig. 2).

In this study we were interested in the comparison between human performance and a specific model: the ideal observer. This model is uniquely defined, and it specifies the level of performance that is expected from a device with perfect knowledge of the statistics of the stimuli and the methods by which they were manipulated. This level of performance represents an upper bound for the performance of any model. Because solving our task optimally involved a simple process of Bayesian estimation, our approach shares superficial similarities with other models of text classification (e.g. Hale, 2001; Lee & Corlett, 2003; Levy, 2008; Nigam, McCallum, Thrun, & Mitchell, 2000; Norris, 2006; Yang & Liu, 1999). However it should be emphasized that the goal of these previous studies was to model human performance, whereas our goal was to measure human efficiency. From the modeling point of view, the latter goal is simpler because only the ideal model needs to be considered. From the experimental point of view, it imposes specific constraints on stimulus design, SNR manipulation and sensitivity measurements. In our study these constraints are reflected in the choice of a simple behavioral measure (2AFC binary response) that may have overlooked other important indicators of text-processing such as confidence (Vickers & Lee, 1998) or compensation (Lee & Corlett, 2003). However, our approach has the benefit that it allows a direct comparison of text-processing measurements with studies from other cognitive domains, notably vision. Our estimates of efficiency for processing natural text fall between 3% and 50%, a range that overlaps with similar measurements for visual tasks such as signal detection in noise (Burgess & Colborne, 1988), density discrimination (Barlow, 1978), motion processing (Barlow & Tripathy, 1997), symmetry perception (Barlow, 1980) and object recognition (Tjan, Braje, Legge, & Kersten, 1995). We found two main effects for efficiency, which we discuss in detail below.

First, in some conditions efficiency increases as the text sample becomes longer. This would not be expected if the participants processed text using quasi-ideal estimation subsequently corrupted only by late decisional noise: for this class of models (commonly used in vision) efficiency is expected to drop below 1 (due to noise) by the same amount for all sample lengths. We divided our data into short (≤ 5 elements) and long (> 5), but this arbitrary distinction (based on the average length of an English word being five letters) was only adopted to bring out the effect with better statistical reliability. The underlying trend is that efficiency increases gradually with increasing sample length. We do not have a ready explanation for this phenomenon, and for why it only applies to Western/Chinese characters and not to words. This result may appear to contradict Pelli et al. (2003)'s finding that efficiency drops with increasing number of letters, but the two sets of findings are not directly comparable. Pelli and collaborators studied how letters are seen; we studied how letters and words, once seen, are assembled for the purpose of classifying them as linguistic constructs. The two sets of efficiency measurements are therefore entirely distinct, even though (as mentioned above) they span similar overall efficiency ranges (but differ in the way efficiency varies as a function of string length).

Second, efficiency spans a very similar range (20–60%) across participants for all three tasks when samples are long, but efficiency for short samples of letters spans a much wider range (3–70%) than words or Chinese characters. Moreover, efficiency for short samples of letters is weakly correlated with efficiency for long samples (0.50, $p = 0.25$), whereas the correlation is strong for the other two conditions (0.81, $p < 0.03$ for words and 0.98, $p < 0.0002$ for Chinese characters). In other words, if a specific individual is better than average at processing long

samples of words or Chinese text, it can be predicted that he/she will also be better than average at processing short samples of this written material. In contrast, the ability to process long samples of letters does not allow reliable prediction of the ability to process shorter samples. We attempted to model this effect by assuming that different participants had different degrees of access to the statistics of the database, possibly as a consequence of different exposure to text during their lifetime. As expected, a reduction in the access to database statistics does result in decreased performance of the Bayesian model, but equally for long and short strings (see Section 2) failing to provide an account for the large variation that we observed only for short strings of letters. Regardless of possible interpretations, this empirical result has direct implications for future studies using short strings of letters (e.g. the word superiority effect (Cattell, 1886; Reicher, 1969)). It is also interesting that, in contrast to human thresholds, ideal thresholds for 'word' and 'character' conditions did not differ between recognition and repetition tasks (Fig. 2A and C, thin and thick grey lines overlap). This demonstrates that tasks and signals that may appear substantially different from a perceptual standpoint (like recognition and repetition tasks) may correspond to identical thresholds from a statistical standpoint.

In conclusion, the Bayesian estimator provides a satisfactory account of some aspects of human text-processing within the context of our stimuli, tasks and measurements, but fails in other important respects. As expected, the performance of the Bayesian estimator is superior to the human in all conditions (with the exception of the repetition task for which efficiency ~ 1), but this difference is easily accounted for by reducing the overall reliability of the Bayesian estimator (e.g. by adding response noise or by reducing access to the database). The critical result is that the same Bayesian estimator generates similar efficiency estimates for short and long sequences of words (black symbols in Fig. 3A), thus capturing two independent sets of data and providing a unified account of both conditions. However, as the unit of manipulation varied from word to character (via Chinese characters which represent an intermediate step) a difference in efficiency between short and long strings became more evident, indicating that other phenomena not captured by the Bayesian estimator were at play. Despite these departures from the data, it should be noted that the model nicely captured the slopes of the noise-versus-signal functions in Fig. 2. This non-trivial result indicates that the Bayesian estimator offers a promising framework for understanding and characterizing the human ability to identify the statistics of natural text, although clearly it only represents a first coarse step (Chater & Manning, 2006).

Acknowledgment

Supported by NIH (R01EY01728 to DML), Royal Society (University Research Fellowship to PN) and Medical Research Council (New Investigator Research Grant to PN).

References

- Barlow, H. B. (1978). The efficiency of detecting changes of density in random dot patterns. *Vision Research*, 18, 637–650.
- Barlow, H. B. (1980). The absolute efficiency of perceptual decisions. *Philosophical Transactions of the Royal Society of London Series B*, 290, 71–82.
- Barlow, H. B., & Tripathy, S. P. (1997). Correspondence noise and signal pooling in the detection of coherent visual motion. *Journal of Neuroscience*, 17, 7954–7966.
- Burgess, A. E., & Colborne, B. (1988). Visual signal detection. IV. Observer inconsistency. *Journal of the Optical Society of America A*, 5, 617–627.
- Burgess, A. E., Wagner, R. F., Jennings, R. J., & Barlow, H. B. (1981). Efficiency of human visual signal discrimination. *Science*, 214, 93–94.
- Cattell, J. M. (1886). The time taken up by cerebral operations. *Mind*, 11, 220–242.

- Chater, N., & Manning, C. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10, 287–291.
- Damashke, M. (1995). Gauging similarity with n -grams: Language-independent categorization of text. *Science*, 267, 843–848.
- Ehrlich, S. F., & Rainer, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20, 641–655.
- Geisler, W. (2003). Ideal observer analysis. In L. Chalupa & J. Werner (Eds.), *The visual neurosciences* (pp. 825–837). Cambridge, MA: MIT Press.
- Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.
- Grainger, J., & Whitney, C. (2004). Does the human mind read words as a whole? *Trends in Cognitive Sciences*, 8, 58–59.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of NAACL*, 2, 159–166.
- Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics*. Cambridge, MA: MIT Press.
- Lee, M. D., & Corlett, E. Y. (2003). Sequential sampling models of human text classification. *Cognitive Science*, 27, 159–193.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106, 1126–1177.
- Neri, P., Morrone, M. C., & Burr, D. C. (1998). Seeing biological motion. *Nature*, 395, 894–896.
- Nigam, K., McCallum, A., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39, 103–1234.
- Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, 113, 327–357.
- Pelli, D. G., Farell, B., & Moore, D. C. (2003). The remarkable inefficiency of word recognition. *Nature*, 423, 752–756.
- Reicher, G. M. (1969). Perceptual recognition as a function of the meaningfulness of stimulus material. *Journal of Experimental Psychology*, 81, 275–280.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423.
- Tanner, W. P., & Birdsall, T. G. (1958). Definitions of d' and η as psychophysical measures. *JASA*, 30, 922–928.
- Taylor, W. L. (1953). A new tool for measuring readability. *Journalism Quarterly*, 30, 415.
- Tjan, B. S., Braje, W. L., Legge, G. E., & Kersten, D. (1995). Human efficiency for recognizing 3-D objects in luminance noise. *Vision Research*, 35, 3053–3069.
- Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgments. I. Properties of a self-regulating accumulator module. *Nonlinear Dynamics, Psychology, and Life Sciences*, 2, 169–194.
- Ward, D. J., & MacKay, D. J. (2002). Artificial intelligence: Fast hands-free writing by gaze direction. *Nature*, 418, 838.
- Wong, W., & Barlow, H. B. (2000). Tunes and templates. *Nature*, 404, 952–953.
- Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In: *SIGIR '99 Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*, pp. 42–49.