The eye tracks the aesthetic appeal of sentences

Hideyuki Hoshi
Department of Language and Literature,
Max Planck Institute for Empirical Aesthetics,
Frankfurt am Main, Germany

Winfried Menninghaus
Department of Language and Literature,
Max Planck Institute for Empirical Aesthetics,
Frankfurt am Main, Germany

Eye-tracking parameters (fixation and pupillary responses) have been shown to be modulated by the aesthetic perception and evaluation of visual and auditory artworks (e.g., paintings, music). The present study investigated whether similar effects can be found in visual text processing. Participants read four groups of short sentences in which a key predictor of aesthetic liking, i.e., familiarity, was systematically modified to four degrees. Across all four groups, the sentences moreover varied with regard to featuring or not featuring meter. During reading, pupil sizes and eye movements were recorded. Aesthetic ratings of all sentences were collected afterwards, and the relationships between the ratings, levels of familiarity, meter, and eye-tracking datasets were tested. The results showed that the rating scores were interactively modulated by both familiarity-driven and meter-driven fluency. Using factor analysis, we extracted two key factors of the aesthetic appeal of the texts: an affective and a cognitive factor. The cognitive factor comprised the rating items "succinctness" and "familiarity," whereas the affective factor reflected the ratings for "beauty" and "liking." A higher cognitive factor predicted shorter dwelling time. Moreover, the two factors modulated the pupillary data antagonistically: A higher affective factor predicted larger pupil dilations, whereas a higher cognitive factor predicted smaller pupil dilations. The study shows a possible application of the eye-tracking method for capturing aesthetically evaluative dimensions of processing sentences.

Introduction

Pupillary measures in empirical aesthetics

Empirical studies in aesthetics often use eye-tracking as a means of capturing indicators of subjective experience. In particular, pupillary responses have been shown to reflect emotional arousal (Granholm & Steinhauer, 2004) and art-elicited chills (Laeng, Eidet, Sulutvedt, & Panksepp, 2016). Given that pupillary dilations are spontaneous reactions and not under the control of the respondent (Laeng, Sirois, & Gredebäck, 2012; Laeng & Sulutvedt, 2014; Loewenfeld, 1999), they can be considered as an objective index of aesthetic perception and evaluation. Previous studies have found pupillary dilations to be modulated by the aesthetic appeal of sexual stimuli (Dabbs, 1997; Hess & Polt, 1960; Rieger & Savin-Williams, 2012), visual artworks (Blackburn & Schirillo, 2012; Johnson, Muday, & Schirillo, 2010; Kuchinke, Trapp, Jacobs, & Leder, 2009; Powell & Schirillo, 2011), and music (Laeng et al., 2016).

The present study examined whether or not these findings can be extended to linguistic stimuli, specifically to the processing of single sentences that feature characteristics often found in proverbs, slogans and commercial ads, and also in poetry (cf. Menninghaus et al., 2015; Menninghaus et al., 2017).

Key theoretical hypothesis: Familiarity and Parallelistic patterning enhance cognitive fluency which in turn modifies aesthetic liking

Since Zajonc (1968) reported the mere-exposure effect (e.g., more frequently presented stimuli are preferred over less frequently presented ones), a positive relationship between familiarity-driven ease of cognitive processing and aesthetic liking/preference, has been repeatedly confirmed (Bornstein, 1989; Martindale & Moore, 1988). The respective findings have been summarized as the cognitive fluency hypothesis (Reber, Schwarz, & Winkielman, 2004; Reber, Winkielman, & Schwarz, 1998). Regarding linguistic stimuli, it has been shown that not just familiarity (i.e., repeated prior exposure), but also several features of

Citation: Hoshi, H., & Menninghaus, W. (2018). The eye tracks the aesthetic appeal of sentences. *Journal of Vision*, *18*(3):19, 1–22, https://doi.org/10.1167/18.3.19.

https://doi.org/10.1167/18.3.19 Received July 17, 2017; published March 29, 2018 ISSN 1534-7362

phonological and prosodic "parallelism" (such as rhyme, meter, alliteration, etc.; for the concept of "parallelism," see Fabb, 2015; Jakobson, 1960; Menninghaus et al., 2017) enhance the perceptual ease of processing and thereby also render a sentence more aesthetically appealing (Kuchinke et al., 2009; McGlone & Tofighbakhsh, 1999; McGlone & Tofighbakhsh, 2000; Menninghaus, Bohrn, Altmann, Lubrich, & Jacobs, 2014; Menninghaus et al., 2015; Reber et al., 2004).

Our study was designed to gain access to the eye-tracking signature of aesthetic liking for linguistic stimuli by tapping into the well-established potential of the two variables *familiarity* and *meter* to modify *aesthetic liking*. To this end, we used a stimulus set that features four systematic gradations of familiarity across four sentence categories, each comprising 40 items (Bohrn, Altmann, Lubrich, Menninghaus, & Jacobs, 2012). At the same time, as the corpus includes both metered and nonmetered sentences across all four sentence categories, it also allows for investigating the influence of regular meter on aesthetic evaluation (for all details, see Methods, Stimuli section).

We expected that high familiarity of sentences as well as metrical sentence patterning should enhance the overall processing fluency, and hence modulate the perceived aesthetic appeal of language. We also expected that correlating the eye-tracking data with the subjective ratings for aesthetic appeal—as modified by four degrees of familiarity and the difference meter versus nonmeter—should allow us to detect eye-movement patterns distinctive of aesthetic liking.

Subjective ratings capturing the aesthetic appeal of sentences

Regarding subjective ratings, we decided to collect familiarity, succinctness, beauty, and liking ratings. We included a *Beauty* rating, because beauty is *the* preeminent category used for evaluating aesthetic appeal (Augustin, Wagemans, & Carbon, 2012; Jacobsen, Buchta, Köhler, & Schröger, 2004; Knoop, Wagner, Jacobsen, & Menninghaus, 2016). We collected *Liking* ratings, because liking is widely used as the most general indicator of positive aesthetic evaluation, even in cases where beauty is not an issue (such as in regard to horror films). Moreover, Liking ratings have repeatedly been shown to reflect the rewards of ease of processing (cf. Silvia, 2007; Winkielman & Cacioppo, 2001; Winkielman, Schwarz, Fazendeiro, & Reber, 2003; see Reber et al., 2004 for a review).

Given our focus on modifying degrees of familiarity, we followed previous studies on proverbs and other single sentences (Bohrn et al., 2012; Bohrn, Altmann, Lubrich, Menninghaus, & Jacobs, 2013) in also collecting *Familiarity* ratings. Finally, and again following previous studies using proverbs (Bohrn et al., 2012; Bohrn et al., 2013; Menninghaus et al., 2015), we included a *Succinctness* rating. The German term that we here translate as "succinctness" is the word "Praegnanz." "The law of *praegnanz*" is the most basic principle proposed in the earlier 20th century by German *gestalt* psychology (Koffka, 1935; Wertheimer, 1923). It was so internationally successful at its time that it paved the way for the inclusion of the word "Praegnanz" into English dictionaries; the German word *gestalt* was likewise adopted into the English language in this very context. Still, in order not to alienate readers unfamiliar with this tradition, we prefer to use the term "succinctness" as a decent (though not perfect) English translation of the German word throughout this manuscript.

Two dimensions of the meaning of praegnanz/succinctness are particularly noteworthy here. First, a succinct sentence, or statement, should drive home its message in as *short* a form as possible, avoiding all detours and superfluous pieces of information. Second, it should leave a *strong imprint/impression* in the reader not only through its meaning, but also as a distinct verbal *gestalt*, i.e., by virtue of its very form, thereby gaining access to a privileged storage in memory. Proverbs clearly aim at these goals associated with succinctness. The English proverbial saying "East or West, home is best" is a good example of a sentence that meet these requirements (for a detailed analysis of this sentence as well as of a German analogue, see the Methods, Stimuli). The saying could barely be any shorter and more memorable. A standard variant of this sentence would be "Whether it's located in the East or West, one's own home is always the best place to be." While including expectable syntactic components missing in the original proverb (most notably, a verb phrase in its first part), this normalized variant lacks the rigid parallelistic structure of the original proverb and consequently appears to be far less succinct and memorable. In line with this understanding, Menninghaus et al. (2015) reported that in the assessment of short sentences, specifically proverbs, Succinctness ratings capture an important aesthetic appeal dimension and that German-speaking participants readily have an intuitive understanding of the task to rate a proverb for succinctness.

Essentially, the difference between the two sentence variants analyzed above is what separates lexicalized proverbs—as well as other nonlexicalized aphorisms, slogans and commercial ads— from ordinary sentences. Our study tapped into the cognitive, affective, and aesthetic implications of this very difference.

It is a key assumption of the cognitive fluency hypothesis that select dimensions of the *cognitive* processing of stimuli support the emergence of aes-

thetically evaluative responses that are of a strongly *affective* nature. For obvious reasons, this applies to Liking ratings. Beauty judgments, too, have been considered to be strongly affective, or emotional, at least since Kant systematically spoke of "feelings of beauty" (Kant 1790/2001; for a recent account of the emotional nature of beauty, see Armstrong & Detweiler-Bedell, 2008). By contrast, Familiarity ratings are primarily of a cognitive nature. Similarly, we expected that Succinctness ratings should primarily, though not exclusively, reflect cognitive processing dimensions.

Eye-tracking studies of linguistic stimuli

Previous studies (Hess & Polt, 1960; Kuchinke et al., 2009; Laeng et al., 2016) have reported that stronger affective responses evoke larger pupillary dilation. As the present study uses linguistic stimuli, effects caused by the cognitive processing of language must also be taken into account. The relations between oculomotor parameters and the cognitive processing of language have been studied for about a century (Huey, 1908; see Clifton et al., 2016 and Rayner, 1998, for comprehensive reviews). It is widely assumed that the number of fixations and their duration reflect the cognitive processing effort required by the word fixated upon—Just and Carpenter (1980) famously called this the "eye–mind assumption." Not only the foveal word but also the texts in parafoveal locations (e.g., upcoming words) are processed orthographically, phonologically, and semantically (Heister, Würzner & Kliegl, 2012; Schotter, Angele, & Rayner, 2012) during a given fixation period.

The relationship between pupillary dilations and language processing is also well-studied. Several studies (Just & Carpenter, 1993; Schluroff, 1982; Schluroff et al., 1986) have reported that syntactically complex and ambiguous sentences produce larger amplitudes in pupillary responses. Using semantic processing tasks with different task demands (e.g., translating, shadowing, and listening), Hyönä, Tommola, and Alaja (1995) found that more demanding tasks were associated with larger pupillary dilations. Furthermore, language processing is dependent on a wide variety of cognitive functions, such as sensory detection, long- and short-term memory, and attention. These processing dimensions also correlate with pupillary dilations (see Beatty & Lucero-Wagoner, 2000, for a review).

In sum, both cognitive demand and affective impact are likely to affect oculomotor and pupillary responses to linguistic stimuli. Based on the assumption that the aesthetic appeal of language also has cognitive and affective dimensions (Jacobs, 2015; Menninghaus et al., 2015), the present study investigated how this dual innervation specifically applies to aesthetic language processing.

Affective and cognitive processing modulate pupil size differently

To start with, potentially co-occurrent affective and cognitive innervations of the pupillary dilation make the pupillary response waveforms more complex. According to the cognitive fluency hypothesis, easy-to-read texts should in general be preferred by readers. If so, such texts should generate smaller pupillary dilation because they are cognitively less demanding. At the same time, they should generate larger dilation because they are also affectively preferred. Theoretically, these antagonistic effects could thus cancel each other out.

However, the temporal trajectories of the affective and the cognitive processing dimensions are likely not to be fully convergent; each may be predominant at different points in time, thereby precluding a cancellation of their hypothetically antagonistic effects. Specifically, the cognitive effect on pupil size is usually detected during the actual reading (Kuchinke, Võ, Hofmann, & Jacobs, 2007), whereas the latency of affective influence is relatively long (Johnson et al., 2010; Kuchinke et al., 2009). Therefore, even if the cognitive and affective processing were to have antagonistic effects on pupillary measurements, texts with different cognitive demands would show different pupillary dilation patterns over the course of time.

To capture the temporal dynamics, the present study used polynomial-curve fitting, which is often applied to time-course pupillary data (e.g., Kojima et al., 2004; Kristjansson, Stern, Brown & Rohrbaugh, 2009). Pupillary response curves were fitted into a two-dimensional polynomial curve, and the linear and quadratic components were estimated. Modelling the estimated coefficients as a function of aesthetic rating scores enabled us to analyze the relationship between the pupillary response dynamics and the aesthetic appeal of the text.

In sum, the present study investigated the eye-tracking signature of perceived aesthetic appeal in sentence processing, with a specific focus on the moderating roles of familiarity and meter.

Methods

Participants

Thirty-seven healthy adult volunteers took part in the study, and data from 29 participants (23 female, six male; one left-handed; mean age 24.7 ± 3.9 years) were

analyzed. Eight participants were excluded from the analysis because of the large number of missing data points (due to artefacts), excessive blinking, making multiple error trials in the checking task (see Task section for details), or application of a wrong setting in the data collection. All participants were native German speakers and had normal or corrected-to-normal vision. All experimental procedures (e.g., verbal instructions, screen messages) were conducted in German. None of the participants had a history of neurological or psychiatric disorders (e.g., dyslexia). All experimental procedures were ethically approved by the Ethics Council of the Max Planck Society and were undertaken with the written informed consent of each participant.

Stimuli

Overview of stimuli

For the present study, we reused a set of stimuli comprising four categories of sentences: original *Familiar Proverbs*, *Synonyms,* and *Creative Alterations* of these proverbs, and original *Unfamiliar Proverbs*. This stimulus set was first introduced by Bohrn et al. (2012; stimuli are listed in Supplementary Figure S1). Bohrn and colleagues performed several pretests on a stimulus pool that originally included 800 proverbs mostly dating back to the 19th century. They asked 14 participants to give dichotomous familiarity judgments ("known" or "unknown"), and ended up selecting 40 familiar proverbs and 40 unfamiliar proverbs (judged as "known" and "unknown," respectively, by all of the 14 participants).

Across categories, the stimuli were controlled for potential differences in lexical parameters (i.e., number of words, syllables, and digits), mean word frequency, and emotional valence. At the same time, semantic meaning was not kept identical or near-identical throughout all sentence variants. Rather, differences in wording and semantic meaning were systematically used to generate differences in perceived familiarity. The original *Familiar Proverbs* (for instance, the iambic proverb "Wer *wágt*, gewínnt," which has a direct English counterpart—"He who *dares*, wins"—that is, however, not metered) make up the highly familiar end of the spectrum while the obsolete and hence unfamiliar proverbs make up the highly unfamiliar end. The two other sentence categories feature complementary blends of familiarity and nonfamiliarity.

In the *Synonyms*, the content of the underlying familiar proverb is retained, while the wording is changed. This change often implies the disruption of meter (and/or of other rhetorical features for which we did not separately control for want of sufficiently regular occurrence). Getting back to our English example ("East or West, home is best"), a semantically roughly synonymous variant would be "East or West, home is the greatest." While this variant retains features of syntactic ellipsis, the meter is destroyed (and in this case also the rhyme). Regarding our German example ("Wer wagt, gewinnt"), the synonym "Wer riskiert, gewinnt" ("He who *takes risks*, wins") likewise ruins the meter and the prosodic parallelism between the two syntactic phrases, because one syllable is added in replacing "wagt" through "riskiert."

In the *Creative Alterations*, the formal template of the familiar proverb is retained and clearly recognizable, but the content is drastically changed, often in a somewhat witty way. Thus, the variant "Wer *klágt*, gewinnt" retains the meter (nonmetered translation: "He who *sues*, wins") and differs from the original only with regard to two phonemes, but the original meaning is deliberately and drastically altered. A Creative Alteration of "East or West, home is best" would, for instance, be "East or West, sex is best." Again, both the multiple parallelistic and the elliptic structures remain in place, and the underlying lexicalized template is readily recognizable, but the meaning is markedly changed and hence unfamiliar.

The fourth sentence category includes completely semantically unrelated proverbs that are no longer in use and hence were rated as entirely *Unfamiliar*. Still, regarding lexical parameters, mean word frequency, and emotional valence, the lexical material of each of these sentences did not differ statistically from the three other sentences with which it was grouped into a four-category item.

Thus, the four sentence categories feature four systematic gradations of familiarity. As noted earlier, we expected that this would allow us to take advantage of the psychological mechanism explained by the cognitive fluency hypothesis (see Introduction, Key theoretical hypothesis): Different levels of the familiarity of the stimuli should generate different levels of processing ease, which, in turn, should lead to differences in perceived aesthetic appeal and potentially correlative eye-movement patterns. Theoretically, sentences conveying a convergent meaning can differ as strongly in beauty and succinctness as sentences that are divergent in meaning. Therefore, if ratings for the aesthetic appeal of sentences are used as general predictors of eye-tracking measures (and vice versa), differences in semantic meaning, while clearly adding more variance to the stimuli, should not compromise the search for correlations between the two dimensions of sentence processing that are targeted in the present study. Accordingly, a previous study (Bohrn et al., 2013) has shown that Beauty ratings for semantically, grammatically, and phonologically very diverse single sentences consistently covary with distinctive neural activation patterns.

New parameter: Meter

While reusing the stimulus set of Bohrn et al. (2012), we added a new variable to its analysis. Bohrn et al. (2012) neither analyzed nor systematically modified any parallelistic features of the proverbs they used. However, such features are clearly important in sentences of the proverb type. Our English example "Eást or Wést, hóme is bést," features both meter (two metrically identical groups, i.e. "cretics," which consist of two stressed syllables with an unstressed one in between), rhyme (West, best), and an ongoing parallelistic series of monosyllabic words. The German proverb "Énde gút, álles gút" (which is included in our proverb set) likewise features two groups of cretics, (identical) rhyme and rigidly parallel word structure of the two syntactic-prosodic cola. Multiple parallelistic patterning of this type has since been shown to influence the cognitive and aesthetic processing of proverbs (Menninghaus et al., 2015). We here exclusively focused on the presence versus absence of meter as an additional covariate in our analysis, because only meter was found across a greater number of the stimulus sentences. Prosodic hyper-regularity due to meter has been shown to have strong effects on both cognitive and aesthetic processing (Menninghaus et al., 2017; Obermeier et al., 2016; Obermeier et al., 2013;).

A repeated check of all stimulus sentences for the presence versus absence of meter led us to retroactively exclude four items (across all four sentence categories) from the present study. Two of them are three-syllable proverbs ("Zeít ist Géld" and "Spórt ist Mórd"). If repeated, the metrical structure of these proverbs would be identical with "Énde gút, álles gút" and hence feature two cretics. As they are, however, the two sentences do not allow for an unambiguous assignment to either the meter or the nonmeter condition. On the one hand, they do feature a metrical building block used in many proverbs; on the other hand, they lack the repetition that first turns a single metrical foot (here, a cretic) into an ongoing meter. For the same reason, two four syllable-proverbs featuring a sequence of a trochaic and an iambic foot (e.g., "Wíssen ist Mácht" and "Ráche ist suéss") were also removed from the stimulus set. If repeated, these prosodic groups could readily be interpreted as two choriambs (each consisting of two stressed syllables surrounding two unstressed ones), just as in the case of other German proverbs or the English ad "Mélts in your moúth,/ nót in your hánd." However, in the absence of such repetition, it cannot be sufficiently determined whether "Zeit ist Geld" is actually a choriamb or simply missing any pronounced metrical structure.

Independent of the fact that we excluded four items across all four sentence categories, the number of metered versus unmetered sentences was not fully balanced across the four categories, as Bohrn et al. had not considered meter in their choice and experimental modification of sentence items. Nevertheless, we decided to use the remaining 36 original stimuli sets of four items each, because all additional changes we might have performed on the set would invariably have had negative influences on the variety of factors that were effectively controlled and confirmed not to be different statistically across the four groups of sentences, such as lexical parameters (e.g., number of words, digits, and syllables), mean word frequency, and emotional valence. Furthermore, using 90% of the original set still put us in a position to test the replicability of the previous rating results. The ratio of metered versus nonmetered items thus ended up being 25 versus 11 for the Familiar Proverb category, 14 versus 22 for the Synonyms, 19 versus 17 for the Creative Alterations, and 17 versus 19 for the Unfamiliar Proverbs.

Expected interaction between familiarity and meter in aesthetic ratings

Bohrn et al. (2012) reported a nonlinear relationship between familiarity and aesthetic rating scores for the sentences that we reused in the present study. The easiest-to-read sentences (i.e., familiar sentences) with the lowest cognitive demand were preferred over all other versions—i.e., over the unfamiliar rewordings of the original proverbs, the more innovative, but also more difficult to understand, Creative Alterations of the familiar proverbs, and wholly Unfamiliar sentences. The authors concluded that the cognitive fluency hypothesis was supported, although this hypothesis did not fully explain the results.

In the present study, we reexamined the interaction between familiarity-driven fluency and aesthetic ratings. Menninghaus et al. (2015) have shown that the presence of parallelistic features can enhance *perceptual fluency*, while at the same time reducing the ease of semantic comprehension (*conceptual fluency*). Getting back to the English example "East or West, home is best" and the German "Ende gut, alles gut:" The sustained parallelistic patterning of these sentences is achieved at the expense of omitting mandatory sentence parts (in these cases, the verb phrases). Such ellipses are likely to render the understanding more demanding, at least as long as the respective sentences have not become familiar and lexicalized as proverbs. Parallelistic structures in both poetry and proverbs and also prose routinely imply ellipses of this sort, albeit not necessarily as strongly as an ellipsis of the entire verb phrase.

Importantly, such ellipses and other cognitively demanding features often associated with highly parallelistic sentences are actually more than just handicaps. To be sure, as "negative prediction errors" in the sense of the predictive coding hypothesis of the

human brain (Friston, 2010; Rao & Ballard, 1999), they have been shown to be costly, as their processing recruits extra efforts of the brain (Thierry et al., 2008). But it is precisely this costliness that contributes to their memorability.

In order for parallelistic sentences that come with such cognitive handicaps to yield positive effects on processing fluency, the enhancing effects of parallelistic patterning on perceptual (prosodic) processing must in the end overcompensate its adverse effect on conceptual (semantic) processing ease. A previous study has provided first evidence for this assumption (Menninghaus et al., 2015). Accordingly, the authors suggested a variant of the cognitive fluency hypothesis that would account for antagonistic effects (enhancing *and* handicapping effects) of a given rhetorical feature on perceptual and conceptual processing ease, respectively (for this distinction, see also Reber et al., 2004), with perceived aesthetic appeal being dependent on the specific interaction of these potentially antagonistic effects. In the present study, we drew on this framework, as we also deal with two different types of linguistic fluency: One (i.e. meter) is driven by perceptual (prosodic) fluency only, whereas the other (i.e. familiarity) additionally, or exclusively, tends to affect the ease of processing of familiar semantic meaning. Given this difference in types of fluency, we examined how their interaction influences the perceived aesthetic appeal of sentences.

Finally, Bohrn and colleagues assumed that the Familiar Proverbs should be easiest and the entirely Unfamiliar ones most difficult to process, with the Synonymous versions and Creative Alterations covering the middle ground. However, they did not experimentally test this assumption. In the present study, we tested it by comparing the fixation dataset (dwelling time), which is indicative of cognitive processing effort, across the sentence categories. We controlled for potential significant differences in stimulus brightness. For each stimulus, the number of screen pixels that were colored with the stimulus color (blue) were counted. The numbers of screen pixels were compared by using one-way ANOVA, and we found no significant differences between the four categories, $F(3, 156) = 0.17$, $p = 0.920$.

Apparatus

Data were collected with an eye tracker (EyeLink 1000: SR-Research, Kanata, Canada) and two Windows PCs, one operating the eye tracker and the other controlling the experimental procedure via Psychtoolbox3 (www.psychtoolbox.org) and Cogent Graphics (http://www.vislab.ucl.ac.uk/cogent.php) running in MATLAB (Version 8.5.0; MathWorks, Natick, MA). The eye-tracking data were sampled at 1,000 Hz. A standard algorithm in the Eyelink software was used to detect oculomotor events (fixations, saccades, and blinks). Saccades were identified as periods during which an eye-movement velocity exceeded 30° (in visual angle [VA]) per second, periods missing pupils were classified as blinks, and the other periods were categorized as fixations. The pupil diameter was assessed in an ellipse pupil-tracking mode with a monocular setup (right eye). A five-point calibration followed by a validation was conducted at the beginning of each session, and a drift check and correction were performed after every four trials (see Task section for details). A chin and forehead rest was used to stabilize the participants' heads. Participants were seated 67 cm away from a 24-in. screen (XL2420Z; BenQ, Taipei, Taiwan) with a resolution of 1,920 × 1,080 pixels at 60 Hz. The same monitor was used in the eye-tracking and postrating tasks.

Task

The experimental procedure consisted of two parts, the eye-tracking part and the subsequent rating part. In the eye-tracking part, all data were collected in a dark room. At the beginning of the experiment, participants were instructed about the tasks and familiarized with the eye tracker. They sat comfortably, adjusting the height of the desk and chair. To minimize the effect of the pupillary light reflex (PLR), the intensity of the stimulus color (blue) was adjusted to be the same as that of the background color (dark yellow with RGB [50, 50, 0]) for each participant, using the flicker photometry method (Bartels & Zeki, 2006: http://www.vislab.ucl.ac.uk/cogent.php). All items presented in the experimental procedures that followed (e.g., instructions, fixation cross, stimulus, rating scales) were colored with the same intensity of blue, against the dark yellow background with RGB [50, 50, 0].

The task design largely followed Bohrn et al. (2012). Each participant completed eight sessions, each consisting of 20 trials. A session started with a five-point calibration followed by a validation, and drift check and correction were performed after every four trials. Each trial comprised three phases: the fixation (1,000 ms), stimulus (4,000 ms), and task (1,500 ms) phase. In the fixation phase, participants focused on a fixation cross (0.8° × 0.8° in VA) presented at the central position of the first word of the stimulus that would follow. Next, the stimulus (center-aligned one-line diction; Arial; 30 pt.; height 1.0°, width varying from 5.2° to 24.9°) was presented at the center of the screen, accompanied by another fixation cross (0.8° × 0.8°) at 3.5° below the text. Participants were instructed to read and understand the stimulus silently and to avoid

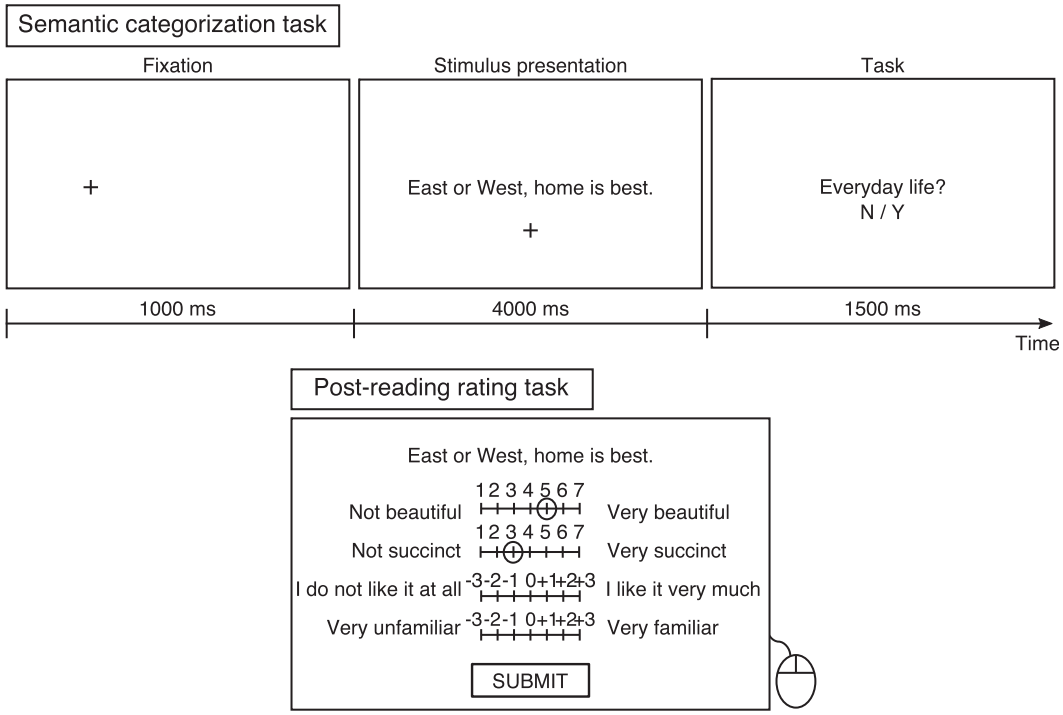

Figure 1. Schematic representation of a trial in the eye-tracking part (upper) and the post-reading rating part (bottom). The stimuli and background colors are shown in black and white in the figure, instead of blue and yellow, for better visibility.

rereading (and hence regressions) as much as possible. They were also instructed to focus on the fixation cross below the text after finishing the reading. Throughout the manuscript, we call this postreading fixation an "ending fixation." While the majority of reading-related eye-tracking studies applies a method in which the stimuli disappear immediately after participants stop reading (e.g., Fernández, Shalom, Kliegl & Sigman, 2014), we applied this "ending fixation" method to keep the presentation time of the stimuli constant (4,000 ms). Since we planned to apply the curve-fitting method to the pupillary waveforms (see Methods, Pupillary data section), we needed to obtain a dataset containing the same number of data points for all trials. In the task phase, participants performed a semantic categorization task for each stimulus. A category cue ("everyday life," "health and well-being," "love and relationship," or "work and success") was presented on the screen with the choices of no or yes ("N / Y"), and participants indicated whether or not the stimulus fitted into that semantic category by pressing a button within a response period of 1,500 ms, during which the category cue was shown. The results of the semantic categorization task were not analyzed further, since the main purpose of the task was to keep participants engaged with the stimuli and motivate the language processing. A schematic representation of a trial is shown in Figure 1. Note that, although participants were instructed to read and understand the stimuli, they did not perform any explicit tasks, such as making decisions or giving aesthetic evaluations *while* the stimuli were presented (the main time window of the data analysis). Each participant completed a total of 160 trials (20 trials × 8 sessions), responding to the 160 stimuli, which were each presented once and in randomized order. The eye-tracking part took approximately 20 to 25 min (varying by resting time between sessions).

In the subsequent rating part, each participant provided four explicit aesthetic judgments (posthoc ratings) for each sentence category. For the reasons explained in the Introduction, we collected Beauty, Liking, Familiarity, and Succinctness ratings. An additional goal in collecting the Familiarity ratings was to confirm that the categorization of the Familiar versus Unfamiliar Proverbs was valid.

Participants rated each stimulus on 7-point Likert scales; the Beauty and Succinctness ratings ranged from 1 (e.g., *not beautiful at all*) to 7 (e.g., *very beautiful*), and the Liking and Familiarity ratings ranged from −3 (e.g., *I do not like it at all*) to +3 (e.g., *I like it very much*). No further instructions were given concerning the definition of the four ratings. Following the studies by Bohrn et al. (2012) and Menninghaus et al. (2015), the anchor-points of the ratings for Beauty and Succinctness (+1 to +7) differed from those used for Liking and Familiarity (−3 to +3). To check participants' full engagement in the task, a dummy task was inserted after every 20 +/− 6 trials (seven trials in total) in which, instead of the stimulus, an instructional message such as "Please select the third option from the

right for all ratings" was presented (the location, e.g., "third from the right," was randomly chosen from seven options). To safeguard the good quality of the data, participants who made errors in more than one of these trial tasks were excluded from the analysis. The entire experimental procedure (eye-tracking and rating parts) was completed within 90 min.

Behavioral data processing

First, the rating scores were averaged across participants. To visualize the variance patterns across the rating scales, we plotted probability density maps (Figure 2). For each rating, each stimulus category, and each of the meter conditions (presence vs. absence), we estimated probability density using Kernel distribution with a bandwidth of 0.5 for the smoothing window (we used 0.5 since the difference between two adjacent levels is one, see also a manual for likert function of R software; https://cran.r-project.org/web/packages/likert/likert.pdf). When the ratings for a stimulus were nested, the Familiarity ratings were not distributed normally. Therefore, Kendall's Tau coefficient, which is applicable to nonnormal datasets, was calculated to confirm the relationships between the ratings. Kendall's method was used here, but not a Spearman correlation, following the method used by Bohrn et al. (2012).

Next, the effect of the stimulus category and the presence of meter on the rating score were tested by using a linear mixed-effect model (LMEM). The LMEM is often applied to behavioral and eye-tracking data (e.g., Demberg & Keller, 2008; Hohenstein, Laubrock & Kliegl, 2010). Each rating score (Beauty, Succinctness, Liking, and Familiarity) was subjected to LMEM with the fixed factors Category (four levels: Familiar Proverbs, Synonymous versions, Creative Alterations, and Unfamiliar Proverbs) and Meter (two levels: meter present vs. absent) and the interaction term between these factors. To take individual differences into account, a random intercept and random slopes (for all fixed predictors) were entered in the model for each participant (the LMEM introduced here is named "Model A"). The model was estimated using a maximum likelihood (ML) method from the Statistics and Machine Learning Toolbox running in MATLAB. Estimated fixed coefficients of fixed predictors were tested for the null hypothesis that the coefficients are equal to zero using a t test; we also assessed whether or not the predictor made a significant contribution to the model. Pairwise comparisons were conducted using Tukey-corrected tests.

As strong correlations between the ratings give rise to a multicollinearity problem when they are entered as predictors into the same regression model, we scrutinized these for underlying patterns using the axis-factoring method (factor analysis) with promax rotation. Factor analysis allows reducing dimensions of the subjective rating data (Beavers et al., 2013; Costello & Osborne, 2005). The original dataset (144 stimuli × 29 participants) with four variables (Beauty, Succinctness, Liking, and Familiarity scores) was used as input, and two principle factors were extracted which we named Affective and Cognitive factor, respectively (see Results, Behavioral results). The factor scores of the two extracted factors were calculated for each trial for each participant and used in the following analysis. We performed the factor analysis using SPSS software (Version 23.0.0.0; IBM, Armonk, NY).

Eye-tracking data processing

Fixation and saccadic data

In the stimulus phase (4,000 ms) of each trial, we calculated (a) the total duration of a fixation (dwelling time, including refixations) located in a region of analysis (ROA), and (b) the number of regressive saccades which were performed (i.e., launched and landed) within the ROA. The ROA was defined as a rectangular area covering the entire sentence line with a margin of 0.5° (in VA; half the size of the stimulus height) along all four sides. To test the influence of the stimulus category and the presence of meter, the dwelling time and regressive saccades were separately regressed by the LMEM (Model A). We also tested the effect of the rating scores (as broken down to the factor scores, see below) on the dwelling time and regressions, using another model (Model B) with two fixed covariates, the Affective factor and the Cognitive factor scores, a fixed factor, Meter (two levels: meter present vs. absent), and the interaction terms between each covariate and Meter. The model also had two random predictors: Participant and Category (four levels: Familiar Proverbs, Synonymous versions, Creative Alterations, and Unfamiliar Proverbs).

Apart from the analysis using LMEM, approximate reading times were calculated using the fixation dataset. To prevent participants from rereading and keep the duration of stimuli constant, each stimulus was presented with a fixation cross beneath it (see Task section for details). For each trial, we determined the latency of the first fixation spotted near the fixation cross (also called the "ending fixation"); this provides rough information about the time taken to read the stimulus. The ending fixation was defined as a fixation meeting the two following criteria: (a) The average location of the fixation is lower than 1.8° (in VA; half the distance from the text to the fixation cross), and (b) a saccade just before the fixation is in the right-to-left direction (starting from the text's end). In the following sections, we report the results

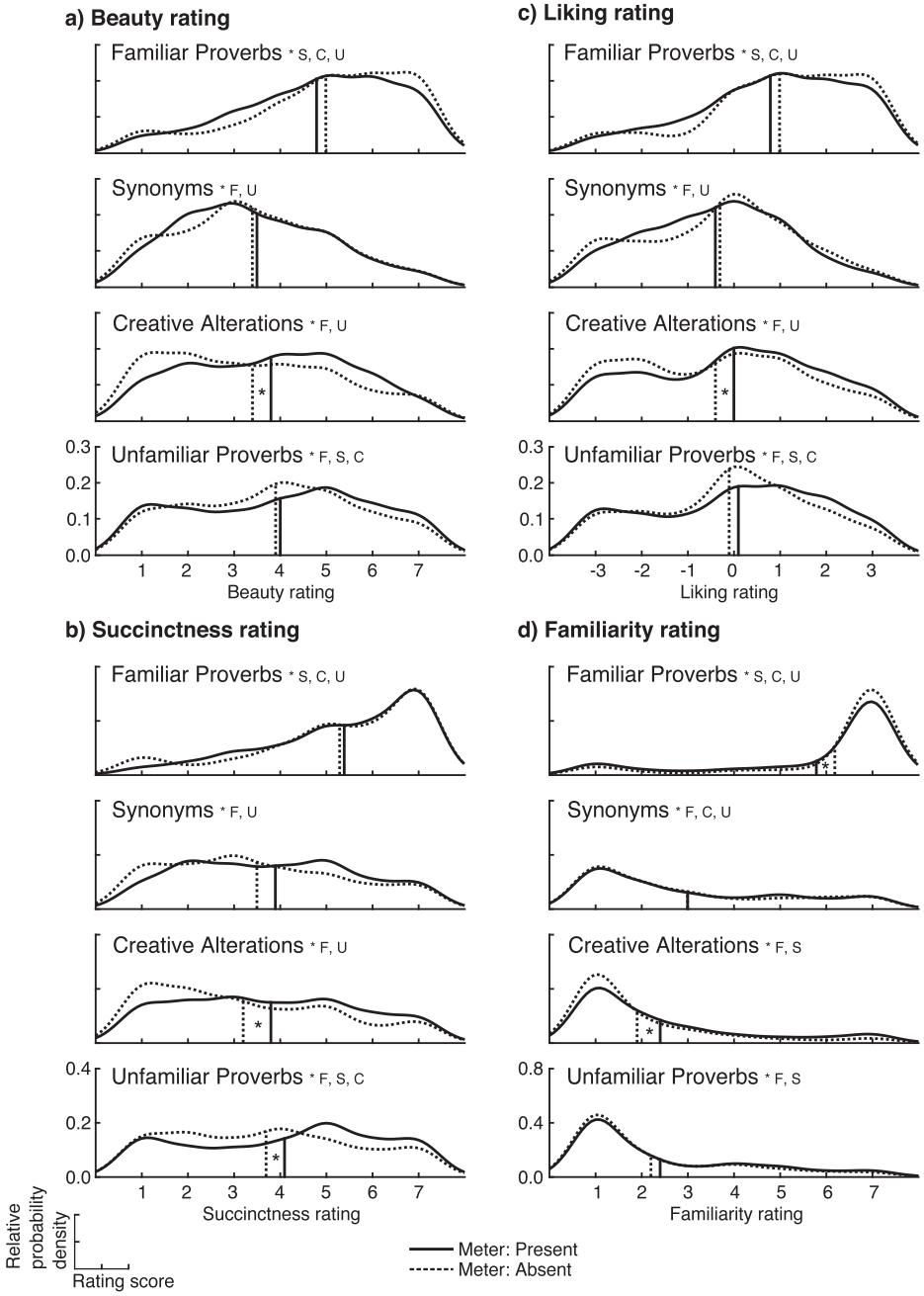

Figure 2. Probability density plots shown for each stimulus category and presence versus absence of meter. Vertical lines are drawn at the mean value of each distribution. Character(s) attached to the category name indicate that the average rating score of the respective category is significantly different from the scores of the other category(-ies) referred to by the initial characters (F = Familiar Proverbs; S = Synonyms; C = Creative Alterations; U = Unfamiliar Proverbs). Asterisks within the figures indicate significant differences dependent on the presence versus absence of meter.

for all trials, independent of the presence versus absence of an ending fixation of the above-defined type (see Results, Fixation and saccades data section and Discussion, Limitations section for more details). However, to identify potential differences, we performed additional analyses using the trials with ending fixations only (see Additional analysis section in the Supplementary materials, and Supplementary Figures S4–S6 and Supplementary Tables S8–S10).

Pupillary data: Preprocessing

Pupillary data were analyzed using MATLAB. Possible distortions of the pupil size measurements, which could be caused by the gaze position, were corrected by using trigonometry algorithms as suggested by Gagl, Hawelka, and Hutzler (2011) and Hayes and Petrov (2016; see Supplementary Figure S2 for details). Corrected pupillary data were divided into 5,000-ms

epochs covering the stimulus phase (4,000 ms) and the preceding fixation phase (1,000 ms). We used the manufacturer's standard algorithms with default settings to detect the blinking periods, and we discarded and linearly interpolated the pupillary data for those periods using the values measured 50 ms before and after each identified blink. The interpolated pupil time series were z-scored, based on the average and *SD* of the pupil diameter across the entire time series, and low-pass filtered (third-order Butterworth, 4 Hz, with the parameters following de Gee, Knapen, & Donner, 2014). Next, we baseline-corrected the data, using the 500-ms period preceding stimulus onset as the baseline. To test the global effects of the stimulus category on pupillary dilation and rating scores, as well as the effects of the presence of meter on the pupillary measures, the baseline-corrected pupillary data were averaged across the time window (stimulus phase: 4,000 ms) and then subjected to the LMEMs (Model A and Model B). For visual inspection, we also averaged them for each stimulus category and each rating score and plotted the pupillary waveforms (see Results, Pupillary data section).

Pupillary data: Polynomial curve fitting

To capture the dynamics of the pupillary responses, the response curve (during the stimulus phase: 4,000 ms) was fitted into a two-dimensional polynomial curve for each trial of each participant. We performed the analysis in two steps. First, using MATLAB's "polyfit" function, the coefficients for the polynomials (linear and quadratic) were calculated for each trial, such that the function best fits (in a least-squares sense) the observed pupillary response curve in the trial. To evaluate the accuracy of fitting, polynomial curves were drawn by using estimated coefficients, which were averaged between trials and subjects (Supplementary Figure S3). Second, each of the calculated coefficients (for the linear and quadratic components) was subjected to the LMEMs (Model A and Model B) separately. We thereby tested whether the quadratic/linear feature of the pupillary response curve was significantly modulated by the stimulus category, the presence of meter, and/or the perceived aesthetic appeal of the texts.

Results

Behavioral results

Kendall's Tau coefficients showed that the post-reading rating scores were significantly correlated between all pairs of the ratings (all p values < 0.001; Supplementary Table S1). In particular, the Beauty and Liking ratings correlated very strongly ($\tau = 0.81$). These results closely replicated those of the study by Bohrn et al. (2012).

The LMEMs and multiple comparisons revealed that the scoring patterns were similar for the Beauty and Liking ratings (see Figure 2a and 2c and Supplementary Table S2a and S2c): The Familiar Proverbs scored higher than all other categories, the Unfamiliar ones received the second highest scores, and the other stimulus types (Synonymous and Creative Alteration of proverbs) received the lowest scores. The presence of Meter modulated the Beauty and Liking scores only when the stimuli belonged to the Creative Alteration category (the interaction term between the Creative Alteration category and Meter was significant in the LMEMs, and the pairwise comparisons that followed showed a significant difference); it did not have a general effect on the rating scores (i.e., the main effect of Meter was not significant). On the Succinctness rating (see Figure 2b and Supplementary Table S2b), the Familiar Proverbs scored highest, the Unfamiliar Proverbs received the second highest score, and the Synonyms and Creative Alterations received the lowest score.

Regarding the Succinctness ratings, the Meter factor heightened the scores for Creative Alterations and Unfamiliar Proverbs (the interaction terms in the LMEMs and the multiple comparisons show the significance), but not for the Familiar Proverbs and Synonyms. Finally, regarding the Familiarity rating (see Supplementary Figure S2d and Supplementary Table S2d), the Familiar Proverbs scored highest and the Unfamiliar ones scored lowest, suggesting that the categorization was reasonable. The Creative Alterations scored as low in Familiarity as the Unfamiliar ones. Furthermore, the presence of meter significantly increased the Familiarity score for Creative Alterations, whereas it reduced the score for the Familiar category. We thus replicated the finding of Bohrn et al. (2012) that the Familiar Proverbs scored very highly on Succinctness and Familiarity; the distribution patterns were heavily skewed in a positive direction (see Figure 2b and 2d). In contrast, Familiarity ratings for the other three categories (Synonyms, Creative Alterations, and Unfamiliar Proverbs) yielded negatively skewed distribution patterns. This implies that the Succinctness and Familiarity ratings accounted for a greater portion of the differences between the four categories than the other two ratings.

We used factor analysis on the four rating-data sets to create factors which could subsequently be used as predictors in the regression analysis. First, the factor analysis was run without any specification of the number of factors (exploratory method). Three factors were extracted with eigenvalues 2.72, 0.70, and 0.38, and these explained 68.1%, 17.4%, and 9.4% of the overall variance, respectively. The maximum factor

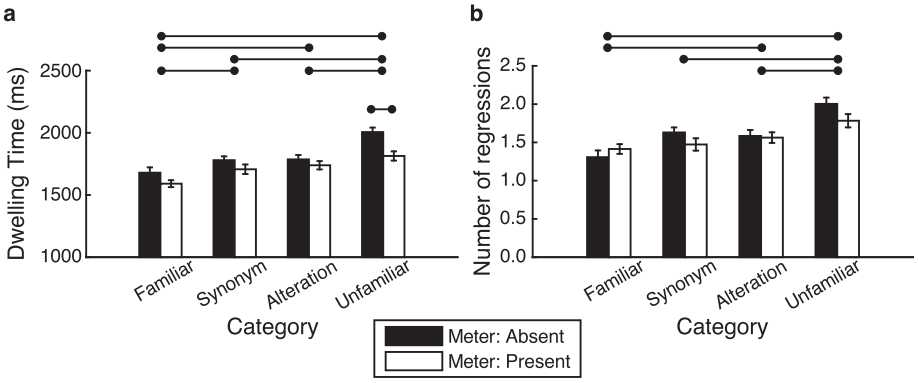

Figure 3. (a) Average dwelling time shown for each stimulus category and presence versus absence of meter. (b) Average number of regressive saccades shown for each stimulus category and presence versus absence of meter. Error bars indicate the *SE*s. The results of pairwise comparisons are represented by horizontal bars ($p < 0.05$, Tukey corrected).

loadings were also inspected (0.87, 0.67, and 0.27; see Supplementary Table S3). The eigenvalue of the second factor, 0.70, failed to meet the Kaiser criterion (Kaiser, 1960), the popular cut-out threshold for factor extraction, which suggests that factors with an eigenvalue greater than one (>1.0) be taken as primary factors. However, our choice of the two-factor model seems reasonable for three reasons. First, the explained variance and the maximum loading factor declined considerably in the third factor (Supplementary Table S3). Second, the two-factor model is in good fit with our theoretical assumptions advanced in the Introduction, Subjective ratings section, namely, that familiarity and succinctness judgments are of a more cognitive nature, whereas beauty and liking judgements are widely considered as being more affective. Third, several studies suggest limitations of the Kaiser criterion (for reviews, see Beavers et al., 2013; Costello & Osborne, 2005), showing that the application of the criteria could overextract or underextract an appropriate number of factors.

Therefore, the factor analysis was run again, this time with the specification that the number of factors to be extracted be two. The extracted factors explained 85.5% of the overall variance (primary factor: 68.1%, secondary factor: 17.4%; see Supplementary Table S4). A factor-loading matrix and the correlation between factors are shown in Supplementary Table S4. There is a clear discrimination between the primary (I) and secondary (II) factors, the former showing remarkably high factor loadings for the Beauty (0.94) and Liking (0.83) ratings and the latter high loadings for the Succinctness (0.67) and Familiarity (0.72) ratings. The other factor loadings were 0.21 or less.

Because Liking is by its very definition an affective judgment and Beauty has time and again been analyzed for its affective implications and qualities (Armstrong & Detweiler-Bedell, 2008; Schindler et al., 2017), the primary (I) factor was labeled as the "Affective" (emotional) factor. Similarly, because Familiarity is the prime variable underlying the cognitive fluency hypothesis of aesthetic judgment and Succinctness (*praegnanz*) is a key driver of an image's or message's cognitive strikingness (cf. the "law of praegnanz"; Koffka, 1935; Wertheimer, 1923), the secondary (II) factor was labeled as the "Cognitive" factor. The factor scores for each factor are henceforth referred to as the "Affective score" and the "Cognitive score." Notably, the inter-factor correlation is still high (0.73); this may cause problems of multicollinearity when the two factors are used as regressors in the same model (see Discussion, Limitations section, for details).

Eye-tracking results

Fixation and saccades data

We used the LMEMs (Model A and Model B) to regress the average dwelling time (see Figure 3a and Supplementary Table S5a) and number of regressive saccades (see Figure 3b and Supplementary Table S5b). For the dwelling time data, the stimulus category factor was significant in Model A: Familiar Proverbs had the shortest dwelling times and the Unfamiliar ones the longest, with the other sentence types ending up in between. The Meter factor had general influence on the dwelling time; dwelling times were shorter for metered than nonmetered stimuli (the meter factor was significant in Model A in Supplementary Table S5a). The posthoc tests revealed a significant difference of dwelling times for metered and nonmetered Unfamiliar Proverbs, but not for the other sentence categories (Figure 3a), indicating that the meter influence was particularly salient for the Unfamiliar sentence variants.

The significant influence of the Meter factor was also evident in Model B (Model B in Supplementary Table S5a); in this model, dwelling times were in general shorter for metered than nonmetered stimuli. More

interestingly, Model B showed that the Cognitive factor, as a fixed covariate, had a significant negative influence on the dwelling time ($\beta = -123.21$), whereas the Affective factor did not have a significant effect.

The stimulus category also modulated the number of regressions significantly (Model A of Supplementary Table S5b): Familiar Proverbs had the fewest regressions and the Unfamiliar ones the highest number of regressions, with the other sentence types ending up in between. However, we did not find significant effects of Meter and aesthetic appeal (as measured by the Affective and the Cognitive factor) on the regressions.

Average stimulus reading time was estimated by detecting the ending fixation. The ending fixation was successfully detected, on average, for 70.2% of the trials per participant. In the other trials, participants either kept reading or remained fixating on the respective sentence until the end of the stimulus phase. As already indicated, we here report the results for all trials. Additional analyses performed on the trials with ending fixation only (70.2% of all trials) did not show major differences compared to the overall results (see Additional analysis section in the Supplementary materials, Supplementary Figures S4–S6 and Supplementary Tables S8–S10, and Discussion, Limitations section). The average reading time calculated from the 70.2% of the trials was 2.8 s ($SD = 0.3$ s).

Pupillary data: Visual inspection and average size

We averaged the preprocessed pupillary waveforms for each stimulus category and each rating score (see Figure 4). They showed several common features: rapid mydriasis and following miosis in the very early period of the stimulus phase (from 0 to 750 ms; possibly corresponding to the PLR), separation of the waveforms (for each category or each rating score) after around 750 ms, mild miosis (at around 750 to 2,500 ms), and a following mydriasis phase (after 2,500 ms). The separated waveform for each stimulus category (Figure 4a) showed a clear correspondence to the fixation data: For the Unfamiliar Proverbs, which had the longest dwelling times, the smallest pupil size was observed; for the Familiar Proverbs, which had the shortest dwelling times, we observed the largest pupil size. Responses to the other two sentence types covered a middle ground. The pupillary response waveforms for the aesthetic ratings (Figure 4b through e) also diverged for the respective rating scores after 750 ms, but the pattern was not as clear as the waveforms for the stimulus category (Figure 4a), possibly because multiple underlying factors were simultaneously operative and caused a mixed response of miosis and mydriasis.

To capture the global effect of the stimulus category, the aesthetic appeal, and the presence versus absence of meter on the pupil size, we used the LMEMs to regress the average pupil size (see Figure 5a and Supplementary Table S6). In Model A, the Category factor was marginally significant: The pupil size for Familiar Proverbs was smaller than for Unfamiliar ones (Model A in Supplementary Table S6). The Meter factor did not predict any significant modification. In Model B, the two fixed covariates, the Affective and the Cognitive factor, showed (marginally) significant modulations on the average pupil size (see Supplementary Table S6). Furthermore, the estimated coefficients pointed to opposite roles of these two covariates (Affective score: $\beta = 0.106$, Cognitive score: $\beta = -0.111$), indicating that they modulated the pupil size in an antagonistic fashion.

Pupillary data: Polynomial curve fitting

We further investigated these data using the curve-fitting method. The pupillary waveforms from the stimulus presentation period were fitted into a two-dimensional polynomial curve for each trial. The polynomial curves were drawn using estimated coefficients (Supplementary Figure S3); they confirm that the curvatures in the original waveforms (Figure 4) were adequately captured by polynomial coefficients. Subsequently, we regressed the estimated coefficient vectors separately for the linear and quadratic components, using the LMEMs. The linear component of the pupillary waveforms was modulated antagonistically by the Affective factor and the Cognitive factor of processing the stimuli (Affective score: $\beta = 0.076$, Cognitive score: $\beta = -0.094$; see Model B in Supplementary Table S7a).

Regarding the quadratic component, the Unfamiliar Proverbs showed significantly smaller components than the other three sentence categories (see Figure 5c and Model A in Supplementary Table S7b). This corresponds to the actual pupillary waveforms visualized in Figure 4a; the pupillary data after 750 ms in Unfamiliar Proverb category showed relatively milder miosis compared to the other three categories. The Meter factor modulated the quadratic component significantly (see Model A and Model B in Supplementary Table S7b), indicating that the metered stimuli generated pupillary response curves with steeper slopes than the nonmetered ones.

Discussion

The present study investigated the relations between eye-tracking parameters (fixations and pupil measurements), familiarity, meter (hyper-regular prosody), and the perceived aesthetic appeal of four sentence variants that are all based on the template of proverbs. We report

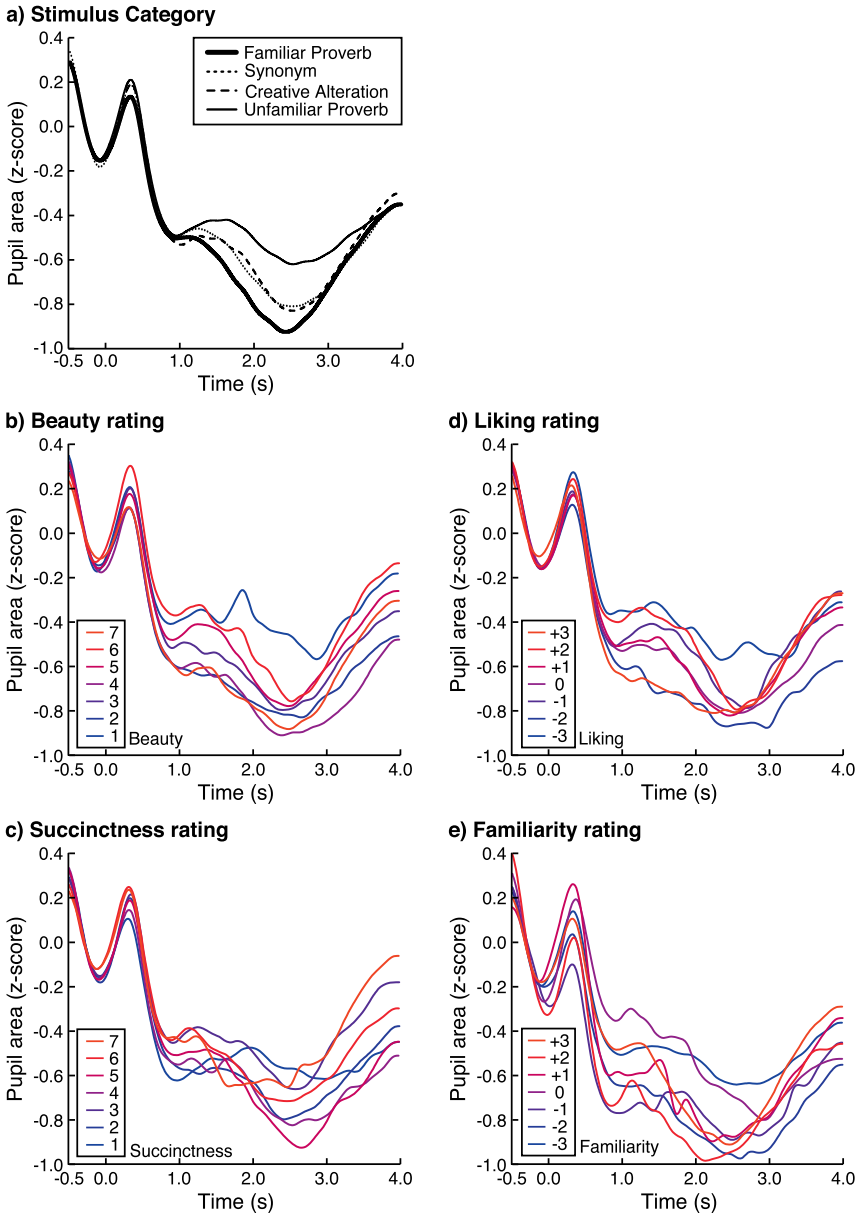

Figure 4. Pupillary waveforms averaged across 29 participants, for (a) each stimulus category, (b) Beauty ratings, (c) Succinctness ratings, (d) Liking ratings, and (e) Familiarity ratings. The negative time window corresponds to the fixation phase, the time zero is stimulus onset, and the positive time window corresponds to the stimulus phase. The data were baseline corrected using the fixation phase (−500 to 0 ms) as baseline.

two major findings. First, the ratings for subjectively perceived aesthetic appeal—as measured by Beauty, Succinctness (*Praegnanz*), and Liking ratings—are modulated by the interaction of two factors: the level of familiarity and the presence versus absence of meter. This corroborates previous results that either focused on ratings of familiarity only (Bohrn et al., 2012) or on meter effects only (Menninghaus et al., 2015), but not on their interaction, and that also did not include eye-tracking measures. Second, and even more importantly, we provide evidence that both the affective and the cognitive dimensions of the perceived aesthetic appeal of the sentences modulate the eye-tracking parameters. Specifically, we found that the two dimensions influenced the pupillary data antagonistically. We further discuss each finding in the following subsections.

Interaction of familiarity- and meter-driven ease of processing

Previous studies have used Beauty and Liking ratings as umbrella terms that broadly capture perceived aesthetic appeal. The present results confirm that

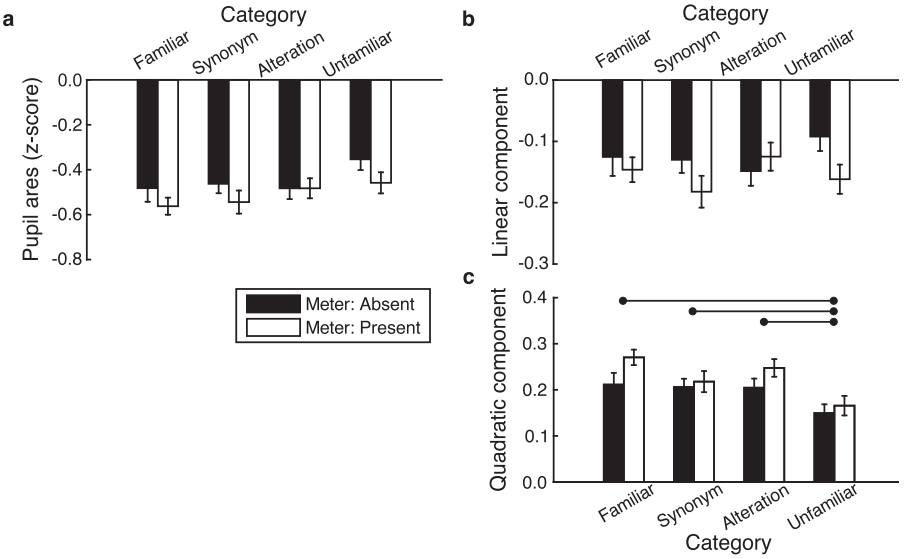

Figure 5. (a) Average pupil size (*z*-scored), (b) estimated linear component, and (c) quadratic component of the pupillary response curve during the stimulus phase, shown for each stimulus category and presence versus absence of meter. Error bars indicate the *SE*s. The results of pairwise comparisons are represented by horizontal bars ($p < 0.05$, Tukey corrected).

the two ratings are closely correlated ($\tau = 0.82$; Supplementary Table S1). Replicating the study of Bohrn et al. (2012), we observed a nonlinear correlation between aesthetic rating scores and familiarity-driven ease of processing (Figures 2a and 2c and Supplementary Table S2a and S2c): The Familiar Proverbs were liked most among all categories, supporting the cognitive fluency hypothesis.

The Unfamiliar original Proverbs ranked second in the Beauty and Liking ratings. At the same time, they had the lowest level of familiarity and longest dwelling time, suggesting that they are most difficult to understand and hence cognitively most challenging. This finding implies that familiarity-driven ease of processing alone by no means predicts aesthetic liking in a linear fashion. Rather, our results are in line with the study by Menninghaus et al. (2015), which experimentally modified a set of 30 unfamiliar proverbs and showed, regarding four versions of the same unfamiliar proverbs, that higher rather than lower cognitive demand went in tandem with the highest scores for both Beauty and Succinctness. The underlying rationale was that only the original proverbs feature meter and rhyme, and that meter and rhyme specifically enhance prosodic ease of processing, while often simultaneously rendering semantic understanding more demanding. After all, meter and rhyme exert significant constraints on both word choice and word order; their implementation therefore often requires a choice of words with less than optimal semantic fit and noncanonical syntactic order (e.g., inversion, ellipses; Fitzgerald, 2007; Leech, 1969; Levin, 1962; Rice, 1997; Youmans, 1983; see also the example given in Stimuli, New parameter section). Still, these cognitive downsides of implementing meter and rhyme schemes were overcompensated by the meter and rhyme-driven benefits, as these render the cognitively more demanding sentence versions more succinct, more beautiful, and also more persuasive than the less demanding versions.

Our findings regarding the Unfamiliar Proverbs used as the fourth sentence category in the present study can be readily explained along these lines. After all, both Familiar and outdated, and hence Unfamiliar, proverbs heavily rely on parallelistic patterns as message enhancement devices (Menninghaus et al., 2015) and are therefore likely to score highly in our aesthetic ratings, while simultaneously requiring high cognitive processing effort. Precisely this cognitive adverse effect of making a sentence both more parallelistic and more succinct–which in Familiar Proverbs is alleviated or even erased by virtue of their repeated prior exposure—is indicated by the fact that these original Unfamiliar Proverbs show the longest dwelling time among all categories (see Eye-movement data reflect the cognitive processing of language). The fact that, in our study, the Familiar Proverbs still scored more highly in aesthetic ratings than the Unfamiliar ones is therefore likely to be entirely due to their familiarity.

The finding that the two modified variants of proverbs (proverb Synonyms and Creative Alterations) scored lower in aesthetic ratings than both Familiar and Unfamiliar original proverbs is also in line with the assumptions mentioned above. Both types of proverb modifications reactivate the syntactic and semantic template of the underlying original proverb while at the same time deliberately departing from its familiar wording. This experience of deviation constitutes a

marked negative prediction error known to exert extra demand on cognitive processing (Pickering & Garrod, 2007). At the same time, the modified proverb variants featured meter (and other rhetorical features, such as rhyme and syntactic parallelism) to a lesser degree than the original proverbs, be they familiar or unfamiliar. As a result of both the mismatch experience and the reduced level of rhetorical features, both kinds of modified proverb variants had lower aesthetic ratings than the two original kinds of proverbs.

Influence on Succinctness ratings

Succinctness ratings were higher for metered than for nonmetered sentences that are particularly low in familiarity (i.e., Creative Alterations and Unfamiliar Proverb categories; see Figure 2b and Supplementary Table S2b). In the study by Menninghaus et al. (2015), the presence of meter also had a positive effect on succinctness. However, this likewise applied—and even to the most pronounced degree—to the original proverbs. We suggest two reasons for this divergence in findings. The first and preeminent reason is that Menninghaus et al. (2015) exclusively used *nonfamiliar* proverbs dating back to the 19th century; hence, the study could not possibly report reduced Succinctness ratings for highly familiar metered compared to highly familiar nonmetered proverbs. Moreover, the study compared original metered proverbs with experimentally altered nonmetered versions, whereas the present study did not perform within proverb-comparisons for proverbs of the same content, but compared genuine proverbs of different content which were either metered or not.

Secondly, the finding may also allow for another explanation. According to the cognitive fluency hypothesis, familiarity and parallelistic patterning separately enhance ease of processing, the former in accordance with the mere exposure-effect, and the latter because parallelistic diction renders the perceptual (mostly prosodic) processing of sentences more smoothly and hence facilitates it. At the same time, many "successful" proverbs—i.e., those that ended up becoming lexicalized—do not feature meter. 34 of the 80 original proverbs (Familiar or Unfamiliar) included in the original stimulus set of Bohrn et al. (2012) are actually nonmetered.

In the English language, a preliminary inspection of a collection of proverbs suggests that the percentage of nonmetered proverbs appears to be even higher than in German. This implies that levels of succinctness characteristic of successful proverbs do not mandatorily require meter. Hence well-accepted proverbs without meter are likely to reach similar levels of succinctness by way of other properties. They can feature alliterations and other parallelistic features which we did not systematically include in our analysis. They can also enhance succinctness by rhetorical means other than parallelistic patterning, for instance, by employing striking metaphors.

Influence on Beauty and Liking ratings

Meter also had an effect on the Beauty and Liking ratings, albeit less generally than on the Succinctness ratings (Figure 2a and 2c in Results, Behavioral section). Specifically, a significant difference between the presence and absence of meter was exclusively found for sentence variants in the Creative Alteration category—the category that involves a particularly striking semantic deviation from the underlying familiar proverbs. Genuine proverbs stress the importance of both moral and social values and of concrete rules of behavior (their lessons being, for instance, "work hard," "don't be lazy," "be honest," etc.). Many of these rules are not exactly objects of liking, but rather of norm compliance; therefore, they are occasionally followed only unwillingly, and certainly not with enthusiasm. In this regard, the original proverbs (both Familiar and Unfamiliar) are not different at all, and the Synonymous variants of the original proverbs are at most marginally different.

The Creative Alterations, by contrast, are enjoyable both for their creativity and for giving the conventional lessons taught by proverbs a decidedly antiauthoritarian twist. They are, in other words, the only sentence variants in our study that offer rewards that are at least remotely comparable to reading an original work of literature. A study on humoristic verses (Menninghaus et al., 2014) has shown that the presence versus absence of meter in these verses significantly increases liking and humor ratings. The humor underlying the anticonventional Creative Alterations of proverbs draws on similar resources. This may explain why meter here also pushed the affective factor to higher levels.

Moreover, the latent evocation of Creative Alterations of a proverb is likely to profit from the double support of both converging syntax and converging prosody, the latter particularly in cases of metered prosody. This double bond to the original, yet semantically distorted, template should be more effective than a single associative bond. Given that meter has been shown to facilitate memory retrieval (Tillmann & Dowling, 2007), it should in these cases have an additional facilitating effect on the spontaneous recall of the original proverbs. Following the cognitive fluency hypothesis, this facilitation effect driven by meter should translate into both higher perceived familiarity (Figure 2d) and higher perceived aesthetic appeal (Figure 2a and 2c), thus explaining the two significant meter effects found specifically for the Creative Alterations.

Influence on Familiarity ratings

Metered Familiar Proverbs scored lower in Familiarity than those without meter (Figure 2d). Here we offer two possible explanations for this finding which seems counterintuitive at first sight. One potential source of the reduced perceived familiarity found for the metered proverbs may be the very nature of rigidly metered diction. Such diction by definition deviates from standard language use. As a result, if participants responding to the familiarity question do not focus on the familiarity of the sentence as a lexicalized expression, but on the familiarity of the special type of diction it employs, then the proverbs are actually, by virtue of their "artificial" meter, less familiar than nonmetered language use.

Another explanation could, inversely, understand the reduced Familiarity ratings for metered proverbs as a negative effect of the relatively high frequency of rhymed and/or metered sentences among German proverbs. To the extent that meter is a highly expectable feature rather than an exceptional virtue (which applies to German proverbs), proverbs that are successful and lexicalized *without* relying on meter are likely to feature less generic and hence more individualized wording. In this sense, nonmetered proverbs could be more unique than metered ones, and precisely this higher distinctiveness compared to the metered standard could well support a better storage in memory and hence higher Familiarity ratings.

Summary of the rating data

In sum, in this study we provide the first evidence that both familiarity- and meter-driven ease of processing interactively influences ratings for aesthetic appeal. In line with Menninghaus et al. (2015), we emphasize that the perceived fluency of sentence processing cannot be explained by a single property of the sentences. Rather, perceived fluency is driven by a complex interaction of semantic, syntactic, and prosodic features and the level of familiarity with a sentence. All of these processing dimensions underlie the overall perceived processing ease and influence the perception of aesthetic appeal. In the following subsections, we discuss how perceived fluency and perceived aesthetic appeal correlate with the eye-tracking measures we collected.

Eye-movement data reflect the cognitive processing of language

The oculomotor parameters, the dwelling time and regression data, show a linear-pattern relationship with familiarity-driven ease of processing (Figure 3 and Model A in Supplementary Table S5a and S5b). Among the four sentence categories, Familiar Proverbs had the shortest dwelling times and the fewest regressions, and Unfamiliar ones the longest dwelling times and the highest number of regressions. Synonymous versions and Creative Alterations of the familiar proverbs ended up between these extremes, presumably because as novel variants of familiar proverbs they are neither completely familiar nor completely unfamiliar. This finding confirms that the stimuli in each category systematically exerted different levels of cognitive demand. Please note that the stimuli were controlled for potential differences in lexical parameters, mean word frequency, and emotional valence (see Stimuli, Overview section). Hence any differences that we observed between the four sentence categories are not accounted for by these factors.

The results for the dwelling times are in line with the classic study by Ehrlich and Rayner (1981) which showed that sentences containing a less predictable word (in our case, Synonyms and Creative Alterations) generate longer dwelling times. The results shown for the regressive saccades also corroborate previous studies suggesting that the number of regressions increases when readers struggle or fail to identify a word (Pollatsek & Rayner, 1990; Shebilske, 1975), when they experience comprehension failure associated with high-level syntactic and semantic processes in understanding a text (Bouma & DeVoogd, 1974; Just & Carpenter, 1980; Shebilske, 1975; Shebilske & Fisher, 1983), or when they process more cognitively demanding texts (Shreve, Lacruz, & Angelone, 2010). Importantly, we instructed participants to avoid making regressions during reading as much as possible (see Task section). However, despite this instruction, we did find the number of regressions to be modulated significantly by stimulus category such that the less familiar and hence more demanding sentence categories elicited more regressions. This indicates that the differences in textual characteristics were powerful enough that the readers could not fully suppress or control the potential effects on regressive saccades.

Additionally, we showed that the dwelling times were reduced for metered compared to nonmetered stimuli (Figure 3a and Supplementary Table S5a). Contrary to what was found for the aesthetic evaluations, the beneficial effect of meter-driven fluency on the dwelling time parameters seems limited; it was exclusively found for Unfamiliar Proverbs. This may indicate that the familiarity of the syntactic and semantic sentence templates of the original proverbs—that underlie all Synonyms and Creative Alterations and are reactivated while processing these two modified versions—is by itself so strong a predictor of processing fluency (as measured by dwelling time) that meter does not add to it anymore. This provides the first eye-tracking evidence that regular meter enhances pro-

cessing fluency. Moreover, the difference was more pronounced for the category of completely Unfamiliar Proverbs compared to the three categories containing proverbs of three gradations of Familiarity. This result shows an interaction of meter- and familiarity-driven fluency such that the effect of one source of fluency (here, Familiarity) on dwelling time parameters can override, or even cancel out, the effect of another potential source of fluency (here, meter) on the same dependent variable.

Regarding the relationship between the rating scores and dwelling times, the factor analysis and LMEM (Model B in Supplementary Table S5a) offer an interesting result. They show that the Cognitive score has a negative effect on the dwelling time: As the Cognitive factor (i.e., Familiarity and Succinctness ratings) increases, the dwelling time decreases. This is in line with the cognitive fluency hypothesis: Easier-to-read texts, which require shorter dwelling time for semantic comprehension, are more aesthetically appealing.

In sum, we found that oculomotor parameters of sentence reading are systematically modulated by our experimental modification of Familiarity in four gradations (familiarity-driven fluency) as well as by the presence versus absence of Meter (i.e., meter-driven fluency). Specifically, combined high ratings for Familiarity and Succinctness (i.e., the Cognitive factor) predict lower dwelling time, thus confirming the cognitive fluency hypothesis also on the level of eye movement parameters.

Dynamic modulation of pupillary response by the aesthetic appeal of the stimulus sentences

The most obvious result in the pupillary data is the modulation by the stimulus category at around 500 to 3,500 ms after stimulus onset. The waveforms are clearly distinct for each category (Figure 4a); this significant difference between the sentence categories is also reflected in the quadratic component of the response curve (Figure 5c and Supplementary Table S7b). The pattern was similar to that obtained for the eye-movement parameter, i.e., we found a linear relationship with familiarity-driven ease of processing. This confirmed that not only the eye-movements, but also the pupil size reflected the amount of cognitive load.

Remarkably, the extracted Affective and Cognitive factors each evoked different pupillary dilation patterns. The LMEM on the average pupil size and linear component of the response curve (Model B in Supplementary Table S7a) showed that the two factors had antagonistic influences. The Affective scores increased the pupil size ($\beta = 0.105$, $p = 0.023$) and linear component ($\beta = 0.076$, $p = 0.006$), whereas the Cognitive scores decreased them (pupil size: $\beta = -0.112$, $p = 0.063$, linear component: $\beta = -0.094$, $p = 0.001$). This implies that the overall size and dilation of the pupil (the slope) while reading increased when the respective sentence evoked positive aesthetic appreciation (with a high Affective factor score), but decreased when it primarily reflected cognitive fluency (easy to read, with a high Cognitive factor score). This dual innervation of the two aesthetic appeal factors generated the complex waveforms (Figure 4b through e). Importantly, we used the LMEM with random intercepts and slopes for each stimulus category and participant, thereby ruling out the possibility that the effects were due to the intervention of the stimulus category and/or individual differences.

Summing up, our study's principal finding is that sentences differing in perceived aesthetic appeal as measured by subjective ratings also produced significantly different pupillary response curves over the course of time. The subjective aesthetic ratings were in accordance with both the cognitive fluency hypothesis (Reber et al., 2004; Reber et al., 1998; Reber & Schwarz, 1999) and the expectation that aesthetic appeal also relies on some departure from more common language use (Giora et al., 2004; Miall & Kuiken, 1994; Miall & Kuiken, 1998; Shklovsky, 1917) and hence may also involve higher cognitive effort (Menninghaus et al., 2015). Taken together, these findings suggest two partly antagonistic sources underlying aesthetic appeal. Consistent with these behavioral findings, the affective and cognitive dimensions of the perceived aesthetic appeal yielded contradictory effects on the objective measure of pupil dilation, as well.

Throughout the history of aesthetics, the production and response to artworks have both been conceived as involving an ongoing antagonism between reason (learned skills, cognitive and technical abilities) and affective engagement (often treated under the rubrics of "passion," "enthusiasm," and even "frenzy" and "madness"). Nietzsche's antagonism of the "Apollonian" and the "Dionysian" (Nietzsche, 1872 [trans. 2000]) famously emphasized that the two poles—for all their antagonism—do not designate two separable entities, but co-occur in different combinations in individual artworks. Our findings suggest that the pupillary dilation of readers may reflect both sides of this antagonism in the processing of single sentences.

Limitations

First, the major limitation of the present study is the limited presentation time of the stimulus (4,000 ms). A replication of our results with longer sentences/texts is clearly called for.

Second, the ending fixation was not successfully detected for 29.8 % of all trials. We suggest two possible reasons for this result: (a) Participants forgot or disregarded the instruction, "please fixate on the fixation cross beneath the text after you have finished reading;" or (b), the respective stimuli were cognitively so challenging for them that they could not but keep reading until the end of each trial. To be sure, our additional analyses (see Results, Fixation and saccades data section, Additional analysis section in the Supplementary materials, and Supplementary Figures S4–S6 and Supplementary Tables S8–S10) show that the results for the data with ending fixations do not show major differences from the overall results. Still, the absence of ending fixations in nearly 30% of the trials does call for further investigation.

Another limitation is the unavoidable multicollinearity issue in the LMEM regression analysis. The two nonindependent factor scores, for Affective and Cognitive factors, were used as predictors. However, as already discussed in the Introduction, it is not realistic to assume that the two predictors—the affective and cognitive aspects of the perceived aesthetic appeal of the stimulus sentences—could actually be independent from one another. After all, we can hardly be emotionally affected by sentences without some level of cognitive understanding. Use of nonsubjective aesthetic measures of the texts, such as certain linguistic features correlated to the text's aesthetic appeal, would be a possible solution to this problem, one that is worth testing in the future.

Conclusion

The present study aimed to identify characteristic features of eye-tracking parameters (the pupillary and fixation dataset) that reflect the aesthetic appeal dimensions of single sentences of the proverb-type. The two dimensions of aesthetic appeal, the Affective and the Cognitive factor, influenced both the fixation and the pupillary measures in distinct and significant ways. In line with the cognitive fluency hypothesis, a higher Cognitive factor predicted shorter dwelling times and smaller pupil dilations. By contrast, a higher Affective factor predicted larger pupil dilations. Moreover, our study reveals that a significant interaction of familiarity-driven and meter-driven ease of processing influences perceived aesthetic appeal. Extending previous findings by Menninghaus et al. (2015) and Bohrn et al. (2012), our study thus provides a comprehensive account of the complex and nonlinear interaction of processing fluency and aesthetic appreciation, and illustrates a possible application of the eye-tracking method as an objective and convenient measure capable of capturing the aesthetic appeal of language.

Keywords: pupillary response, eye-tracking, aesthetic appeal, language processing

Acknowledgments

We thank Dr. Wolff Schlotz and Dr. Sebastian Wallot for invaluable advice on the statistical methods and comments on an earlier version of the manuscript. We also thank Dr. Jan Auracher for comments on an earlier version of the manuscript.

Commercial relationships: none.
Corresponding author: Hideyuki Hoshi.
Email: hideyuki.hoshi@ae.mpg.de.
Address: Department of Language and Literature, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany.

References

Armstrong, T., & Detweiler-Bedell, B. (2008). Beauty as an emotion: The exhilarating prospect of mastering a challenging world. *Review of General Psychology*, *12*(4), 305–329. Retrieved from https://doi.org/10.1037/a0012558

Augustin, M. D., Wagemans, J., & Carbon, C. C. (2012). All is beautiful? Generality vs. specificity of word usage in visual aesthetics. *Acta Psychologica*, *139*(1), 187–201.

Bartels, A., & Zeki, S. (2006). The temporal order of binding visual attributes. *Vision Research*, *46*(14), 2280–2286. Retrieved from https://doi.org/10.1016/j.visres.2005.11.017

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (2nd ed., pp. 142–162). Cambridge, UK: Cambridge University Press.

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research & Evaluation*, *18*:6.

Blackburn, K., & Schirillo, J. (2012). Emotive hemispheric differences measured in real-life portraits using pupil diameter and subjective aesthetic preferences. *Experimental Brain Research*, *219*(4),

447–455. Retrieved from https://doi.org/10.1007/s00221-012-3091-y

Bohrn, I. C., Altmann, U., Lubrich, O., Menninghaus, W., & Jacobs, A. M. (2012). Old proverbs in new skins: An fMRI study on defamiliarization. *Frontiers in Psychology*, *3*:204. Retrieved from https://doi.org/10.3389/fpsyg.2012.00204

Bohrn, I. C., Altmann, U., Lubrich, O., Menninghaus, W., & Jacobs, A. M. (2013). When we like what we know: A parametric fMRI analysis of beauty and familiarity. *Brain and Language*, *124*(1), 1–8. Retrieved from https://doi.org/10.1016/j.bandl.2012.10.003

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin*, *106*, 265–289.

Bouma, H., & De Voogd, A. H. (1974). On the control of eye saccades in reading. *Vision Research*, *14*(4), 273–284.

Clifton, C., Ferreira, F., Henderson, J. M., Inhoff, A. W., Liversedge, S. P., Reichle, E. D., & Schotter, E. R. (2016). Eye movements in reading and information processing: Keith Rayner's 40 year legacy. *Journal of Memory and Language*, *86*, 1–19. Retrieved from https://doi.org/10.1016/j.jml.2015.07.004

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(7), 1–9.

Dabbs, J. M. (1997). Testosterone and pupillary response to auditory sexual stimuli. *Physiology and Behavior*, *62*(4), 909–912. Retrieved from https://doi.org/10.1016/S0031-9384(97)00268-0

de Gee, J. W., Knapen, T., & Donner, T. H. (2014). Decision-related pupil dilation reflects upcoming choice and individual bias. *Proceedings of the National Academy of Sciences, USA*, *111*(5), E618–E625. Retrieved from https://doi.org/10.1073/pnas.1317557111

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655. Retrieved from https://doi.org/10.1016/S0022-5371(81)90220-6

Fabb, N. (2015). *What is poetry?: Language and memory in the poems of the world*. Cambridge, UK: Cambridge University Press.

Fernández, G., Shalom, D. E., Kliegl, R., & Sigman, M. (2014). Eye movements during reading proverbs and regular sentences: The incoming word predictability effect. *Language, Cognition and Neuroscience*, *29*(3), 260–273.

Fitzgerald, C. M. (2007). An optimality treatment of syntactic inversions in English verse. *Language Sciences*, *29*(2–3), 203–217. Retrieved from https://doi.org/10.1016/j.langsci.2006.12.020

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Gagl, B., Hawelka, S., & Hutzler, F. (2011). Systematic influence of gaze position on pupil size measurement: Analysis and correction. *Behavior Research Methods*, *43*(4), 1171–1181. Retrieved from https://doi.org/10.3758/s13428-011-0109-5

Giora, R., Fein, O., Kronrod, A., Elnatan, I., Shuval, N., & Zur, A. (2004). Weapons of mass distraction: Optimal innovation and pleasure ratings. *Metaphor and Symbol*, *19*(2), 115–141. Retrieved from https://doi.org/10.1207/s15327868ms1902_2

Granholm, E., & Steinhauer, S. R. (2004). Pupillometric measures of cognitive and emotional processes. *International Journal of Psychophysiology*, *52*(1), 1–6. Retrieved from https://doi.org/10.1016/j.ijpsycho.2003.12.001

Hayes, T. R., & Petrov, A. A. (2016). Mapping and correcting the influence of gaze position on pupil size measurements. *Behavior Research Methods*, *48*(2), 510–527. Retrieved from https://doi.org/10.3758/s13428-015-0588-x

Heister, J., Würzner, K. M., & Kliegl, R. (2012). Analysing large datasets of eye movements during reading. *Visual Word Recognition*, *2*, 102–130.

Hess, E. H., & Polt, J. M. (1960, August 5). Pupil size as related to interest value of visual stimuli. *Science*, *132*(3423), 349–350. Retrieved from https://doi.org/10.1126/science.132.3423.349

Hohenstein, S., Laubrock, J., & Kliegl, R. (2010). Semantic preview benefit in eye movements during reading: A parafoveal fast-priming study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1150–1170.

Huey, E. B. (1908). *The psychology and pedagogy of reading*. New York, NY: Macmillan.

Hyönä, J., Tommola, J., & Alaja, A.-M. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology, Section A*, *48*(3), 598–612. Retrieved from https://doi.org/10.1080/14640749508401407

Jacobs, A. M. (2015). Neurocognitive poetics: Methods and models for investigating the neuronal and cognitive–affective bases of literature reception. *Frontiers in Human Neuroscience*, *9*:186. Retrieved from https://doi.org/10.3389/fnhum.2015.00186

Jacobsen, T., Buchta, K., Köhler, M., & Schröger, E. (2004). The primacy of beauty in judging the aesthetics of objects. *Psychological Reports*, *94*(3), 1253–1260.

Jakobson, R. (1960). Linguistics and poetics. In T. A. Sebeok (Ed.), *Style in language* (pp. 350–377). New York, NY: Wiley.

Johnson, M. G., Muday, J. A., & Schirillo, J. A. (2010). When viewing variations in paintings by Mondrian, aesthetic preferences correlate with pupil size. *Psychology of Aesthetics, Creativity, and the Arts*, *4*(3), 161–167. Retrieved from https://doi.org/10.1037/a0018155

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354. Retrieved from https://doi.org/10.1037/0033-295X.87.4.329

Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *47*(2), 310–339. Retrieved from https://doi.org/10.1037/h0078820

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. Retrieved from https://doi.org/10.1177/001316446002000116

Kant, I. (2001). *Critique of the power of judgment*. (P. Guyer & E. Matthews, Trans.) Cambridge, UK: Cambridge University Press.

Knoop, C. A., Wagner, V., Jacobsen, T., & Menninghaus, W. (2016). Mapping the aesthetic space of literature from "below." *Poetics*, *56*, 35–49. Retrieved from https://doi.org/10.1016/j.poetic.2016.02.001

Koffka, K. (1935). *Principles of gestalt psychology*. New York, NY: Harcourt, Brace.

Kojima, M., Shioiri, T., Hosoki, T., Kitamura, H., Bando, T., & Someya, T. (2004). Pupillary light reflex in panic disorder. *European Archives of Psychiatry and Clinical Neuroscience*, *254*(4), 242–244.

Kristjansson, S. D., Stern, J. A., Brown, T. B., & Rohrbaugh, J. W. (2009). Detecting phasic lapses in alertness using pupillometric measures. *Applied Ergonomics*, *40*(6), 978–986.

Kuchinke, L., Trapp, S., Jacobs, A. M., & Leder, H. (2009). Pupillary responses in art appreciation: Effects of aesthetic emotions. *Psychology of Aesthetics, Creativity, and the Arts*, *3*(3), 156–163. Retrieved from https://doi.org/10.1037/a0014464

Kuchinke, L., Võ, M. L. H., Hofmann, M., & Jacobs, A. M. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *International Journal of Psychophysiology*, *65*(2), 132–140. Retrieved from https://doi.org/10.1016/j.ijpsycho.2007.04.004

Laeng, B., Eidet, L. M., Sulutvedt, U., & Panksepp, J. (2016). Music chills: The eye pupil as a mirror to music's soul. *Consciousness and Cognition*, *44*, 161–178. Retrieved from https://doi.org/10.1016/j.concog.2016.07.009

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science 7*(1), 18–27. Retrieved from https://doi.org/10.1177/1745691611427305

Laeng, B., & Sulutvedt, U. (2014). The eye pupil adjusts to imaginary light. *Psychological Science*, *25*(1), 188–197. Retrieved from https://doi.org/10.1177/0956797613503556

Leech, G. N. (1969). *A linguistic guide to English poetry*. Harlow, UK: Longman.

Levin, S. R. (1962). *Linguistic structures in poetry*. The Hague, Netherlands: Mouton.

Loewenfeld, I. (1999). *The pupil: Anatomy, physiology, and clinical applications*. Oxford, UK: Butterworth-Heinemann.

Martindale, C., & Moore, K. (1988). Priming, prototypicality, and preference. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 661–670.

McGlone, M. S., & Tofighbakhsh, J. (1999). The Keats heuristic: Rhyme as reason in aphorism interpretation. *Poetics*, *26*(4), 235–244. Retrieved from http://doi.org/10.1016/s0304-422x(99)00003-0

McGlone, M. S., & Tofighbakhsh, J. (2000). Birds of a feather flock conjointly: Rhyme as reason in aphorisms. *Psychological Science*, *11*(5), 424–428. Retrieved from http://doi.org/10.1111/1467-9280.00282

Menninghaus, W., Bohrn, I. C., Altmann, U., Lubrich, O., & Jacobs, A. M. (2014). Sounds funny? Humor effects of phonological and prosodic figures of speech. *Psychology of Aesthetics, Creativity, and the Arts*, *8*(1), 71–76. Retrieved from https://doi.org/10.1037/a0035309

Menninghaus, W., Bohrn, I. C., Knoop, C. A., Kotz, S.

A., Schlotz, W., & Jacobs, A. M. (2015). Rhetorical features facilitate prosodic processing while handicapping ease of semantic comprehension. *Cognition*, *143*, 48–60. Retrieved from https://doi.org/10.1016/j.cognition.2015.05.026

Menninghaus, W., Wagner, V., Hanich, J., Wassiliwizky, E., Jacobsen, T., & Koelsch, S. (2017). The distancing-embracing model of the enjoyment of negative emotions in art reception. *Behavioral and Brain Sciences*, *40*:e347. Retrieved from https://doi.org/10.1017/S0140525X17000309

Miall, D. S., & Kuiken, D. (1994). Foregrounding, defamiliarization, and affect: Response to literary stories. *Poetics*, *22*(5), 389–407. Retrieved from https://doi.org/10.1016/0304-422X(94)00011-5

Miall, D. S., & Kuiken, D. (1998). The form of reading: Empirical studies of literariness. *Poetics*, *25*(6), 327–341. Retrieved from https://doi.org/10.1016/S0304-422X(98)00002-3

Nietzsche, F. W. (2000). *The birth of tragedy* (D. Smith, Trans.) Oxford, UK: Oxford University Press.

Obermeier, C., Kotz, S. A., Jessen, S., Raettig, T., von Koppenfels, M., & Menninghaus, W. (2016). Aesthetic appreciation of poetry correlates with ease of processing in event-related potentials. *Cognitive, Affective, & Behavioral Neuroscience*, *16*(2), 362–373. Retrieved from https://doi.org/10.3758/s13415-015-0396-x

Obermeier, C., Menninghaus, W., Koppenfels M. von., Raettig, T., Schmidt-Kassow, M., Otterbein, S., & Kotz, S. A. (2013). Aesthetic and emotional effects of meter and rhyme in poetry. *Frontiers in Psychology*, *4*:10. Retrieved from https://doi.org/10.3389/fpsyg.2013.00010

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105–110.

Pollatsek, A., & Rayner, K. (1990). Eye movements and lexical access in reading. In D. A. Balota, G. B. Flores d'Arcais, & K. Rayner (Eds.), *Comprehension Processes in Reading* (pp. 143–163). New York: Routledge.

Powell, W. R., & Schirillo, J. A. (2011). Hemispheric laterality measured in Rembrandt's portraits using pupil diameter and aesthetic verbal judgements. *Cognition & Emotion*, *25*(5), 868–885. Retrieved from https://doi.org/10.1080/02699931.2010.515709

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Rayner, K. (1998). Eye movements in reading and information processing: 20 Years of Research. *Psychological Bulletin*, *124*(3), 372–422.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, *8*(3), 338–342.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, *8*(4), 364–382. Retrieved from https://doi.org/10.1207/s15327957pspr0804_3

Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science*, *9*(1), 45–48. Retrieved from https://doi.org/10.1111/1467-9280.00008

Rice, C. (1997). Ranking components: The grammar of poetry. In G. Booij & J. van de Weijer (Eds.), *Phonology in progress: Progress in phonology* (pp. 321–332). The Hague, Netherlands: Holland Academic Graphics.

Rieger, G., & Savin-Williams, R. C. (2012). The eyes have it: Sex and sexual orientation differences in pupil dilation patterns. *PLoS One*, *7*(8):e40256. Retrieved from https://doi.org/10.1371/journal.pone.0040256

Schindler, I., Hosoya, G., Menninghaus, W., Beermann, U., Wagner, V., Eid, M., & Scherer, K. R. (2017). Measuring aesthetic emotions: A review of the literature and a new assessment tool. *PLoS One*, *12*(6):e0178899. Retrieved from https://doi.org/10.1371/journal. pone.0178899

Schluroff, M. (1982). Pupil responses to grammatical complexity of sentences. *Brain and Language*, *17*(1), 133–145. Retrieved from https://doi.org/10.1016/0093-934X(82)90010-4

Schluroff, M., Zimmermann, T. E., Freeman, R. B., Hofmeister, K., Lorscheid, T., & Weber, A. (1986). Pupillary responses to syntactic ambiguity of sentences. *Brain and Language*, *27*(2), 322–344. Retrieved from https://doi.org/10.1016/0093-934X(86)90023-4

Schotter, E. R., Angele, B., & Rayner, K. (2012). Parafoveal processing in reading. *Attention, Perception, & Psychophysics*, *74*(1), 5–35. Retrieved from https://doi.org/10.3758/s13414-011-0219-2

Shebilske, W. (1975). Reading eye movements from an information-processing point of view. In D. W. Massaro (Ed.), *Understanding language* (pp. 291–311). London: Academic Press.

Shebilske, W. L., & Fisher, D. F. (1983). Eye movements and context effects during reading of

extended discourse. In K. Rayner (Ed.), *Eye Movements in Reading: Percetual and Language Processes* (pp. 153–179). New York: Academic Press.

Shklovsky, V. (1917). Art as technique. In L. T. Lemon & M. J. Reis (Eds. and Trans.), *Russian formalist criticism: Four essays* (pp. 3–24). Lincoln, NE: University of Nebraska Press.

Shreve, G. M., Lacruz, I., & Angelone, E. (2010). Cognitive effort, syntactic disruption, and visual interference in a sight translation task. In G. M. Shreve & E. Angelone (Eds.), *Translation and cognition* (pp. 63–84). Amsterdam, Netgerlands: John Benjamins Publishing.

Silvia, P. J. (2007). Knowledge-based assessment of expertise in the arts: Exploring aesthetic fluency: Correction to Silvia (2007). *Psychology of Aesthetics, Creativity, and the Arts*, *1*(4), 247–249. Retrieved from https://doi.org/10.1037/1931-3896.2.1.33

Thierry, G., Martin, C. D., Gonzalez-Diaz, V., Rezaie, R., Roberts, N., Davis, Ph. M. (2008). Event-related potential characterisation of the Shakespearean functional shift in narrative sentence structure. *NeuroImage*, *40*, 923–931.

Tillmann, B., & Dowling, W. J. (2007). Memory decreases for prose, but not for poetry. *Memory & Cognition*, *35*(4), 628–639. Retrieved from https://doi.org/10.3758/BF03193301

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt II. *Psycologische Forschung*, *4*, 301–350.

Winkielman, P., & Cacioppo, J. T. (2001). Mind at ease puts a smile on the face: Psychophysiological evidence that processing facilitation elicits positive affect. *Journal of Personality and Social Psychology*, *81*(6), 989–1000. Retrieved from https://doi.org/10.1037/0022-3514.81.6.989

Winkielman, P., Schwarz, N., Fazendeiro, T. A., & Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 189–217). Mahwah, NJ: Lawrence Erlbaum.

Youmans, G. (1983). Generative tests for generative meter. *Language*, *59*(1), 67–92. Retrieved from https://doi.org/10.2307/414061

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*(2), 1–27.