RESEARCH ARTICLE

REVISED In poetry, if meter has to help memory, it takes its time [version 2; peer review: 1 approved, 1 approved with reservations, 2 not approved]

Sara Andreetta[1], Oleksandra Soldatkina[1], Vezha Boboeva[1,2], Alessandro Treves[1]

[1]Cognitive Neuroscience, SISSA, Trieste, 34136, Italy
[2]Bioengineering, Imperial College London, London, SW7 2AZ, UK

v2 **First published:** 28 May 2021, **1**:59
https://doi.org/10.12688/openreseurope.13663.1
Latest published: 23 Feb 2023, **1**:59
https://doi.org/10.12688/openreseurope.13663.2

Abstract

To test the idea that poetic meter emerged as a cognitive schema to aid verbal memory, we focused on classical Italian poetry and on three components of meter: rhyme, accent, and verse length. Meaningless poems were generated by introducing prosody-invariant non-words into passages from Dante's Divina Commedia and Ariosto's Orlando Furioso. We then ablated rhymes, modified accent patterns, or altered the number of syllables. The resulting versions of each non-poem were presented to Italian native speakers, who were then asked to retrieve three target non-words. Surprisingly, we found that the integrity of Dante's meter has no significant effect on memory performance. With Ariosto, instead, removing each component downgrades memory proportionally to its contribution to perceived metric plausibility. Counterintuitively, the fully metric versions required longer reaction times, implying that activating metric schemata involves a cognitive cost. Within schema theories, this finding provides evidence for high-level interactions between procedural and episodic memory.

Keywords

Schema theory, Memory retrieval, Hendecasyllables, Sequence replay, Dynamical attractors.

This article is included in the Languages and Literature gateway.

Open Peer Review

Approval Status ✓ ✗ ? ✗

	1	2	3	4
version 2 (revision) 23 Feb 2023				
version 1 28 May 2021	✓ view	✗ view	? view	✗ view

1. **Lev Blumenfeld**, Carleton University, Ottawa, Canada

2. **Stefan Blohm**, Radboud University, Nijmegen, The Netherlands
Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany
Stefano Versace, University of Birmingham, Birmingham, UK

3. **Johann-Mattis List**, Max Planck Institute for the Science of Human History, Jena, Germany

4. **Mara Breen**, Mount Holyoke College, South Hadley, USA

Any reports and responses or comments on the

article can be found at the end of the article.

This article is included in the Marie-Sklodowska-Curie Actions (MSCA) gateway.

This article is included in the The Mind, Mental Health, and Behaviour collection.

Corresponding author: Alessandro Treves (ale@sissa.it)

Author roles: Andreetta S: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Writing – Original Draft Preparation; **Soldatkina O**: Formal Analysis, Resources, Software, Writing – Review & Editing; **Boboeva V**: Software, Visualization, Writing – Review & Editing; **Treves A**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No [765549], (Memory research: Ground-breaking, Applied, and Technological Exchanges [M-GATE]) and by the Human Frontier Science Program grant RGP0057/2016.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Andreetta S, Soldatkina O, Boboeva V and Treves A. **In poetry, if meter has to help memory, it takes its time [version 2; peer review: 1 approved, 1 approved with reservations, 2 not approved]** Open Research Europe 2023, **1**:59 https://doi.org/10.12688/openreseurope.13663.2

First published: 28 May 2021, **1**:59 https://doi.org/10.12688/openreseurope.13663.1

REVISED **Amendments from Version 1**

We have extensively revised the text to respond to the numerous criticism, requests for clarifications and suggestions by all four reviewers.

In particular, we have

- Clarified that we use the word meter in a broad sense, to include rhyme.
- Emphasized the delicate nature of our manipulations, which made them all largely acceptable to our participants' ears. This is shown in a new figure in the Extended Data, which has the plausibility of all versions fall within the variability of a shuffled distribution, with the marginal exception of the NPR versions of Ariosto's passages. Likewise, the memory performance obtained with all versions is similar, and largely within the variability of the shuffled distribution – what stands out is the correlation between the two measures, already reported in Figure 4, which is highly significant for Ariosto and null for Dante.
- Thoroughly clarified the procedure, which allowed for a meaningful interaction with participants, necessary to fully engage them in a task that taps on their cultural enjoyment, while enabling access to a relatively large number of them during the pandemics, in sufficiently standardized conditions.
- Cases of inconsistent terminology have been rectified and references to the relant literature have been added or updated.

Any further responses from the reviewers can be found at the end of the article

Plain language summary

We have tested the common idea that poetic meter has emerged in order to help verbal memory. We have focused on classical Italian poetry and on three components of meter: rhyme, accent and verse length. We selected four passages from Dante's Divina Commedia and four from Ariosto's Orlando Furioso, and from them derived meaningless poems by replacing key words with non-words in such a way as to retain the original prosody. Alongside these eight "original" non-poems, we created three metrically defective variants of each, by either ablating rhymes or modifying accent patterns or altering the number of syllables. The resulting four versions of each non-poem were presented repeatedly to Italian native speakers, who were then asked, a day later, to remember some of the non-words. Surprisingly, we found that the integrity of Dante's meter has no significant effect on memory performance. With Ariosto, instead, removing each component downgrades memory proportionally to its contribution to perceived metric plausibility, with rhymes the most important one. Counterintuitively, the fully metric versions required longer reaction times, implying that activating a given metric pattern, an example of a recurring schema, is helpful but involves a cognitive cost. Within schema theories, this finding provides evidence for high-level interactions between procedural and episodic memory.

Introduction

Poems, nursery rhymes, traditional songs: they are found in every culture, and they have been around for ages, well before the advent of writing systems. Sometimes, they have or had the crucial mission of carrying an important message for the listeners: a list to know by heart, an event happening every year, a warning of a potential danger. What do these texts have in common? At least one aspect: they adopt a variety of devices that help hold verbal material in memory.

Human memory can, in fact, fail spectacularly at times. Writing systems have helped safely store verbal information, in a format relatively difficult to tamper with; before, when our ancestors had to rely on their fallible memory, a number of linguistic devices crystallized to help them remember words and verbal material. Cultural transmission, then, has depended for ages on these devices, which in poetry we can broadly refer to as "meter". These devices may range from the use of repeated metaphors: "rosy-fingered dawn" in Homer (Reece, 2011), to the *ring composition* in the Zoroastrian Yasna (Hintze, 2002) to semantic repetition as in Biblical poetry: "In the way of righteousness is life; and in the pathway thereof there is no death.**"** in Prov. 12:28 (King James Version).

In several Western literary traditions, including the Italian one, the local structure of poetry revolves around the verse, and includes a constant number of syllables, a limited variability in the pattern of accents, and a specific organization of rhymes. These components of meter (again, intended in a very broad sense) have gradually lost their centrality or at least their perceived necessity over the course of several centuries, but they were in full sway at least between the 13th and 16th centuries, from the emergence of modern Italian (so called "volgare", the language of the people) as an acceptable literary language to the diffusion of the printing press. The *Divina Commedia* by Dante Alighieri and the *Orlando Furioso* by Ludovico Ariosto are two lengthy masterpieces towards the beginning and, respectively, the end of this golden age. With 14,233 verses in the *Commedia* and 38,736 in the *Orlando Furioso*, neither of which contains material which is absolutely necessary to remember in order to carry on with one's life, it may be asked whether their metric structure still retained a primary memory function, or is already a purely esthetic ornament for cultured readers (Rubin, 1995).

Can the role of metrical features be explained from a neurocognitive point of view, with respect to memory? Recently, in memory literature the notion of schemata, long seen as important (see e.g., (Brewer & Dupree, 1983), has been discussed again (Ghosh & Gilboa, 2014), stimulated also by the analysis of its neurobiological basis in rodents (Tse *et al.*, 2007). A schema, whether directly functional like those involved in preparing coffee (Norman & Shallice, 1986) or social/ornamental, like rituals of salutations (Taylor & Crocker, 1981), can be considered as a set of regularities that help organize and retrieve information (Van Kesteren *et al.*, 2012). Its neural basis would be the memory trace of those regularities, that helps funnel neural activity so as to reinstate them. In language processing,

"narrative schemata" have often been described, as the features of stories which make them easier to remember, sometimes called "story grammars" (Rumelhart, 1975). In the present context, however we are considering metrical schemata: patterns which, by somewhat restricting options and encouraging expectations, facilitate the recall of verses. Thus, meter (broadly defined) should help us recruit, and possibly produce, the next element of a sequence stored in our memory.

In facilitating verbal sequence replay, metrical features appear to be effective with extended "trajectories", lasting even several verses. These are extended relative to the short trajectories thought to be produced by the phonological loop of Baddeley's model of working memory, which are presumed to last only a couple of seconds, precisely because of the lack of specific devices that extend their range (Baddeley, 1992; Haluts *et al.*, 2020). To the best of our knowledge, though, the effectiveness of these features has never been quantified. In this study, we aim at measuring the strength of some metric devices. Specifically, we focus on the three main characteristics of classical Italian meter: rhyme, pattern of accents, and verse length.

Methods

We extracted passages from two masterpieces of Italian literature: the *Divina Commedia* by Dante Alighieri (1265–1321), and the *Orlando Furioso* by Ludovico Ariosto (1474–1533). From the latter we chose *ottave* (octaves, stanzas of eight verses) from *canti* XIII, XV, XIX and XXX, and one from *canto* I to train subjects, while from the former we selected sequences of three consecutive *terzine* (hence nine verses) from two *canti* from *Inferno* (XXIV for the experiment, V for training), two from *Purgatorio* (VI and XVI) and one from *Paradiso* (XXVII). All passages had only proper (Italian) hendecasyllables with an accented 10th syllable, and were, to our arbitrary judgement, devoid of explicit or easily reconstructed memorable content or references.

Poem manipulation

The original texts were manipulated in a number of different ways. Firstly, most content words were converted into non-words in order to eliminate discernible semantic content, hence semantic effects on memory; an effort was made to minimize the impact of this manipulation (like for the subsequent ones below) by changing phonemes with similar ones, while maintaining the original prosody. Function words were not modified. This applied equally to all passages and resulted in "original non-poems" (ONPs).

The second stage of manipulations focused on metrical patterns. We created three conditions:

1) a condition where we eliminated rhymes ("NPR" – Non-Poem without Rhyme)
2) a condition where the accent patterns of four-five verses per passage were replaced with less standard ones ("NPA" – Non-Poem with modified Accents). This manipulation was particularly non-trivial, since accent patterns are not rigidly defined. However, to validate proper original accents, we consulted with an expert scholar for *Orlando Furioso*, whereas for *Divina Commedia* we referred to the "Archivio Metrico Italiano", a database collecting masterpieces of Italian literature with their accents annotated (www.maldura.unipd.it). On these bases, we altered the "original accents" by putting them in different positions within the verse, taking advantage in particular of non-words with no mandatory accent.
3) a condition where the number of syllables per verse, which in the ONPs were strictly 11 throughout (regular hendecasyllables) was altered again in four-five verses, to nine, 10, 12 or 13 ("NPS" – Non-Poem with wrong numbers of Syllables). Note that by adding or subtracting one or two syllables, also the pattern of accents was perforce altered, but we attempted to make the alteration less noticeable than the number change, in contrast to the NPA condition in which, while there were strictly 11 syllables/verse, the accents followed more unusual patterns.

These manipulations were applied to all passages. All texts were then recited by a professional actor and audio recorded.

For the experiment, every subject was administered four texts in total, by the same (original) author: one per passage and one per condition (a Latin square design). Therefore, twenty-four combinations were created.

An example of an NPS we used, from the *Commedia*, is presented in Figure 1 together with its original spectrogram, and all non-poems can be inspected in the Extended data, with a descriptive README file detailing how they were used (Andreetta *et al.*, 2021).

Ranking

We conducted an online survey about how these manipulated poems were perceived by a group of Italian native speakers. Participants were asked to listen to the four conditions and give a ranking of preference, from the one that sounded the most "poetically plausible" to them, to the one they perceived as the strangest.

Consent statement. Written informed consent for participation was obtained in advance from all participants.

Subjects. 62 people participated in the online survey for Ariosto (F=32, M=30, mean age = 29.06, sd= 8.13) and 65 people for Dante (F=35, M=30, mean age = 26.48, sd = 6.26). Part of either cohort were the participants in the main experiment below, but tested with the other author, and they were asked to complete this survey after the end of the second session of the main experiment. Another group of participants was recruited through the online platform Prolific (www.prolific.co). This last group was compensated with five euros. We had

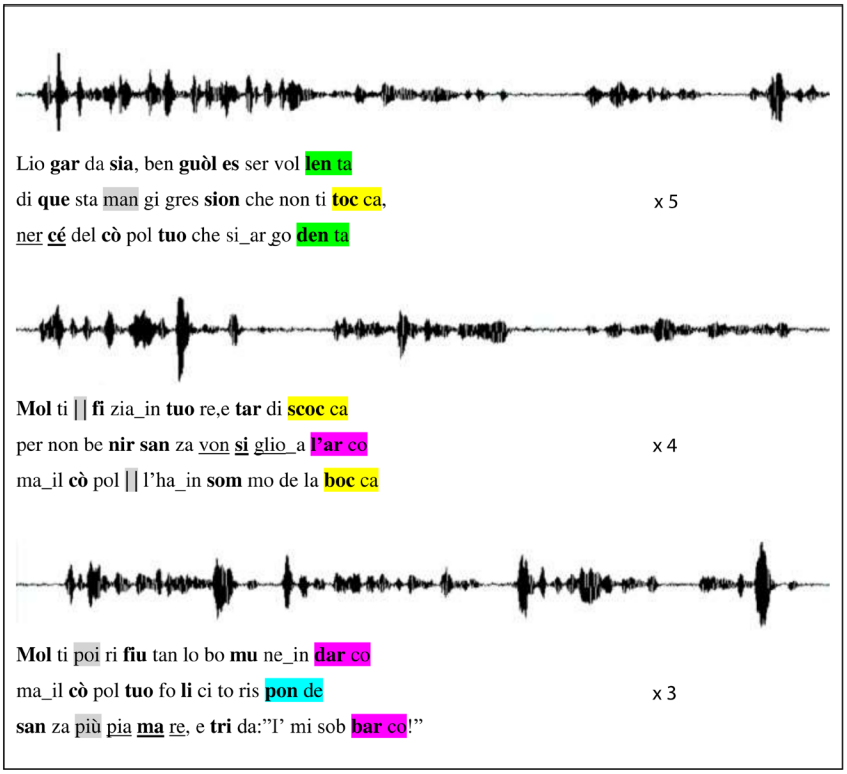

Figure 1. NPS example derived from Purgatorio, canto VI. For each *terzina* (vv.127–135) the NPS text, shown below the sound wave by the professional actor, maintains rhymes, in color, and accents, in boldface, as in the ONP version; whereas overall three syllables have been added and two taken away, in gray. The underlined non-words were the targets of the memory test; underlined blanks denote *synizesis* (when two syllables are pronounced as one).

aimed for 72 rankers in each cohort, to have 3 for any of the 24 passage-condition combinations, but had to exclude some *a posteriori*, who failed to complete the ranking in full. To maintain the balance in the averages, if a combination ended up with only 2 rankers, we gave them weight 1.5 (also in the shuffled distributions, see the Extended Data).

Experimental design. The online survey was designed with the open-source toolkit Psytoolkit (Stoet, 2017). After an example, presented as training also in the main experiment below, they listened to the four poems one at a time. Every poem was associated with a name, in order to help participants refer to that specific condition. If they wanted, they were allowed to listen again and again to the same poem before proceeding to the next.

At the end they were asked to rank the four poems: from the one they perceived as the best, to the one that sounded worse to them. From the rankings by all participating subjects we extracted an average index of metric plausibility by assigning a value 0.6 to the first-ranked condition (e.g., NPA), 0.3 to the second, 0.1 to the third, and 0 to the fourth. The logic behind this assignment is that subjects occasionally reported being unsure as to which passage sounded the strangest. The rankings were collapsed across passages, with the relatively large number of participants ensuring approximately even sampling (each passage was presented originally 18 times per condition, which came down to 16+/-2 after the exclusions). As a result, the average metric plausibility of each condition could in principle range from 0 to 0.6, but in practice was much more restricted, particularly with passages from Dante, to values around the average of 0.25 (see Figure 2, and the Extended Data).

Memory experiment

The main experiment testing the effect of the manipulations utilized two groups of 24 participants each, who were later included in the cohorts for the ranking (of the texts from the other author). Ethical approval for this study was granted by the SISSA Ethics Committee at the Scuola Internazionale Superiore di Studi Avanzati with deliberation 2018/16/ib on Nov. 5th, 2018 transmitted by act prot. 15534-III/13.

Subjects. 48 native Italian speakers who had been exposed to Italian literature through one of the national high school curricula were recruited through the SISSA recruiting platform and social media. Half of them were administered material from Ariosto (F= 15, M=9, mean age= 26.34, sd=4.02), the other half from Dante (F=15, M=9, mean age= 26.12, sd=3.61). None of them had a previous history of psychiatric or neurological illness, learning disabilities, nor hearing or visual loss. They were asked to participate in a study on memory and

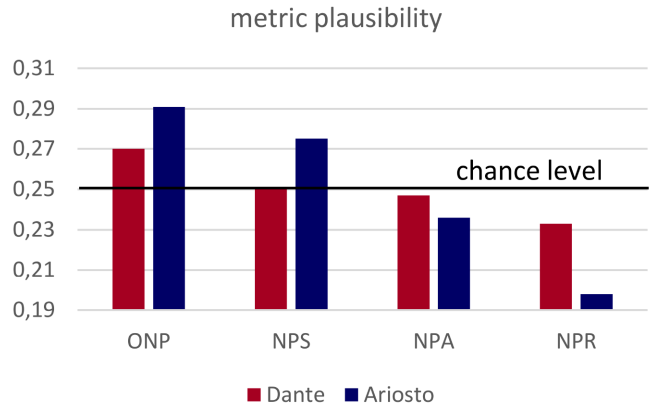

Figure 2. Relative metric plausibility. The different versions of the same four passages from the *Divina Commedia* (red) and *Orlando Furioso* (blue) were ranked in the same order, but the plausibility index (see Methods) is more spread out for Ariosto.

poetry, which would have involved them for two consecutive days, for about 30 minutes the first and about 10 minutes the second. Due to the pandemic situation, they were asked to be connected remotely with their own devices.

Experimental design. The experiment was designed with Psychopy Builder (Peirce *et al.*, 2019). It included a study phase of about 30 minutes the first day, and the test of about 10 minutes the second day.

We aimed at an almost exclusively auditory experiment, in order to assess how memory relies on meter if listening is the only available channel to learn from (Najme *et al.*, 2020). Indeed, the material included audio files only, with the sole exception of written material when a fill in the gap task appeared.

Besides the four passages, verses from two other *canti* were used for training, as indicated above. However, these verses were presented only in the ONP condition, leaving meter intact.

Every passage, including the training, was associated with an image, taken from among Gustave Doré's illustrations of the *Divina Commedia*. The images were the same for each passage in the different conditions and were intended to help engage memory without, at the same time, biasing the linguistic material (see Extended data).

Every poem, including the training, was presented divided into three consecutive portions.

Notice on the use of the Zoom platform. The pandemic of 2020 forced us to find new methods to administer our experiment to subjects in remote mode.

After evaluating several options, we decided that, for this study, it was important to keep a degree of dialogue with participants. Also, we wanted to make sure that they were focused on the task and that they did the second part at the same time the following day.

For these reasons, among others, we thought that a good option was to have them on streaming in an open source platform. We chose Zoom, for which we had an institutional account.

Unfortunately, this meant that lab conditions could not be fully guaranteed. To overcome potential biases, we gave participants specific instructions:

- be connected with a computer or a tablet. Smartphones were not allowed, because of the small size of their screens and because they could potentially be distracting in case of notifications during the experiment
- be in a silent room with no disturbances
- be in the same room during the experiment for both days

Moreover, the accompanying images also helped focus the visual attention of the participants away from distracting visual stimuli in the rooms they were in. The images, it should be noted, accompanied passages derived from *canti* in the *Commedia* unrelated to those Gustave Doré referred to, as well as the *a priori* unrelated passages derived from the *Orlando Furioso*, and they were further enriched with graded color hues. Each image was consistently paired to each passage, whichever version, ONP, NPS, NPA or NPR, was presented to the subject. The variability across images, enhanced also by leaving each image in its original non-commensurate format, was thus adding an independent component to the natural variability across

passages, but not contributing to the metric plausibility effect; and possibly helped reduce the variability, across participants, due to their heterogeneous testing-at-home conditions.

The dialogue over zoom was always conducted by the first author, but the experiment was self-paced by the participants using the Psychopy platform, with the images displayed on the shared screen and the passages played auditorily from the prior recordings by actor Sara Alzetta. The recordings are available upon request. When tested with the muted non-word (see below), participants would read aloud the left, central or right alternative appearing on the screen, and the experimenter would press the corresponding key (leftward, downward or rightward) on her own keyboard.

Study phase. The study phase started with the training ONP. First, participants listened once to all verses. Then they listened to the first part (three verses) repeated five times, with a three second pause between repetitions. Afterwards, another repetition of the same verses followed, but this time a non-word was muted. Muted non-words were usually positioned in the third verse. The task for the participant was then to retrieve the correct non-word. Three written alternatives appeared on the screen and the participant had to read aloud his/her choice.

After this training, each of the experimental passages was played, in a separate block, in the one of the four versions that had been assigned to that subject in the design. Then, after listening once to the entire poem (nine verses in Dante, eight verses in Ariosto), participants listened to the first part (three verses for both authors) five times. Next, they had to complete the task with the muted non-words.

The same happened with the second part of the poem, repeated four times. In this case the audio started from silence with the first part ramping up linearly in intensity, until it continued smoothly into the second part at the standard volume. Therefore, this allowed them to have feedback about the test they just completed.

For the third part they just listened to the repetitions (three verses for Dante, two verses for Ariosto; repeated three times), starting now with an acoustically smoothed version of the second part, but there was no test.

The alternatives to the correct non-word were chosen by maintaining the same number of syllables, and the same accent. Typically, stems and intermediate vowels or consonants changed. Again, target non-words were generally in the third verse, aiming not to overload working memory from the moment they listened to the silent word until the test time. In a few cases where this was not possible (e.g., because there were no appropriate non-words in the third verse) a non-word in the second verse was chosen, towards the end. Notably, for every passage we chose options which were consistent across all conditions, allowing a fair comparison in the results.

Test phase. The following day participants were connected at the same time, so 24 hours had passed. They listened directly to the test parts of the text, so the verses with a muted non-word, and they were asked to identify the muted target in all three parts, including then the third one. For the first and second part, tested also the previous day, target non-words were different. After completing the three tests per passage, they listened to the entire poem, to receive feedback.

Data analysis. The outcome of interest is essentially the presence of a significant correlation between the ranking produced by our intentionally delicate manipulations (expressed by the plausibility index, see Figure 2 and Extended Data) and the memory score in the main experiment (as well as the reaction times, see Figure 4), that would indicate a joint dependence on the type of manipulation. Correlations were considered significant at $p < 0.05$. Additional methodological considerations and controls are detailed below.

Entropy in the accent distribution. Two simple entropy measures were used to quantify the variability in the pattern of accents in the eight passages from the *Divina Commedia* and *Orlando Furioso* from which we derived the non-poems used in the experiment. First, the pattern of accents in each verse (from a total of 36 verses from the *Commedia* and 32 from the *Orlando*) was codified, based on the consensus in the literature, as a binary string of length 11, where each syllable was assigned a 1 if accented and a 0 if not. Since all 68 verses were regular hendecasyllables with the 10^{th} syllable accented and the 11^{th} not, we focused on the first nine digits in each string.

The first measure is based on the simplifying assumption of independent accents on neighboring syllables and calculates, for each author, the sum of the binary entropies for the syllables in each ordinal position, a sum which can range from 0 to nine bits.

The second measure is the entropy of the distribution, for each author, of distinct binary strings, and it ranges from 0 to $\log_2(36)$ for Dante and from 0 to 5 bits for Ariosto.

These two entropy measures appear no less sensitive, with our passages, than others recently proposed (see e.g. Sela & Gronas, 2022).

Word frequency. The targets were derived from words of widely different frequency, covering the entire spectrum from 5×10-8 to 1.5×10-3 in SUBTLEX-IT, the corpus of Italian Subtitle-based Word Frequency Estimates, containing 517.564 entries (poster presented at the Annual Meeting of the Italian Ass Expl Psychol, Rovereto, Sept. 2015: Crepaldi, Amenta, Mandera, Keuleers and Brysbaert, SUBTLEX-IT. Subtitle-based word frequency estimates for Italian. Available online: https://lrlac.sissa.it/publications/frequency-estimates-different-registers-explain-different-aspects-visual-word).

Results

The contribution of distinct components to metric plausibility. Two separate cohorts of rankers, for Dante- and Ariosto-derived non-poems, were presented with a combination of the four passages from the same author, one in each of the ONP, NPS, NPS and NPR versions, and were asked to rank them in order of metric plausibility. The fully balanced design allowed us to extract a passage-independent plausibility score.

Both when derived from passages by Dante and Ariosto, non-sense poems were found most plausible in their fully metric ONP versions, somewhat less when the number of syllables was manipulated (NPS), even less when the pattern of accents was altered in the NPA renditions, and the least when rhymes were removed, NPR. Remarkably, however, differences in the plausibility index are shown in Figure 2 to be quite limited (see also the comparison with shuffled responses in the Extended Data), confirming the soft impact of our manipulations and making the fully balanced design essential. The variance was particularly limited in passages from the *Commedia*, which may be due to Dante's taking more liberties with the meter he had adopted (the same hendecasyllables as Ariosto, but in *terzine* rather than *ottave*). To quantify this perception, at least in relation to accent patterns, which are more accessible to analysis, we applied two independent measures of accent variability to the four original passages by each author.

Dante appears to be slightly more variable in his accent patterns relative to Ariosto, but the main observation that can be gleaned off Figure 3 is that both poets are far from using a fixed pattern, utilizing over half of the maximum entropy they had available in terms of accenting those passages.

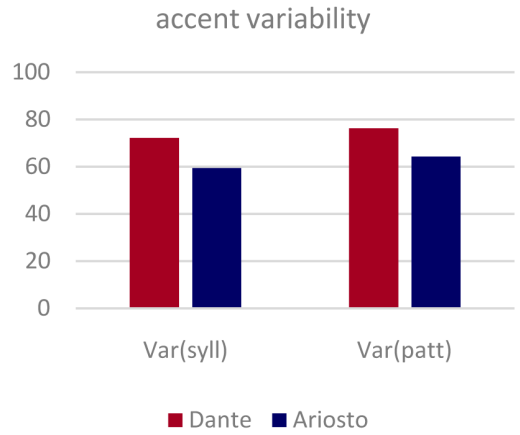

Figure 3. Variability in the pattern of accented syllables in the eight original passages by the two authors. Two independent entropy measures of variability, per syllable and per verse (see Methods) are both normalized to range from 0 (a single fixed pattern) to 100% (i.e., each syllable in the verse is accented half the time; or each verse follows a different accent pattern).

Meter can facilitate memory. Does such a loose structure help remember individual words? Table 1 and Figure 4 (upper) show that it does, only for the non-poems derived from Ariosto's octaves. Twenty-four subjects per author were asked, one day after repeatedly listening to one version of each passage,

Table 1. Correct responses out of a total of 24 participants for the first, second, and third query.

	Dante				Ariosto			
	NPR	NPA	NPS	ONP	NPR	NPA	NPS	ONP
1	5	15	13	10	11	16	19	13
2	13	13	9	10	9	10	9	15
3	15	17	16	12	11	8	12	14
All	33	45	38	32	31	34	40	42

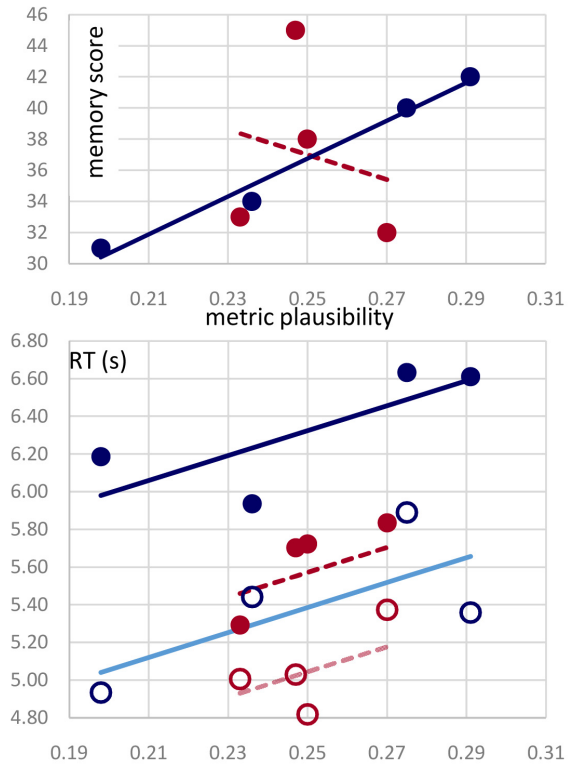

Figure 4. Memory and reaction times both increase with metric plausibility. (Upper) Overall correct responses (out of 72) for each condition, ordered in terms of their metric plausibility, as in Figure 2, for passages from Dante (red) and Ariosto (blue). (Lower) Reaction times (in seconds) for correct (circles) and wrong responses (dots) are regressed against plausibility for each author, with a single slope parameter. The slope is significant and similar to that characterizing the Ariosto data alone, whereas it is denoted with a dashed line for the Dante data, because the latter would not produce a significant correlation on its own.

to remember non-word targets out of three alternatives, upon listening to the non-poem with selected non-words muted. There were three such targets in each non-poem. While in the case of those derived from passages in the *Divina Commedia* the overall number of correct responses per condition was unrelated to its metric plausibility (r^2=0.04), seemingly fluctuating as much as the correct responses to the first, second, and third query taken alone (Table 1), for the passages from *Orlando Furioso* the correlation with metric plausibility was remarkable (r^2=0.98) and highly significant ($p<0.01$). Interestingly, the total score of the two cohorts was nearly identical, 147 for Ariosto and 148 for Dante, out of a total of 288 (24×4×3) and the memory scores per condition, with the exception of the NPA for Dante, were not significantly different from those obtained by randomly shuffling conditions across subjects (see Extended Data, Andreetta *et al.*, 2021).

Meter helps, but not for free. The analysis of reaction times helps interpret the above results. As shown in Figure 4 (lower), overall it took longer for participants to pick a wrong answer over the correct answer (on average, 733ms more), and it took longer for participants tested with Ariosto, relative to those tested with Dante, to respond (on average, 547ms more). Most importantly, in each of the four types of trials above, the more metrically plausible the passage, the longer the reaction time. However, the trend is significant only with Ariosto, if data from the two authors are analyzed separately, and it is significant overall ($p<0.004$) with a slope mainly determined by the Ariosto data, if analyzed together, as shown in Figure 4. The slope for the Dante data alone would be higher, but not significant, likely because of the limited plausibility range spanned.

The overall distribution of reaction times is reported in Figure 5. Note that to avoid biasing RT results with the occasional outliers, only RTs < 10s where included in the averages in Figure 4, leaving out 14 trials for Dante and 20 for Ariosto, each out of 288. Including them (or alternatively excluding the three trials with RT< 2.8s), does not change the results, in fact it widens the RT gap between Dante and Ariosto.

These findings suggest that processing meter in order to help retrieve a non-word heard the day before has a cognitive cost, and takes the order of hundreds of ms extra time, depending on exactly how much meter there is "used up" in the process. For passages derived from Dante, it appears that although outwardly the metric structure is essentially the same (with the slight qualification reported in Figure 3, and the note that a passage is a sequence of three *terzine* rather than a single *ottava*), meter is used less, and the very same memory performance is attained on average in less time.

Word frequency does not have major effects. While targets derived from more frequent words tended to be remembered marginally better, the same trend was observed for both authors (Figure 6), and each target appeared by design in all four conditions.

A strong bias makes subject favor the left alternative, among the three non-word options, but mainly in their wrong responses and the extent of the bias does not correlate with metric plausibility (Figure 7).

Discussion

The connection between meter and memory is not new to cognitive science: in a seminal book Rubin described oral traditions and the linguistic devices they use, highlighting in particular their role in memory as limiting the choice (one could say the entropy, (Shannon & Weaver, 1949) of larger units: by indicating a specific word ending, for instance, choices will be limited to those words which have the same ending, if a rhyme is expected (Rubin, 1995).

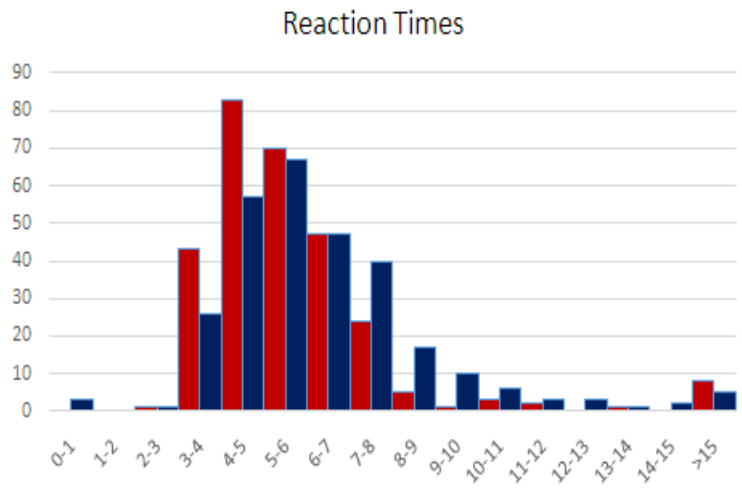

Figure 5. Distribution of reaction times. As explained in the results section, only RTs < 10s where included, leaving out 14 trials for Dante (red) and 20 for Ariosto (blue).

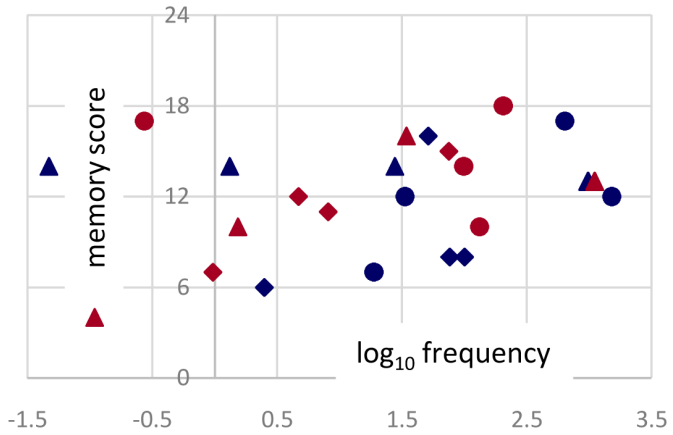

Figure 6. Frequencies. Word frequencies (log of occurrencies per million) in SUBTLEX-IT for the 24 (8×3) target words used in non-poems derived from Dante (red) and Ariosto (blue). On the y-axis the memory score is the number of times each target word has been correctly selected, by 24 participants.

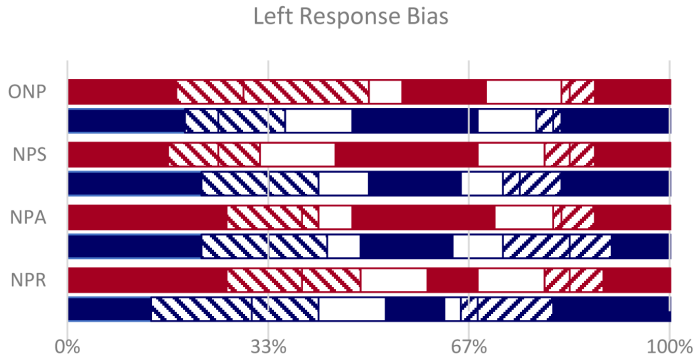

Figure 7. Left Bias. In full color (red for Dante, blue for Ariosto) the correct responses on the left, central and right non-words. Left alternative, among the three non-word options, was often chosen when the correct non-word was central or on the right (left-tilted striped segments). Blank segments are responses in the center, when the correct non-word was left or right. Right-tilted striped responses are on the right, when the correct non-word was left or center.

In music, Schulkind has investigated memory mechanisms by having participants listen to well-known and novel songs which were altered in their rhythm. Results showed that unaltered versions were identified significantly better than the altered ones, and this applied to both known and novel songs (Schulkind, 1999).

Analogously, Sachs investigated the retention of semantic and syntactic information in discourse by having participants listen to short prose stories. By selectively manipulating the meaning or the syntactic form of a target sentence, she could show that meaning is remembered, in prose, better and for longer that meaning-irrelevant sentence form (Sachs, 1967).

In a similar design, Tillman and colleagues have tested short term memory in prose and poetry. Also in this case, a sentence, considered the target, was changed in its form or in meaning. While with prose memory for surface characteristics declined over time, as expected, the same did not happen with poetry, for which form, in addition to content, was efficiently retrieved (Tillmann & Dowling, 2007).

With this study, we had hoped to be able to quantify, in rather absolute terms, the contribution of different aspects of meter to memory retention, using "material" from the classical period of Italian literature, before the advent of the printing press diminished the perceived value of memorability *per se*, and promoted the further ritualization of the written verse into a primarily esthetic construct. The results belied our naïve expectation, in that meter seems to be 'perceived' much more (in terms of our plausibility index) in passages derived from Ariosto than in those from Dante, and to contribute to memory in the former but not in the latter. Yet the meter employed by the two authors is nearly identical, with a discrete difference in the concatenation of hendecasyllables (in *terzine* in the *Divina Commedia*, in *ottave* in the *Orlando Furioso*) and a presumably small quantitative difference in the variability with which the common meter is used (Figure 3). Therefore, one would

expect that the listener, or the reader, activates the very same cognitive schema, at least locally, within the few verses of a single *ottava* or three *terzine*.

The time the subjects from the two statistically indistinguishable cohorts needed to react to the memory tests suggest an account of the main finding: the metric schema is the same, but it is activated to a different extent. Somewhat counterintuitively, it appears to be activated less with Dante, an author with whom most people who have been in high school in Italy are rather familiar, than with Ariosto, who has been relegated, especially in the last few decades, to a marginal niche in standard Italian curricula. This appears to discount a possible interpretation of this difference, i.e., that we are seeing two competing effects, whereby both congruence and incongruence with established schemata can enhance memory, the latter a novelty effect (Bonasia *et al.*, 2018): novelty presupposes the activation of the schema it contradicts. An alternative interpretation is that Dante's verses are just more interesting and tend to focus one's attention to other aspects than the components of meter. Even if this interpretation were to be shown to be correct, it is quite surprising that it would apply, in our paradigm, to verses that have been deprived of their meaning. Moreover, our replacing several of the key words chosen by Dante with our untalented choice of non-words would have been expected to remove other potentially useful devices from the poet's bag-of-tricks, like alliteration, onomatopoeia, use of liquids, of newly crafted words, etc. (Robey, 1985). Still, the wide-ranging contribution of sound 'shape' to cognitive processes has been noted, in particular in poetry (Blohm *et al.* 2021) and it might play a role in our findings, in forms not readily evident. In line with previous studies, we also acknowledge that a potential factor could also be the individual participants' expertise with poetry and/or music. Such data were not considered, here, and presumably individual differences average out, but these factors are being investigated in a related study.

Can the hypothesis of differential schema activation be tested experimentally? In principle, yes, and one approach would be by looking at evoked response potentials (ERPs), which have been widely used to reveal brain signals that reflect violations of expectation, whether (in the language domain) syntactic, semantic, or just phonological (Brown & Hagoort, 1993; Hagoort, 2003). With poetry, there have been ERP studies of aesthetic appreciation and ease of processing (Obermeier *et al.*, 2016) and of brain activity during poem composition (Liu *et al.*, 2015). For an experimental design like ours, however, one challenge is how to obtain the large number of trials per condition needed in order to obtain valid ERP measures. Another one is to what extent one can rely on single ERPs to characterize a heterogeneous variety of metric components. It is possible that addressing both challenges will require a change in perspective from whole brain dynamics to one which articulates the cortex into a plurality of interacting local networks, as embodied e.g., by the Potts model (Naim *et al.*, 2018). Distinct processes, among the many that concur to the overall perception, appreciation and memory of a poem, including the components of meter, are likely reflected differentially in the dynamics of distinct cortical networks, just like other, better studied types of memory such as episodic and spatial memories (Robin & Moscovitch, 2017), which have stimulated theories about the interactions between the medial temporal lobe and medial prefrontal cortex (Van Kesteren *et al.*, 2012). Meter, with its multiple components, indicates the need to go beyond the somewhat coarse distinctions available to neuropsychology, what is at the moment accessible only through mathematical models (Spalla *et al.*, 2021), with which one can study forms of partial coherence among multiple local networks, reminiscent of that of systems unable to attain long-range order (Stella *et al.*, 2020).

While partial coherence might seem to detract from the wholeness attributed to conscious processing (Dehaene *et al.*, 1998), it is entirely consistent with the idea of a mixture of automatic and controlled components concurring to memory encoding and retrieval (Wang & Morris, 2010). With meter, the notion that different schemata might be activated only optionally, at times, and then partially and incoherently with others, and when activated might offer only an incremental contribution, suggests a more nuanced take on high-level cognition in general. Using many filters to interpret reality, and to a variable degree, implies check-and-balances and minimal recourse to prevailing or dominant schemata, those that often reflect biases or prejudice.

Data availability

Underlying data

Repository: Meter and Memory. DOI: 10.17605/OSF.IO/A825X (Andreetta *et al.*, 2021).

The project contains the following underlying data:

- all Ariosto.xlsx (Responses to the non-poems derived from the *Orlando Furioso*).
- all Dante.xlsx (Responses to the non-poems derived from the *Divina Commedia*).

Extended data

Repository: Meter and Memory. DOI: 10.17605/OSF.IO/A825X (Andreetta *et al.*, 2021).

The project contains the following extended data:

- Andreetta_ExtendedData.pdf (the non-poems and the associated images)
- README file

Acknowledgements

An earlier version of this article can be found on bioRxiv (https://doi.org/10.1101/2021.03.14.435310)

We are grateful to the HFSP collaboration on Analog computations underlying language mechanisms, in particular to Yair Lakretz, with whom the original idea for this study was discussed, and to Elisa Ciaramelli, Rodolfo Zucco, Sergio Bozzola, Andrea Tabarroni and Sara Alzetta, who offered their different perspectives.

Author contributions

SA and AT conceived the experiment and generated the non-poems. SA and VB implemented it in Psychopy and Psytoolkit, and SA run it, assisted by OS with Prolific and with data analysis. SA and AT have written the report with inputs from VB and OS.

References

Andreetta S, Soldatkina O, Boboeva V, *et al.*: **Meter and memory.** 2021.
http://www.doi.org/10.17605/OSF.IO/A825X

Baddeley A: **Working Memory Alan Baddeley.** *Science.* 1992; **255**(5044): 556–559.
PubMed Abstract | **Publisher Full Text**

Blohm S, Kraxenberger M, Knoop CA, *et al.*: **Sound shape and sound effects of literary texts.** In: Kuiken, D. and Jacobs, A. M. (ed. by), *Handbook of Empirical Literary Studies*, De Gruyter, Berlin-Munich-Boston. 2021.
Publisher Full Text

Bonasia K, Sekeres MJ, Gilboa A, *et al.*: **Prior knowledge modulates the neural substrates of encoding and retrieving naturalistic events at short and long delays.** *Neurobiol Learn Mem.* 2018; **153**(Pt A): 26–39.
PubMed Abstract | **Publisher Full Text**

Brewer WF, Dupree DA: **Use of plan schemata in the recall and recognition of goal-directed actions.** *J Exp Psychol Learn Mem Cogn.* 1983; **9**: 117–129.
Publisher Full Text

Brown C, Hagoort P: **The processing nature of the N400: Evidence from masked priming.** *J Cogn Neurosci.* 1993; **5**(1): 34–44.
PubMed Abstract | **Publisher Full Text**

Dehaene S, Kerszberg M, Changeux JP: **A neuronal model of a global workspace in effortful cognitive tasks.** *Proc Natl Acad Sci U S A.* 1998; **95**(24): 14529–14534.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Ghosh VE, Gilboa A: **What is a memory schema? A historical perspective on current neuroscience literature.** *Neuropsychologia.* 2014; **53**(1): 104–114.
PubMed Abstract | **Publisher Full Text**

Hagoort P: **Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations.** *J Cogn Neurosci.* 2003; **15**(6): 883–899.
PubMed Abstract | **Publisher Full Text**

Haluts N, Trippa M, Friedmann N, *et al.*: **Professional or amateur? the phonological output buffer as a working memory operator.** *Entropy (Basel).* 2020; **22**(6): 662.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Hintze A: **On the literary structure of the older Avesta.** *Bull Sch Orient Afr Stud.* 2002; **65**(1): 31–51.
Publisher Full Text

Liu S, Erkkinen MG, Healey ML, *et al.*: **Brain Activity and Connectivity During Poetry Composition: Toward a Multidimensional Model of the Creative Process.** 2015.
Publisher Full Text

Naim M, Boboeva V, Kang CJ, *et al.*: **Reducing a cortical network to a Potts model yields storage capacity estimates.** *J Stat Mech Theory Exp.* 2018; **2018**(4): 043304.
Publisher Full Text

Najme D, Mahdi P, Majid ZA: **Oral-Oriented Teaching of Meter through the Use of Music: Proposing a New Method of Teaching Meter in Poetry.** *J Literary Arts.* 2020; **12**(4): 73–96.
Publisher Full Text

Norman DA, Shallice T: **Attention to action: Willed and automatic control of behavior BT - Consciousness and Self-Regulation.** *Consciousness and Self-Regulation.* 1986; **4**: 1–18.
Reference Source

Obermeier C, Kotz SA, Jessen S, *et al.*: **Aesthetic appreciation of poetry correlates with ease of processing in event-related potentials.** *Cogn Affect Behav Neurosci.* 2016; **16**(2): 362–373.
PubMed Abstract | **Publisher Full Text**

Peirce J, Gray JR, Simpson S, *et al.*: **PsychoPy2: Experiments in behavior made easy.** *Behav Res Methods.* 2019; **51**(1): 195–203.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Reece S: **Epithets**. In M. Finkelberg (Ed.), *Homeric Encyclopedia.* Blackwell. 2011; 257–259.

Robey D: **Language and Style in the *Divine Comedy*.** *Roman Stud.* 1985; **3**(1): 112–127.
Publisher Full Text

Robin J, Moscovitch M: **Details, gist and schema: hippocampal-neocortical interactions underlying recent and remote episodic and spatial memory**. In *Curr Opin Behav Sci.* Elsevier Ltd. 2017; **17**: 114–123.
Publisher Full Text

Rubin DC: **Memory in oral traditions: The cognitive psychology of epic, ballads, and counting-out rhymes.** In *Memory in oral traditions: The cognitive psychology of epic, ballads, and counting-out rhymes.*1995.
Reference Source

Rumelhart DE: **Notes on a schema for stories.** In: *Representation and understanding*, Morgan Kaufmann. 1975; 211–236.
Publisher Full Text

Sachs JS: **Recognition memory for syntactic and semantic aspects of connected discourse.** *Percept Psychophys.* 1967; **2**(9): 437–442.
Publisher Full Text

Schulkind MD: **Long-term memory for temporal structure: Evidence from the identification of well-known and novel songs.** *Mem Cognit.* 1999; **27**(5): 896–906.
PubMed Abstract | **Publisher Full Text**

Šeļa A, Gronas M: **Measuring rhythm regularity in verse: entropy of inter-stress intervals.** CHR 2022: Computational Humanities Research Conference, Dec 12-14, Antwerp, Belgium, CEUR Workshop Proceedings (CEUR-WS.org). 2022.
Reference Source

Shannon CE, Weaver W: **The mathematical theory of communication**. University of Illinois Press. 1949.
Reference Source

Spalla D, Cornacchia IM, Treves A: **Continuous attractors for dynamic memories.** *eLife.* 2021; **10**: e69499.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Stella F, Urdapilleta E, Luo Y, *et al.*: **Partial coherence and frustration in self-organizing spherical grids.** *Hippocampus.* 2020; **30**(4): 302–313.
PubMed Abstract | **Publisher Full Text**

Stoet G: **PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments.** *Teach Psychology.* 2017; **44**(1): 24–31.
Publisher Full Text

Taylor SE, Crocker J: **Schematic Bases of Social Information Processing**. In Higgins, E.T. *et al.* (Ed.), *Social Cognition: The Ontario Symposium.* Lawrence Erlbaum Associates. 1981.
Reference Source

Tillmann B, Dowling WJ: **Memory decreases for prose, but not for poetry.** *Mem Cognit.* 2007; **35**(4): 628–39.
PubMed Abstract | **Publisher Full Text**

Tse D, Langston RF, Kakeyama M, *et al.*: **Schemas and memory consolidation.** *Science.* 2007; **316**(5821): 76–82.
PubMed Abstract | **Publisher Full Text**

Van Kesteren MTR, Ruiter DJ, Fernández G, *et al.*: **How schema and novelty augment memory formation.** *Trends Neurosci.* 2012; **35**(4): 211–219.
PubMed Abstract | **Publisher Full Text**

Wang SH, Morris RGM: **Hippocampal-neocortical interactions in memory formation, consolidation, and reconsolidation.** *Annu Rev Psychol.* 2010; **61**: 49–79, C1-4.
PubMed Abstract | **Publisher Full Text**

www.maldura.unipd.it ("Archivio Metrico Italiano", accessed November 2018).

www.prolific.co (accessed December 2020).

Open Peer Review

Current Peer Review Status: ✔ ✘ ? ✘

Version 1

Reviewer Report 14 March 2022

https://doi.org/10.21956/openreseurope.14736.r28462

✘ **Mara Breen**
Department of Psychology and Education, Mount Holyoke College, South Hadley, MA, USA

The authors' major question is whether poetic meter serves as a cognitive schema that supports verbal memory. To assess this claim, they manipulated Italian poems by Dante and Ariosto to have a) consistent meter and rhyme, b) consistent meter but missing rhyme, c) conserved rhyme but incorrect accent pattern, or d) the incorrect # of syllables. These variations were assessed for metric plausibility, and this factor was the predictor in the behavioral assessments. Participants were trained with the poems, then tested on their ability to remember final words. The authors report that metric plausibility predicted reaction time, but only in the manipulations of the Ariosto poems. Greater metric plausibility led to more correct memory of the missing target words, but also led to increased reaction time.

I can't recommend this paper pass peer review. I have many questions about the method and results, including how the experiment was run, and the choice of statistical analysis models. In addition, I'm not sure that results are interpretable given the variability in the experimental materials across authors, verses, and manipulations.

Detailed Comments:

Introduction
I don't find that the Introduction as written effectively introduces the research question; nor does it effectively review the relevant literature.

At the end of the introduction, it's not clear what is being measured and what is being hypothesized. The authors say that they will "aim at measuring the strength of some metric devices." but it's not clear which devices they're specifically talking about, and, moreover, what specific predictions they're making about the devices. It would be helpful to see explicit predictions about the pattern of results.

Methods
I don't understand how the manipulation was implemented in the memory experiment. How were

conditions assigned to participants? Was it a Latin Square design? Did each participant encounter the same number of verses from each condition? How many ill-formed versions were participants presented with in total? Were there well-formed fillers included? I worry that if participants heard equal numbers of items from each condition that means that 75% of the items they heard were ill-formed, meaning they would not have a reason to expect well-formed items. If well-formed items were rare (25% of the time), then it's hard to interpret the reaction time differences as being a result of the manipulation and not a result of metacognitive processes on the part of the participants.

Test phase

- How are the participants choosing the muted word? Is it free response or do they choose from a set of alternatives?

- How is reaction time measured? That is, when does the clock start? Are any trials thrown away for having reaction times that are too long? Relatedly, I am quite surprised by the variation in reaction times and how generally long they are. Was there any instruction to the participants to answer as quickly as they could? What do the authors think participants are doing in these very long delays of six or more seconds?

I also don't understand how the authors are accounting for the fact that the targets across verses are different, both in terms of what the actual words are and the way they were produced across conditions. Speaking as someone who studies the role of prosody in cognitive processing, we know a considerable amount about how the acoustic realization of words contributes to how they are remembered. The fact that the production of the target words is not standardized across conditions (i.e., participants heard different acoustic renditions of the targets across conditions), how can the authors be sure that the differences in memory are due to poetic features?

I don't believe the analyses are appropriate here. As I see it, the memory experiment has two outcome variables – accuracy and reaction time. First, for accuracy, I expect to see a mixed-effects logistic regression, where the authors are analyzing outcomes on a trial-by-trials basis, modeling the probability of an accurate response based a set of fixed and random effects. Fixed effects include the experimental condition (NPR, NPA, NPS, ONP), the author (Dante or Ariosto), the metric plausibility as assessed in the pre-test, and any other variables that might influence the result (e.g., lexical frequency or entropy). The random effects should be participant and item (verse). The fixed effects should also include any interactions the authors want to test – for example, the interaction between condition and author.

For reaction time, the authors should be computing a mixed effects linear regression, modeling the reaction time on a trial-by-trials basis with a response based a set of fixed and random effects. Fixed effects include the experimental condition (NPR, NPA, NPS, ONP), the author (Dante or Ariosto) the metric plausibility as assessed in the pre-test, and any other variables that might influence the result (e.g., lexical frequency, entropy, direction boas). The random effects should be participant and item (verse).

Data analysis

"The outcome of interest is essentially the presence of a significant correlation between the dependent variables measured in the ranking (the plausibility index, see Figure 2) and in the memory experiment (correct responses, and reaction times, see Figure 4), that would indicate a joint dependence on the type of passage manipulation."

Note that there can't be a correlation between a continuous variable (plausibility index) and a binary variable (correct vs. incorrect).

Entropy in the accent distribution.
It's not clear why this is being measured. There needs to be some explanation in advance. I am expecting this measure to be included in the analysis somehow, perhaps as a fixed effect in the regression.

Results
This section is not well-written. The authors are mixing results with interpretation, but they should be separated. In addition, as stated above, I don't believe the analyses are appropriately used here. For example, it's hard to figure out what is being shown in Figure 4. It appears that the authors have averaged reaction times within conditions, then plotted those averages against the metric plausibility scores. But again, I don't believe this is the most appropriate analysis as the authors should be determining the simultaneous effect of all predictors on the outcomes using a regression fit over all trials.

Discussion
The first four paragraphs belong in the Introduction as these papers are framing the motivation for the current study. There is also a lot missing here. I would expect for the authors to restate the research question, the hypothesis, and the results. And then attempt to interpret the results in the theoretical framework established in the Intro. Ultimately, it's not clear what these data are adding to our understanding of how metric structure is processed.

Is the work original in terms of material and argument?
Yes

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?
No

Is the work clearly and cogently presented?
No

Is the argument persuasive and supported by evidence?
No

If any, are all the source data and materials underlying the results available?
Yes

Does the research article contribute to the cultural, historical, social understanding of the field?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Psycholinguistics, statistical modeling, metric processing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 21 Dec 2022

Alessandro Treves, SISSA, Trieste, Italy

Detailed Comments:

Introduction

I don't find that the Introduction as written effectively introduces the research question; nor does it effectively review the relevant literature.

At the end of the introduction, it's not clear what is being measured and what is being hypothesized. The authors say that they will "aim at measuring the strength of some metric devices." but it's not clear which devices they're specifically talking about, and, moreover, what specific predictions they're making about the devices. It would be helpful to see explicit predictions about the pattern of results.

Response: We thank the reviewer for her observation, however we think that the three components we consider, of what we broadly refer to as meter, are rather emphatically salient throughout the paper. As to hypothesis-testing science, it is definitely not our style, but if we can mention expectations, then clearly our expectations about the role of meter in Dante were not met by our findings. We thank the reviewer for stimulating us to make these disappointed expectations clearer in the revision.

Methods

I don't understand how the manipulation was implemented in the memory experiment. How were conditions assigned to participants? Was it a Latin Square design? Did each participant encounter the same number of verses from each condition? How many ill-formed versions were participants presented with in total? Were there well-formed fillers included? I worry that if participants heard equal numbers of items from each condition that means that 75% of the items they heard were ill-formed, meaning they would not have a reason to expect well-formed items. If well-formed items were rare (25% of the time), then it's hard to interpret the reaction time differences as being a result of the manipulation and not a result of metacognitive processes on the part of the participants.

Response: Indeed, the design can be loosely called a Latin Square, although each participant is only tested on a row of the square. The crucial element to make it balanced is that the 24 participants for each author exhaust the 24 possible rows. This is now better explained and the term Latin Square is included to facilitate the reading: For the experiment, every subject was administered four texts in total, by the same (original) author: one per canto passage and one per condition (a Latin square design). Therefore, twenty-four combinations were created. Although 3 of the 4 versions could be construed to be ill-formed on metrical grounds, as the reviewer notes, we did make an effort to make the

manipulations soft – as the revised text emphasizes – and the plausibility rankings show that they were indeed perceived as such.

Test phase

- *How are the participants choosing the muted word? Is it free response or do they choose from a set of alternatives?*

Response: The text says: Three written alternatives appeared on the screen and the participant had to read aloud his/her choice.

- *How is reaction time measured? That is, when does the clock start? Are any trials thrown away for having reaction times that are too long? Relatedly, I am quite surprised by the variation in reaction times and how generally long they are. Was there any instruction to the participants to answer as quickly as they could? What do the authors think participants are doing in these very long delays of six or more seconds?*

Response: There were no instructions that they had to answer as quickly as possible. Considering the task, and the processing cost it takes, we are not particularly surprised by differences in time: some participants enjoyed reverberating the verses more than others; what is indicative, given the balanced design, are the differences across conditions. Outliers were already excluded in the analysis as previously reported.

I also don't understand how the authors are accounting for the fact that the targets across verses are different, both in terms of what the actual words are and the way they were produced across conditions. Speaking as someone who studies the role of prosody in cognitive processing, we know a considerable amount about how the acoustic realization of words contributes to how they are remembered. The fact that the production of the target words is not standardized across conditions (i.e., participants heard different acoustic renditions of the targets across conditions), how can the authors be sure that the differences in memory are due to poetic features?

Response: The muted/target words were actually the same across conditions, as already noted: Notably, for every passage we chose options which were consistent across all conditions, allowing a fair comparison in the results. As the reviewer observes, clearly altering the accent pattern, or the number of syllables, or ablating the rhymes will have affected in subtle ways the way the actor recited the non-poems, despite the target words being overtly untouched by the manipulations. This is part of the effects we aimed to measure: we do not distinguish in this study between the effects of the alterations that would be evident in a written transcription and those that only emerge from the acoustically expressed verse prosody.

I don't believe the analyses are appropriate here. As I see it, the memory experiment has two outcome variables – accuracy and reaction time. First, for accuracy, I expect to see a mixed-effects logistic regression, where the authors are analyzing outcomes on a trial-by-trials basis, modeling the probability of an accurate response based a set of fixed and random effects. Fixed effects include the experimental condition (NPR, NPA, NPS, ONP), the author (Dante or Ariosto), the metric plausibility as assessed in the pre-test, and any other variables that might influence the result (e.g., lexical frequency or entropy). The random effects should be participant and item (verse). The fixed effects should also include any interactions the authors want to test – for example, the interaction between condition and author.

Response: We ask in the Discussion "Can the hypothesis of differential schema activation be tested experimentally?" as a speculation, and to encourage further work, also by others, but our study is hypothesis-free, which we believe to be one of its strengths. As discussed in the response to the other reviewers, the manipulations were intended to be soft, and indeed the shuffling analysis now reported in the extended data confirms that only the NonPoem without Rhymes sounded significantly abnormal: In terms of significance, we do report the significance and non-significance of the main finding, the plausibility-memory score correlations for Ariosto and Dante, respectively, but we believe readers remain free to interpret the data however they prefer. See linked figure here.

For reaction time, the authors should be computing a mixed effects linear regression, modeling the reaction time on a trial-by-trials basis with a response based a set of fixed and random effects. Fixed effects include the experimental condition (NPR, NPA, NPS, ONP), the author (Dante or Ariosto) the metric plausibility as assessed in the pre-test, and any other variables that might influence the result (e.g., lexical frequency, entropy, direction boas). The random effects should be participant and item (verse).

Response: We do report the wide distribution of reaction times, which makes the shuffling analysis redundant: again, we do report the significance of the correlation …in each of the four types of trials above, the more metrically plausible the passage, the longer the reaction time. However, the trend is significant only with Ariosto, if data from the two authors are analyzed separately, and it is significant overall ($p<0.004$) with a slope mainly determined by the Ariosto data, if analyzed together, as shown in Figure 4 (linked here). The slope for the Dante data alone would be higher, but not significant, likely because of the limited plausibility range spanned.

Data analysis

"The outcome of interest is essentially the presence of a significant correlation between the dependent variables measured in the ranking (the plausibility index, see Figure 2) and in the memory experiment (correct responses, and reaction times, see Figure 4), that would indicate a joint dependence on the type of passage manipulation."

Note that there can't be a correlation between a continuous variable (plausibility index) and a binary variable (correct vs. incorrect).

Response: In fact, the memory score is not a binary variable, as shown clearly in Fig. 4: it spans a range from 31 to 45.

Entropy in the accent distribution*.*

It's not clear why this is being measured. There needs to be some explanation in advance. I am expecting this measure to be included in the analysis somehow, perhaps as a fixed effect in the regression.

Response: The two entropy measures we present are two possible synthetic descriptions of the variability of accent patterns across the 4 passages we have selected for each author.

Other measures are of course possible. We now cite another one from a recent paper: These two entropy measures appear no less sensitive, with our passages, than others recently proposed (see e.g. Sela & Gronas, 2022). but of course any such measure can only give a rough indication. It is also worthwhile to note, particularly to anglophone readers, that accent patterns are much less defined and less binary in Italian, a language with less marked consonant dominance (Ramus et al, 1999)

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. Cognition, 73(3), 265-292.

Results

This section is not well-written. The authors are mixing results with interpretation, but they should be separated. In addition, as stated above, I don't believe the analyses are appropriately used here. For example, it's hard to figure out what is being shown in Figure 4. It appears that the authors have averaged reaction times within conditions, then plotted those averages against the metric plausibility scores. But again, I don't believe this is the most appropriate analysis as the authors should be determining the simultaneous effect of all predictors on the outcomes using a regression fit over all trials.

Response: We have taken, as always in our research, a hypothesis-free approach, which leaves readers free to interpret the observations as they wish. If anything, again, our expectations were contradicted by the findings.

Discussion

The first four paragraphs belong in the Introduction as these papers are framing the motivation for the current study. There is also a lot missing here. I would expect for the authors to restate the research question, the hypothesis, and the results. And then attempt to interpret the results in the theoretical framework established in the Intro. Ultimately, it's not clear what these data are adding to our understanding of how metric structure is processed.

Response: We feel the previous response applies to this criticism as well. Maybe one indication emerging from these data is precisely that there is no unique way metric structure is processed, as passages with similar meter are differently affected by our delicate manipulations.

Competing Interests: No competing interests were disclosed.

Reviewer Report 27 January 2022

https://doi.org/10.21956/openreseurope.14736.r28183

Johann-Mattis List
Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena, Germany

I read the study *In poetry, if meter has to help memory, it takes time* by Andreetta *et al*. with great interest. Poetry, reflected in meter and rhyme, plays an important role in historical linguistics, my own field of research. The analysis of rhyme patterns allows us to reconstruct ancient stages of a language which might not be reflected in written sources (see Baxter 1992 for an overview on Old Chinese phonology). Assuming that meter often reflects – at least to a certain degree – the prosody of a language may allow us to reconstruct prosodic structures for languages with attested poetic traditions (Kümmel 2018). Given the importance of poetry, meter, and rhyme for a variety of scientific fields, the study conducted by the authors is of great value and has the potential to inspire follow-up studies.

Due to my background in historical linguistics and computational linguistics, I do not feel confident to comment on questions of the experimental design or the interpretation of the results. As a result, I will try to provide comments which may help the authors to make their study more accessible to a broader audience and improve its transparency.

First, I would suggest that the authors clarify the process of poem manipulation a bit more. They give one figure as an example, but I find it hard to interpret the figure, since I did not really understand which parts are manipulated and which parts are original. Investing more time in the design of the figure is probably worthwhile, as it would help the readers to understand more clearly what happens with the poems, where pseudowords are inserted, etc. Given that the examples are in Italian (as far as I asssume, even in a historical Italian variety that is far from being used today), I'd furthermore suggest to provide translations (maybe even using English pseudowords) of all passages (including the data showing in the supplementary PDF).

Second, I'd like to encourage the authors to explain a bit more about the sociolinguistic role which Dante and Ariosto play in Italy today. If I understand this correctly, both text must look rather archaic to modern language users, so I wonder to which degree the application of ancient rather than modern poetry might have had an impact on the results.

Third, I would like to encourage the authors to share some more information on *Schema theory* and some other terms and frameworks they use. As far as my reading of their text is concerned, I had the impression that the authors assume that their readers will know these terms, so it would not be important to explain them in more detail, but given that their study may be interesting for a broader reading circle, it may be worthwile to rephrase and extend the introduction and also the discussion to be more inclusive with respect to readers from different scientific backgrounds.

Fourth, I found the data not very well explained, as already mentioned in my comment made on the article earlier:

What I miss from the current study, however, are more explicit explanations on the data which you have shared (detailed description of column names, which information is used where in the article, etc.), and also that you share more detailed information on the software that was used for plotting. For example, you mention the use of the SUBTLEX-IT data for assessing word frequencies, but I had to look quite a bit when I was trying to find where in the data files you had

this information provided. In order to avoid that readers interested in the details of your methods have to second-guess what part of the data relates to what part of the article, it is always recommended to be very verbose about the data, ideally providing a README file that provides all necessary information, specifically explaining what one can find in which column. As a scientist who has been struggling a lot with studies in which code is not being shared fully, I'd also recommend to share your plotting code for the individual data plots, also in order to allow young scholars to learn from your expertise.

To summarize my comment here: the supplementary data should be *explained* more transparently, ideally by double-checking with the FAIR principles of data sharing (Wilkinson *et al*. 2016). Additionally, I'd ask the authors to also share the code they used for their plots, as this is a major requirement for replicability and it also helps younger scholars not experienced in doing plots and the like, to learn from the authors in their own work.

As a fifth and last point, I'd like to ask the authors to which degree they have tried to make sure that the pseudowords they use are neutral across the poems: could it be possible that by coincidence they selected pseudowords that differ with respect to their memorizability, e.g., because they are more or less phonetically isolated? I do not know if studies on this question exist, but I could imagine that certain pseudowords are harder to learn than others, maybe because they are phonotactically less common. If the author know of studies that have looked into such differential characteristics of pseudowords, it may be useful to discuss them quickly.

References

1. Baxter W. H: A Handbook of Old Chinese Phonology. *de Gruyter.*1992.
2. Kümmel, M. J: "Silbenstruktur Und Metrik: Neues Zum Altavestischen" In eds O. Hackstein, D.Gunkel Language and Meter. *Brill*. 2018. 129-57
3. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, et al.: The FAIR Guiding Principles for scientific data management and stewardship.*Sci Data*. 2016; **3**: 160018 PubMed Abstract | Publisher Full Text

Is the work original in terms of material and argument?

Yes

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?

Yes

Is the work clearly and cogently presented?

Partly

Is the argument persuasive and supported by evidence?

Partly

If any, are all the source data and materials underlying the results available?

Partly

Does the research article contribute to the cultural, historical, social understanding of the

field?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: historical linguistics, computational linguistics, digital humanities

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 21 Dec 2022

Alessandro Treves, SISSA, Trieste, Italy

First, I would suggest that the authors clarify the process of poem manipulation a bit more. They give one figure as an example, but I find it hard to interpret the figure, since I did not really understand which parts are manipulated and which parts are original. Investing more time in the design of the figure is probably worthwhile, as it would help the readers to understand more clearly what happens with the poems, where pseudowords are inserted, etc. Given that the examples are in Italian (as far as I asssume, even in a historical Italian variety that is far from being used today), I'd furthermore suggest to provide translations (maybe even using English pseudowords) of all passages (including the data showing in the supplementary PDF).

Response: We thank the reviewer for this suggestion, expressed also by the other reviewers, to which we have responded by emphasizing the attempt to keep manipulations minimal, by better explaining the accent alterations, and by clarifying that muted non-words were chosen to be the same across conditions, and in positions other than where the manipulations occurred. See also the annotated pdf in the revised Extended Data. We refrain from providing translations in English of the surviving words, because they would look awkward interspersed with the non-words, but the translations of the original Dante and Ariosto passages are of course readily available online, in a number of different variants – how to translate them has been a major issue in itself.

Second, I'd like to encourage the authors to explain a bit more about the sociolinguistic role which Dante and Ariosto play in Italy today. If I understand this correctly, both text must look rather archaic to modern language users, so I wonder to which degree the application of ancient rather than modern poetry might have had an impact on the results.

Response: Indeed, this is an important point, which it is difficult to discuss in a few sentences, given also the heterogeneity of the young Italian population and of their educational experiences. Broadly speaking, Dante remains, even among many who have really not studied his poetry in school, the respected avuncular figure of a genius who, almost single-handedly, with his creativity made the Florentine dialect into standard Italian. The celebration of 700 years from his death stimulated several festive initiatives around the country, with public readings, etc. Ariosto is nowadays much less read, but easier to read and quite enjoyable. They can both be argued to be associated, in the public imagination, with notions of boundless imagination and freedom, in contrast to the often oppressive

normativity and rule-learning associated e.g. to Latin grammar, in the traditional educational framework. We enjoy the opportunity to exchange this comments with the reviewer, but feel that they would be somewhat out of place, without proper evidence to support them, in a scientific paper.

Third, I would like to encourage the authors to share some more information on Schema theory and some other terms and frameworks they use. As far as my reading of their text is concerned, I had the impression that the authors assume that their readers will know these terms, so it would not be important to explain them in more detail, but given that their study may be interesting for a broader reading circle, it may be worthwile to rephrase and extend the introduction and also the discussion to be more inclusive with respect to readers from different scientific backgrounds.

Response: Thank you for this other important comment, which have addressed by substantially revising the two relevant paragraphs: A schema, whether directly functional like those involved in preparing coffee (Norman & Shallice, 1986) or social/ornamental, like rituals of salutations (Taylor & Crocker, 1981), can be considered as a set of regularities that help organize and retrieve information (Van Kesteren et al., 2012). Its neural basis would be the memory trace of those regularities, that helps funnel neural activity so as to reinstate them. In language processing, "narrative schemata" have often been described, as the features of stories which make them easier to remember, sometimes called "story grammars" (Rumelhart, 1975). In the present context, however we are considering metrical schemata: patterns which, by somewhat restricting options and encouraging expectations, facilitate the recall of verses. Thus, meter (broadly defined) should help us recruit, and possibly produce, the next element of a sequence stored in our memory. In facilitating verbal sequence replay, metrical features appear to be effective with extended "trajectories", lasting even several verses. These are extended relative to the short trajectories thought to be produced by the phonological loop of Baddeley's model of working memory, which are presumed to last only a couple of seconds, precisely because of the lack of specific devices that extend their range (Baddeley, 1992; Haluts et al, 2020).

Fourth, I found the data not very well explained, as already mentioned in my comment made on the article earlier:

What I miss from the current study, however, are more explicit explanations on the data which you have shared (detailed description of column names, which information is used where in the article, etc.), and also that you share more detailed information on the software that was used for plotting . For example, you mention the use of the SUBTLEX-IT data for assessing word frequencies, but I had to look quite a bit when I was trying to find where in the data files you had this information provided. In order to avoid that readers interested in the details of your methods have to second-guess what part of the data relates to what part of the article, it is always recommended to be very verbose about the data, ideally providing a README file that provides all necessary information, specifically explaining what one can find in which column. As a scientist who has been struggling a lot with studies in which code is not being shared fully, I'd also recommend to share your plotting code for the individual data plots, also in order to allow young scholars to learn from your expertise.

Response: We thank the reviewer for this suggestion, a README file is now included in the

supplementary material, where indications about what each column represents are provided. The Extended Data are also now clearer.

To summarize my comment here: the supplementary data should be explained more transparently, ideally by double-checking with the FAIR principles of data sharing (Wilkinson et al. 2016). Additionally, I'd ask the authors to also share the code they used for their plots, as this is a major requirement for replicability and it also helps younger scholars not experienced in doing plots and the like, to learn from the authors in their own work.

Response: Disappointingly perhaps, our plots were simply made in Excel, and assembled in Powerpoint, also for ease of communication among us. We thank the reviewer for this comment, that made us discover a bug in the published Figure 2, perhaps due to our using pedestrian software: the chance level has dropped below its correct 0.25 value! We do not know at what stage in the article production this happened, and apologize. I will be corrected in the revised version (the figure is correct in our folders).

As a fifth and last point, I'd like to ask the authors to which degree they have tried to make sure that the pseudowords they use are neutral across the poems: could it be possible that by coincidence they selected pseudowords that differ with respect to their memorizability, e.g., because they are more or less phonetically isolated? I do not know if studies on this question exist, but I could imagine that certain pseudowords are harder to learn than others, maybe because they are phonotactically less common. If the author know of studies that have looked into such differential characteristics of pseudowords, it may be useful to discuss them quickly.

Response: We gratefully acknowledge this comment by the reviewer: this is indeed an aspect we are aware of, but know of no systematic way to handle it, other than relying on our own best judgement. On the one hand, we did our best to manipulate the passages, while on the other trying to avoid an effect of a totally unrelated schema, as we report in the text: an effort was made to minimize the impact of this manipulation (like for the subsequent ones below) by changing phonemes with similar ones, while maintaining the original prosody. Function words were not modified. The degree to which we succeeded can be assessed by readers by looking at the muted words and at their alternatives in the Underlying Data Excel files. It would indeed be nice to be able to rely on more on more objective criteria, but they seem to be a long way behind. For example, arguably the most advanced automated system for generating literary language, GPT-3, has churned out a pathetic attempt when challenged with the incipit of a mere sonnet by Dante (Floridi and Chiriatti, 2020). Clearly, more research is needed in this respect. Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30(4), 681-694.

Competing Interests: No competing interests were disclosed.

Reviewer Report 15 November 2021

https://doi.org/10.21956/openreseurope.14736.r27840

Stefan Blohm
[1] Centre for Language Studies, Radboud University, Nijmegen, The Netherlands
[2] Department of Language and Literature, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany
Stefano Versace
Leopardi Centre, University of Birmingham, Birmingham, UK

SUMMARY
Focusing on two poetical works of classical Italian literature, the study reported in this article tests the idea that poetic meter and rhyme aid verbal memory.

Stimulus materials: The authors selected text sections of 8-9 verse lines from Dante's *Divina Commedia* and from Ariosto's *Orlando Furioso*. By substituting individual speech sounds in a portion of lexical words, they converted the original text sections into meaningless (but grammatical) jabberwocky versions that preserved the prosodic structure and the rhyme scheme. These were then modified to yield versions with (a) an irregular number of syllables in some verse lines, (b) an irregular distribution of prominent/accented syllables within some verse lines, and (c) no rhyme in any of the verse lines.

Procedure/method: Using audio recordings of the four resulting versions of each text section, the authors conducted two experiments. In one experiment, participants listened to four critical sections of one work (Ariosto: *n*=62; Dante: *n*=65)––each in one of the four experimental conditions––and ranked them according to "poetic plausibility"; resulting rank data were used to calculate the non-linearly weighted average rank per condition: the "metric plausibility index".

The main experiment comprised a study phase and a test phase after 24 hours. During study, participants (*n*=48) listened to four critical sections of one work, each in one of the four experimental conditions. During test, participants were presented with individual lines from the text sections they had heard a day before. In each critical line a single (pseudo-)word was muted, and participants had to choose the muted target word from three alternatives presented on a screen.

Data analysis: The authors calculated correlations between the metric plausibility index per condition and the condition means of both the response latencies and the accuracy rates observed in the memory experiment; correlation analyses were conducted separately for the *Divina Commedia* and for *Orlando Furioso*.

Results: The authors observed that rankings were sensitive to metrical modifications and particularly to the removal of rhyme. The metrical plausibility index of the *Divina Commedia* was unrelated to either reaction times or response accuracy, whereas the metrical plausibility index of *Orlando Furioso* correlated positively with response accuracy and negatively with reaction times.

Conclusion: From these results, the authors concluded that metre facilitates memory retrieval, but that this facilitation requires additional time and cognitive effort.

COMMENTS
Is the work original in terms of material and argument?
YES, because – contrary to prior investigations of accentual-syllabic metre (e.g., Menninghaus et al., 2014; van Peer, 1990) – this study aims to dissociate the metrical constraints on (a) the number of syllables and (b) the distribution of syllable prominence/accent within the verse line.

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?
NO, because...

1. The current version of the article disregards most of the relevant empirical research into memory effects of rhyme and metre (e.g., metre and memory: van Peer, 1990; rhyme and memory: Bower & Bolton, 1969; Lea et al., 2021; Rubin & Wallace, 1989).
> This concern can be addressed by relating the present study to a broader range of relevant findings (for a recent overview of parallelism-induced memory effects, see Blohm *et al*., 2021).

2. The authors applied the conversion into pseudo-word verse inconsistently and less systematically than prior investigations relying on this text-modification method (for poetry, e.g., Obermeier *et al*., 2013). Due to these shortcomings it remains unclear in how far the text conversion successfully removed the meaning of the texts; the reported effect of lexical frequency (p. 8 and Fig. 6) casts further doubt on the effectiveness of this procedure. Specifically, modifications:
    - Failed to convert all target words of the memory experiment (2/12 in Ariosto and 1/12 in Dante remain unchanged).
    - Targeted only a subset of lexical classes, e.g., nouns but not adverbs.
    - Targeted only a subset of lexical words per word class.
    - Did not substitute all consonants, which is the common procedure.
    - Did not substitute speech sounds systematically, e.g., voiceless plosives with voiceless plosives.
    - Sometimes resulted in actual words rather than pseudo-words, e.g., *lascio* [I leave] > *bascio* [I kiss]; *lena* [haste] > *rena* [sand].
    - Sometimes substituted only a single speech sound per verse line, e.g., *a privilegi **l**enduti e mendaci.*

> Unfortunately, we don't see how this concern could be addressed a posteriori.

Is the work clearly and cogently presented?
PARTLY, because...

1. The assumed underlying mechanisms are not made explicit enough.
    - The authors appear to assume that activation/recognition of metrical schemata is crucial for facilitated memory retrieval, but how exactly schema recognition is supposed to facilitate the retrieval of pseudo-words remains opaque. Clearly, the pseudo-word targets cannot be part of the metrical schema.
    - On page 10, the authors further point to Rubin's idea that the constraints of rhyme restrict

the set of alternatives, or possible continuations, in an unfolding sentence/verse (e.g., Bower & Bolton, 1969; Rubin & Wallace, 1989). This explanation seems less plausible for non-lexical items (=pseudo-words) and for the relatively weak constraints of metre, which are consistent with a much larger portion of the lexicon than the constraints of rhyme.

2. The procedure and the analysis are not described in sufficient clarity and detail.
   - For instance, it is unclear how participants indicated their ranking in the ranking experiment.
   - The authors state (p. 6) that participants navigated through the texts in a self-paced manner but that it was the experimenter who logged participants' responses. Thus it remains unclear how the experiment was controlled, e.g., which machine actually ran the experiment (the experimenter's? the participant's?).
   - It is unclear whether the Zoom screen, i.e., the experimenter, was visible at all times during the experiment, and in how far the experimenter interacted with the participants during study and test.

3. Not all decisions regarding the research design and the analysis procedure are sufficiently well motivated. In particular, the current version of the article does not make clear enough...
   - Why the authors chose to present jabberwocky verse rather than the original texts?
   - Why the authors chose the detour via the metrical plausibility index rather than comparing conditions directly?
   - Why the rank data have been transformed? While the authors state (p. 5) that this was intended to accommodate individual participants' uncertainty regarding the ranking, we suspect that this non-linear transformation was applied to make the data fit a priori assumptions: without this non-linear transformation, original pseudo-word versions of *Orlando Furioso* and versions with an irregular number of syllables per line were indistinguishable in terms of metrical plausibility.
   - Why the authors opted for an indirect response collection?
   - Why the authors chose to vary the number of repetitions per text section during study?
   - Why the authors chose to conduct correlation analyses rather than, say, linear and logistic regression analyses?
   - Why the authors chose to analyse data from the *Divina Commedia* and from *Orlando Furioso* separately?
   - Why the authors report analyses of frequency effects and response bias?

4. The statistical analysis is not described in sufficient detail to allow for replication (e.g., which software was used), and the statistical results are not reported according to the conventional standards of the field (e.g. in the format recommended by the APA). Moreover, the reported results seem to be at odds with the values supplied in the data sheets.
> These concerns can be addressed by providing more information about (1) the assumed underlying cognitive mechanisms, (2) the experimental procedure, (3) the motivation behind the authors' decisions, and (4) the statistical analysis.

Is the argument persuasive and supported by evidence?
NO, because of a number of methodological flaws that seriously undermine the validity of the results and of the inferences drawn from them.

1. The most serious issue relates to the indirect response collection employed in this study. Participants reported their choices to the experimenter who then logged responses via button press. This procedure conflates the behaviour (i.e., latencies and accuracy of responses) of both

participants and experimenter. Since it is impossible to dissociate the contributions of participants and experimenter, this conflation renders the behavioural results of the memory experiment uninterpretable.
> Unfortunately, we don't see how this issue could be successfully addressed.

2. The sample sizes (*n*=4) are too small for correlation analyses. This increases the likelihood of both type I and type II errors and inflates the correlation coefficient of significant effects (Aggarwal & Ranganathan, 2016; Knudson & Lindsey, 2014; Makin & Orban de Xivry, 2019).
> This concern can partly be addressed by pooling all data and by calculating correlations per text section. However, we strongly recommend to conduct ANOVA and/or regression analyses instead in order to test for differences between experimental conditions.

3. The between-participant design confounds participant group and author, i.e., observed differences between Dante's and Ariosto's verse could in fact merely reflect differences between participants. While the authors maintain that participant groups were indistinguishable in terms of age and sex, they do not provide evidence that groups were indistinguishable in terms of their prior experience with poetry (as assessed e.g. by self report). Prior experience with poetry is necessary for the emergence of the metrical schemata whose activation is assumed to be crucial for facilitated memory retrieval.
> This concern can be addressed by pooling all data and by not interpreting differences between authors.

If any, are all the source data and materials underlying the results available?
PARTLY, because the authors do provide the materials and the underlying data. However, it is unclear at times how these data relate to the results reported in the article. For instance, the correlation coefficients reported on page 8 do not match the values supplied in the data sheet.
> This concern can be addressed by (additionally) providing data in a more user-friendly format, e.g., in a single comma/tab-delimited text file. The reader comment by Johann-Mattis List generally seems to provide sensible suggestions; please also add participant IDs and the number of repetitions during study to the data table.

Does the research article contribute to the cultural, historical, social understanding of the field?
NO, because – due to the methodological issues outlined above – it remains unclear what we can learn from this study. Unfortunately, the most serious flaw is an inherent feature of the research design and cannot be remedied. We had been enthusiastic to read about this research, but after several thorough readings, we are convinced that the reported results do not support the conclusion that metre takes its time if it has to help memory.

MINOR POINTS

- End rhyme is not commonly considered a component of metre (cf. Lev Blumenfeld's review report) but rather a para-metrical phenomenon or: a structuring part of the orchestration of the metrical line.
- Extended data: Please underline target words in the supplementary file; please place images next to the text excerpts they were presented with.
- Please make sure to use terminology consistently, e.g., canti/poems/texts.

References
1. Menninghaus W, Bohrn I, Altmann U, Lubrich O, et al.: Sounds funny? Humor effects of phonological and prosodic figures of speech.*Psychology of Aesthetics, Creativity, and the Arts*. 2014; **8** (1): 71-76 Publisher Full Text
2. Van Peer W: The measurement of metre. *Poetics*. 1990; **19** (3): 259-275 Publisher Full Text
3. Bower G, Bolton L: Why are rhymes easy to learn?. *Journal of Experimental Psychology*. 1969; **82** (3): 453-461 Publisher Full Text
4. Lea RB, Elfenbein A, Rapp DN: Rhyme as resonance in poetry comprehension: An expert-novice study.*Mem Cognit*. **49** (7): 1285-1299 PubMed Abstract | Publisher Full Text
5. Rubin D, Wallace W: Rhyme and reason: Analyses of dual retrieval cues.*Journal of Experimental Psychology: Learning, Memory, and Cognition*. 1989; **15** (4): 698-709 Publisher Full Text
6. Blohm S, Kraxenberger M, Knoop C, Scharinger M: Sound Shape and Sound Effects of Literary Texts. 2021. 7-38 Publisher Full Text
7. Obermeier C, Menninghaus W, von Koppenfels M, Raettig T, et al.: Aesthetic and emotional effects of meter and rhyme in poetry.*Front Psychol*. 2013; **4**: 10 PubMed Abstract | Publisher Full Text
8. Aggarwal R, Ranganathan P: Common pitfalls in statistical analysis: The use of correlation techniques.*Perspect Clin Res*. **7** (4): 187-190 PubMed Abstract | Publisher Full Text
9. Knudson D, Lindsey C: Type I and Type II Errors in Correlations of Various Sample Sizes. *Comprehensive Psychology*. 2014; **3**. Publisher Full Text
10. Makin T, Orban de Xivry J: Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife*. 2019; **8**. Publisher Full Text

Is the work original in terms of material and argument?

Yes

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?

No

Is the work clearly and cogently presented?

Partly

Is the argument persuasive and supported by evidence?

No

If any, are all the source data and materials underlying the results available?

Partly

Does the research article contribute to the cultural, historical, social understanding of the field?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: SB: language/text comprehension, literary linguistics, poetic form(s); SV:

metrics, literary linguistics, Italian literature

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 05 Dec 2021

Alessandro Treves, SISSA, Trieste, Italy

We thank the two reviewers for their most attentive reading of our manuscript and their useful comments. Many of the issues they raise cannot, as they note, be addressed ***a posteriori***. Still, we would like to better explain some of the main ***a priori*** choices in our study.
First, although it is of course impossible to neatly separate semantics from syntactic and metric structure, we tried our best to use our subjective judgement to create original non-poems that strike a balance between removing meaning as much as possible while retaining structure. This motivated sometimes altering voicing and other phonemic features, masterfully used especially by Dante to convey the sense/tone in some verse. The result is in our view an acceptable compromise.

Second, the choice to conduct the experiment over Zoom sessions was also an acceptable compromise, between the need to reach an adequate number of subjects, in particular during the pandemic, and that of establishing a relationship of trust and common purpose with the experimenter while she (the first author) maintained an objective but caring posture. This is particularly delicate in that we obviously required subjects to have had the exposure to Dante and Ariosto normally available in the Italian school system, without being specialists or, the opposite, harbouring long-lasting negative emotions from such exposure. With our recruitment and Zoom sessions we found that no subject had to be excluded on such basis.

Third, the training procedure, with a variable number of repetitions depending on the position of each verse in the passage, was designed to facilitate encoding, vaguely mimicking a common rote learning used in school. The proof of its validity was in the pudding, in that subjects achieved memory performances in the intended intermediate range, away from floor and ceiling effects, for the entire length of the passages.

Regarding the metric plausibility index (which allows an across-subjects correlation with memory performance, whereas comparisons between conditions were within-subjects) it was developed when the study was still limited to Dante, and then applied also to the Ariosto passages.

Finally, we would like to thank the reviewers for their bibliographic suggestions, some of which we were not aware of, including the recent Blohm *et al* (2021) study, which will undoubtedly be taken into account in our future endeavours.

Competing Interests: None

Reviewer Response 09 Dec 2021

Stefan Blohm, Radboud University, Nijmegen, The Netherlands

Thanks to the authors for their replies to our review. Unfortunately, the major issues have not been addressed in this response (experimenter-mediated response collection, statistical analysis), but we are confident that this will be rectified in the revised manuscript.

Considering that the two reviews so far have come to diverging assessments, it might be best to seek a third opinion. Since our criticism focused almost exclusively on the methodology, we trust that the authors will seek a third opinion from a scholar with a solid background in empirical research (e.g., an expert in language-related memory research) to ensure that the scientific quality of this contribution will be properly assessed.

In the meantime, we provide comments on the points raised in the author response:

1. The authors state that “it is of course impossible to neatly separate semantics from syntactic and metric structure”. The manipulation in question aimed to remove semantics while keeping syntactic and metric structure intact. As prior studies show, it is perfectly possible to do this to the extent that meaning is conveyed by content words, and to do it systematically and exhaustively. Contrary to what is suggested in the reply, there is no need to strike a balance between removing meaning and retaining structure; it is possible to have both if the appropriate procedure is employed. Moreover, the conventional procedure of systematically substituting speech sounds within phoneme classes even allows to keep some sub-phonemic features (e.g., voicing) constant and, in fact, preserves virtually the entire sonority profile of linguistic stimuli.

2. We readily accept that the pandemic situation requires alternative ways of doing experimental research, but we doubt that it necessitates the kinds of choices made in this study (see the comments below). Most importantly: COVID-19 is no excuse for flawed experiment design or inappropriate statistical analysis.
   • It is customary in experimental research that the experimenter refrains from influencing the outcome of the experiment, e.g., by interacting with participants during task performance, or by influencing dependent variables directly. Please note that we do not wish to accuse the first author of consciously manipulating the outcome of the study; however, we feel obliged to point out that it is highly problematic to draw valid conclusions if the experimenter, who is familiar with the hypotheses, modulates the dependent variables
   • As the reported online study demonstrates, it is possible to recruit an adequate number of participants without video conference.
   • Of course, participants should trust that they will be informed about what is expected from them, that the experimenter acts according to accepted scientific criteria, and that the collected data will be handled in compliance with legal and ethical standards. However, we fail to see the need to establish “a relationship of trust and common purpose with the experimenter” beyond that. On the contrary, this is usually seen as an undue influence on participants’ task performance.
   • It is possible to ensure without personal interaction that participants “have had the exposure to Dante and Ariosto normally available in the Italian school system,

without being specialists or, the opposite, harbouring long-lasting negative emotions from such exposure"; collecting appropriate self-report measures would have been a straightforward objective alternative.

3. Repeated exposure helps to consolidate memory, and we assume that performance during test varies as a function of the number of repetitions during study. We recommend that the authors make sure to detail in the revised manuscript how a variable vs. fixed number of repetitions facilitates encoding.

4. The authors disclosed in their response that the study had originally been limited to Dante and that the metric plausibility index was introduced only after the Dante study had been completed.
Frankly, the authors' a posteriori decisions 1) to extend the experiment after learning that the original study did not yield significant results, 2) to opt for an inappropriate statistical analysis that is prone to yield significant results in the absence of an actual effect, 3) to take the detour via the metric plausibility index (between-participant correlations instead of more convincing within-participant contrasts), and 4) to non-linearly adjust the calculated index to obtain differences between conditions do not increase readers' confidence in the scientific rigour of this study. Therefore, we recommend that the authors devote particular care during revision to the motivation of these decisions.
We further recommend that the authors clarify in the article that the Dante sessions were not followed by a ranking experiment (the current version of the article states that participants "were asked to complete this survey after the end of the second session of the main experiment"), and whether all of the ranking data for Ariosto were obtained from participants recruited via Prolific.

5. We are glad to learn that the authors will take the recommended empirical literature into account, and we hope that these suggestions help to better contextualize the current study. However, the authors should not feel obliged to cite our own work unless they consider it relevant to the purpose of their study.

Thanks again to the authors for taking the time to explain some of their decisions in their response. We look forward to reading the revised version of the manuscript.

Competing Interests: No competing interests were disclosed.

Author Response 21 Dec 2022

Alessandro Treves, SISSA, Trieste, Italy

COMMENTS

Is the work original in terms of material and argument?

YES, because – contrary to prior investigations of accentual-syllabic metre (e.g., Menninghaus et al., 2014; van Peer, 1990) – this study aims to dissociate the metrical constraints on (a) the

number of syllables and (b) the distribution of syllable prominence/accent within the verse line.

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?

NO, because…

1. The current version of the article disregards most of the relevant empirical research into memory effects of rhyme and metre (e.g., metre and memory: van Peer, 1990; rhyme and memory: Bower & Bolton, 1969; Lea et al., 2021; Rubin & Wallace, 1989).
> This concern can be addressed by relating the present study to a broader range of relevant findings (for a recent overview of parallelism-induced memory effects, see Blohm et al., 2021).

Response: We thank the reviewer for this suggestion: at the time of submitting this paper we were not aware of this contribution. We have now included it in the discussion: Still, the wide-ranging contribution of sound 'shape' to cognitive processes has been noted, in particular in poetry (Blohm et al. 2021) and it might play a role in our findings, in forms not readily evident.

We have also added the remark that In line with previous studies, we also acknowledge that a potential factor could also be the individual participants' expertise with poetry and/or music. Such data were not considered, here, and presumably individual differences average out, but these factors are being investigated in a related study.

2. *The authors applied the conversion into pseudo-word verse inconsistently and less systematically than prior investigations relying on this text-modification method (for poetry, e.g., Obermeier et al., 2013). Due to these shortcomings it remains unclear in how far the text conversion successfully removed the meaning of the texts; the reported effect of lexical frequency (p. 8 and Fig. 6) casts further doubt on the effectiveness of this procedure. Specifically, modifications:*

- *Failed to convert all target words of the memory experiment (2/12 in Ariosto and 1/12 in Dante remain unchanged).*
- *Targeted only a subset of lexical classes, e.g., nouns but not adverbs.*
- *Targeted only a subset of lexical words per word class.*
- *Did not substitute all consonants, which is the common procedure.*
- *Did not substitute speech sounds systematically, e.g., voiceless plosives with voiceless plosives.*
- Sometimes resulted in actual words rather than pseudo-words, e.g., *lascio* [I leave] > *bascio* [I kiss]; *lena* [haste] > *rena* [sand].
- Sometimes substituted only a single speech sound per verse line, e.g., *a privilegi **l** enduti e mendaci.*

Response: Each of the points above is a valid concern, which is unfortunately difficult to address *a posteriori*. Satisfying all the listed constraints together, however, would have produced in our judgement repellent non-poems, inappropriate or a study intended to engage, however marginally, the aesthetic enjoyment of subjects drawn from the general Italian population. A useful guide, in this respect, is the success enjoyed by the famous non-

poem Il Lonfo, by Fosco Maraini (in F. Maraini, *Gnosi delle fànfole*, Dalai Editore, 1994), in which contagiously enjoyable meta-semantics is achieved by a judicious admixture of words and non-words.

Is the work clearly and cogently presented?
PARTLY, because…

1. The assumed underlying mechanisms are not made explicit enough.
- *The authors appear to assume that activation/recognition of metrical schemata is crucial for facilitated memory retrieval, but how exactly schema recognition is supposed to facilitate the retrieval of pseudo-words remains opaque. Clearly, the pseudo-word targets cannot be part of the metrical schema*.

Response: Our report does not discuss, for they would be out of place, models and conjectures about the neural mechanisms involved in the representation of meter. Some of the neural computation research in our group touches on those issues. One general idea, however, is that schemata in general are expressed as rather loose dynamical attractors, which guide the evolution in time of distinct patterns of neural activity to a partial degree, compete with each other, often fail together, and are therefore quite different structure from the rules, with a well-defined domain of application, typically assumed in linguistic analyses. An interesting discussion to be continued elsewhere.

- *On page 10, the authors further point to Rubin's idea that the constraints of rhyme restrict the set of alternatives, or possible continuations, in an unfolding sentence/verse (e.g., Bower & Bolton, 1969; Rubin & Wallace, 1989). This explanation seems less plausible for non-lexical items (=pseudo-words) and for the relatively weak constraints of metre, which are consistent with a much larger portion of the lexicon than the constraints of rhyme*.

Response: Agreed. The citation of Rubin is not meant to affirm the absolute validity of that explanation in our case, but simply to introduce one relevant and inspiring idea.

2. *The procedure and the analysis are not described in sufficient clarity and detail.*
- *For instance, it is unclear how participants indicated their ranking in the ranking experiment*.

Response: It is actually written that "At the end they were asked to rank the four poems: from the one they perceived as the best, to the one that sounded worse to them, by simply typing in the platform their preferential order."

- *The authors state (p. 6) that participants navigated through the texts in a self-paced manner but that it was the experimenter who logged participants' responses. Thus it remains unclear how the experiment was controlled, e.g., which machine actually ran the experiment (the experimenter's? the participant's?).*

Response: It is actually written that "When tested with the muted non-word (see below), participants would read aloud the left, central or right alternative appearing on the screen, and the experimenter would press the corresponding key (leftward, downward or rightward) on her own keyboard."

- *It is unclear whether the Zoom screen, i.e., the experimenter, was visible at all times during the experiment, and in how far the experimenter interacted with the participants during study and test*.

Response: We believe that any doubt about the experimental procedure has been clarified in the revised text.

3. *Not all decisions regarding the research design and the analysis procedure are sufficiently well motivated. In particular, the current version of the article does not make clear enough…*

- *Why the authors chose to present jabberwocky verse rather than the original texts*?

Response: This was in fact a rather crucial element of the study: to eliminate or strongly suppress the semantic component, which is so prominent when listening to poetry by both authors. We note most content words were converted into non-words in order to eliminate discernible semantic content, hence semantic effects on memory. We were, in fact, interested in the strength of meter devices, and a semantic effect would have been a major distorting bias.

- *Why the authors chose the detour via the metrical plausibility index rather than comparing conditions directly?*
- *Why the rank data have been transformed? While the authors state (p. 5) that this was intended to accommodate individual participants' uncertainty regarding the ranking, we suspect that this non-linear transformation was applied to make the data fit a priori assumptions: without this non-linear transformation, original pseudo-word versions of Orlando Furioso and versions with an irregular number of syllables per line were indistinguishable in terms of metrical plausibility.*
- *Why the authors opted for an indirect response collection?*
- *Why the authors chose to vary the number of repetitions per text section during study?*
- *Why the authors chose to conduct correlation analyses rather than, say, linear and logistic regression analyses?*
- *Why the authors chose to analyse data from the Divina Commedia and from Orlando Furioso separately?*
- *Why the authors report analyses of frequency effects and response bias?*

Response: Overall it feels like the reviewers are asking: why did you prepare couscous with mullet, almonds and saffron rather than spaghetti with anchovies, toasted bread crumbs, capers and olives. They will surely agree that both are nutritious and tasty food to those that like them.

4. *The statistical analysis is not described in sufficient detail to allow for replication (e.g., which software was used), and the statistical results are not reported according to the conventional standards of the field (e.g. in the format recommended by the APA). Moreover, the reported results seem to be at odds with the values supplied in the data sheets.*
> These concerns can be addressed by providing more information about (1) the assumed underlying cognitive mechanisms, (2) the experimental procedure, (3) the motivation behind the authors' decisions, and (4) the statistical analysis.

Response: The new figure added to the Extended data with the shuffling analyses should clarify that the statistical analysis was actually very simple. As already pointed out but now further emphasized, our manipulations were delicate and succeeded in not altering much the meter and prosody of our non-poems: only the NPR versions of the Ariosto passages was rated as significantly less plausible at the modest $p < 0.05$ significance level. The memory score differences were similarly limited (see the Extended Data for a comparison with their shuffled distribution). The resulting correlation between the two measures (tight correlation in the case of Ariosto, no correlation for Dante) are presented transparently in

Figure 4.

Is the argument persuasive and supported by evidence?

NO, because of a number of methodological flaws that seriously undermine the validity of the results and of the inferences drawn from them.

1. The most serious issue relates to the indirect response collection employed in this study. Participants reported their choices to the experimenter who then logged responses via button press. This procedure conflates the behaviour (i.e., latencies and accuracy of responses) of both participants and experimenter. Since it is impossible to dissociate the contributions of participants and experimenter, this conflation renders the behavioural results of the memory experiment uninterpretable.
> Unfortunately, we don't see how this issue could be successfully addressed.

2. The sample sizes (n=4) are too small for correlation analyses. This increases the likelihood of both type I and type II errors and inflates the correlation coefficient of significant effects (Aggarwal & Ranganathan, 2016; Knudson & Lindsey, 2014; Makin & Orban de Xivry, 2019).
> This concern can partly be addressed by pooling all data and by calculating correlations per text section. However, we strongly recommend to conduct ANOVA and/or regression analyses instead in order to test for differences between experimental conditions.

3. The between-participant design confounds participant group and author, i.e., observed differences between Dante's and Ariosto's verse could in fact merely reflect differences between participants. While the authors maintain that participant groups were indistinguishable in terms of age and sex, they do not provide evidence that groups were indistinguishable in terms of their prior experience with poetry (as assessed e.g. by self report). Prior experience with poetry is necessary for the emergence of the metrical schemata whose activation is assumed to be crucial for facilitated memory retrieval.
> This concern can be addressed by pooling all data and by not interpreting differences between authors.

Response: We believe the previous responses addressed the concerns expressed here in somewhat dramatic form. The effect of major first-person expertise at versification is actually a very interesting topic, although we believe not relevant to the participants in this study. It is has been taken up in a separate study by others in our group.

If any, are all the source data and materials underlying the results available?

PARTLY, because the authors do provide the materials and the underlying data. However, it is unclear at times how these data relate to the results reported in the article. For instance, the correlation coefficients reported on page 8 do not match the values supplied in the data sheet.
> This concern can be addressed by (additionally) providing data in a more user-friendly format, e.g., in a single comma/tab-delimited text file. The reader comment by Johann-Mattis List generally seems to provide sensible suggestions; please also add participant IDs and the number of repetitions during study to the data table.

Response: This was done. Thank you for the suggestion.

Does the research article contribute to the cultural, historical, social understanding of the field?

NO, because – due to the methodological issues outlined above – it remains unclear what we can learn from this study. Unfortunately, the most serious flaw is an inherent feature of the research design and cannot be remedied. We had been enthusiastic to read about this research, but after several thorough readings, we are convinced that the reported results do not support the conclusion that metre takes its time if it has to help memory.

Response: We regret the reviewers reached this sad persuasion.

MINOR POINTS

- *End rhyme is not commonly considered a component of metre (cf. Lev Blumenfeld's review report) but rather a para-metrical phenomenon or: a structuring part of the orchestration of the metrical line.*

Response: Agreed. We use the term in a broader sense.

- *Extended data: Please underline target words in the supplementary file; please place images next to the text excerpts they were presented with*.

Response: Thank you for the suggestion. The Extended Data is now presented in a clearer format

- *Please make sure to use terminology consistently, e.g., canti/poems/texts*.

Response: Thank you again. We have revised the wording throughout the text. Now "passages" are the consistently excerpts from specific "canti" and once they are manipulated in different types of "non-poems" they become "texts".

Competing Interests: No competing interests were disclosed.

Reviewer Report 28 October 2021

https://doi.org/10.21956/openreseurope.14736.r27806

Lev Blumenfeld
Carleton University, Ottawa, ON, Canada

The article attempts to test the contributions of various aspects of metricality to memory. Non-word metrical passages derived from Dante and Ariosto were manipulated in one of three ways to render them less regular: syllable count, accent distribution, and rhyme. The resulting passages were ranked by listeners, and a metric "plausibility" rating was extracted from those rankings. In the second experiment, a different subjects were presented with the same passages, and tested

on their memory retention of one of the words in them. The basic results were: (a) metric plausibility positively correlates with memory retention for Ariosto but not for Dante, (b) reaction times are longer for incorrect responses than for correct responses, and (c) reaction times POSITIVELY correlate with metric plausibility.

I believe the paper makes a significant, original contribution to the literature and can pass peer review with very minor clarifications and additions.

GENERAL REMARKS:

1. While the research is framed around the question of meter's role in memory retention, it is worth emphasizing that the experiment does not compare meter with non-meter. Rather, the comparison is between meter and "almost-meter". A passage that has been manipulated to make it less regular in one respect is still regular in its other properties. The authors could discuss whether this issue makes their claims stronger (because their approach isolates specific aspects of meter) or weaker (because it does not test entirely unmetrical passages).

2. If I understand the systems right, the requirements of rhyme and syllable count are absolute, i.e. a line deviating from the 11-syllable count and the ottava or terzina rhyme simply could occur in Dante or Ariosto. On the other hand, accent distribution, other than in the 10th syllable, is not strictly regulated, and various accent patterns may be more or less likely but not absolutely unmetrical. It is then interesting that violations of absolute requirements in the NPS conditions are not ranked worse than violations of soft requirements (NPA). Does this mean that the subjects of the experiments are not in fact proficient in the knowledge of the relevant metrical systems and were not in fact perceiving the intended metrical structure, especially with reference to syllable count? Does this make the results of the paper weaker?

SPECIFIC REMARKS:

1. Where in the lines were the muted non-words? In rhyming position? in the middle? Were the muted non-words the ones causing metrical irregularity in the manipulated passages?

2. More details should be provided on how accent was manipulated in the NPA condition, and how that manipulation relates to the actual preferences or requirements of Dante's and Ariosto's meters.

3. It is odd to refer to rhyme as a subcategory of meter. Normally "meter" refers to the distribution of prominences (accents, syllable weight) in a line. Perhaps "formal" is a better cover term for both meter and rhyme, and "formal plausibility" could be used instead of "metrical plausibility".

Is the work original in terms of material and argument?

Yes

Does it sufficiently engage with relevant methodologies and secondary literature on the topic?

Yes

Is the work clearly and cogently presented?
Partly

Is the argument persuasive and supported by evidence?
Yes

If any, are all the source data and materials underlying the results available?
Yes

Does the research article contribute to the cultural, historical, social understanding of the field?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Phonology, historical linguistics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 21 Dec 2022

Alessandro Treves, SISSA, Trieste, Italy

GENERAL REMARKS:

1. *While the research is framed around the question of meter's role in memory retention, it is worth emphasizing that the experiment does not compare meter with non-meter. Rather, the comparison is between meter and "almost-meter". A passage that has been manipulated to make it less regular in one respect is still regular in its other properties. The authors could discuss whether this issue makes their claims stronger (because their approach isolates specific aspects of meter) or weaker (because it does not test entirely unmetrical passages).*

   We thank the reviewer for bringing up this important point which, with the comments by the other reviewers, made us realize that we had not emphasized enough our efforts to make the manipulations as soft as possible, to keep the text “almost” metrically correct and try to stay within a presumed linear regime. This is now clearer in the main text, and particularly with the addition of the new Fig.8 (in the Extended Data), that shows, through a comparison with shuffled responses, that the effects of the manipulations on plausibility and memory score are essentially small, and hence allow us to focus on their correlation.

2. *If I understand the systems right, the requirements of rhyme and syllable count are absolute, i.e. a line deviating from the 11-syllable count and the ottava or terzina rhyme simply could occur in Dante or Ariosto. On the other hand, accent distribution, other than in the 10th syllable, is not strictly regulated, and various accent patterns may be more or less likely but not absolutely unmetrical. It is then interesting that violations of absolute*

requirements in the NPS conditions are not ranked worse than violations of soft requirements (NPA). Does this mean that the subjects of the experiments are not in fact proficient in the knowledge of the relevant metrical systems and were not in fact perceiving the intended metrical structure, especially with reference to syllable count? Does this make the results of the paper weaker?

Indeed, we regard this as an interesting and somewhat unexpected result of our study. Although syllable number and rhyme are absolute requirements only in theory (a bag of acceptable tricks is available to occasionally elude them) we did select passages where they apply strictly, and it is true that it is much simpler to write a sort code to check them out, rather than to assess, with a computer, the validity of an accent pattern. Because of that, we took extra care in quantifying, as presented in Fig. 3, the amount of variability in those patterns in our material. Still, the intended recipients of Dante's and Ariosto's poetry were not computer codes, but roughly speaking the ancestors, with probably less formal education, of our subject cohorts.

SPECIFIC REMARKS:

1. *Where in the lines were the muted non-words? In rhyming position? in the middle? Were the muted non-words the ones causing metrical irregularity in the manipulated passages?*

   Thank you for noting this point, which we had left unclear. We carefully avoided the rhyming positions and those causing the other metrical irregularities. We have now specified in the Study Phase section: Muted non-words were usually positioned in the third verse. and further below The alternatives to the correct non-word were chosen by maintaining the same number of syllables, and the same accent. Typically, stems and intermediate vowels or consonants changed. Again, target non-words were generally in the third verse, aiming not to overload working memory from the moment they listened to the silent word until the test time. In a few cases where this was not possible (e.g., because there were no appropriate non-words in the third verse) a non-word in the second verse was chosen, towards the end. Notably, for every passage we chose options which were consistent across all conditions, allowing a fair comparison in the results.

2. *More details should be provided on how accent was manipulated in the NPA condition, and how that manipulation relates to the actual preferences or requirements of Dante's and Ariosto's meters.*

   We are aware that accent device was the trickiest one. However, it has a crucial role in the Italian poetic tradition. We addressed this issue at our best, by consulting an expert and by referring to an Italian database collecting several literary masterpieces, and their metrical annotation. This is described also in the text: This manipulation was particularly non-trivial, since accent patterns are not rigidly defined. However, to validate proper original accents, we consulted with an expert scholar for *Orlando Furioso*, whereas for *Divina Commedia* we referred to the "Archivio Metrico Italiano", a database collecting masterpieces of Italian literature with their accents annotated (www.maldura.unipd.it). On these bases, we altered the "original accents" by putting them in different positions within the verse, taking advantage in particular of non-words with no mandatory accent.

3. *It is odd to refer to rhyme as a subcategory of meter. Normally "meter" refers to the distribution of prominences (accents, syllable weight) in a line. Perhaps "formal" is a better cover term for both meter and rhyme, and "formal plausibility" could be used instead of "metrical plausibility".*

   We now better clarify in the text that we refer to “meter” (the word we would mostly use in everyday conversation) in a broader sense.

Competing Interests: No competing interests were disclosed.

Comments on this article

Version 1

Author Response 26 Mar 2022

Alessandro Treves, SISSA, Trieste, Italy

We thank all reviewers for their useful comments. In particular, we have appreciated the suggestion regarding the need of a clearer description of the data, especially in the supplementary files. For this reason, in that section it is now possible to find:

- a readme file
- better annotations about the manipulated features in the texts

In addition, we would like to thank the reviewers also for their bibliographic suggestions, some of which we were not aware of, and we would possibly include them in future developments of this study. We would also like to address some of the comments: some of the concerns cannot be really addressed ***a posteriori***, but we still hope to clarify our approach. There were doubts about the validity of the manipulation of metrical patterns. We hope to have made them clearer by adding details in the supplementary files, and here we would like to better define the aim we had:

- we had no intention to create a completely meter-less version of our poems; on the contrary, we wanted to enable participants to partially activate their metrical schemata. In order to do that, it was crucial that the versions created retained each a specific combination of elements of meter;
- analogously, we did not aim at testing the effect of removing meter altogether: our goal was to quantify the relative importance of those specific elements that we removed one-by-one;
- the altered metric patterns we used are fully described in the “Poem manipulation” section

of the manuscript; these three specific elements were chosen because of their major, traditional role in the Italian literary tradition exemplified by Dante and Ariosto, a choice validated by the literary experts we consulted; but of course in a different cultural context another choice may have been more appropriate;

- admittedly, some of the patterns were harder to manipulate, and this applies in particular to the NPS condition. Indeed, as specified in the paper, "[in the NPS manipulation] by adding or subtracting one or two syllable, also the pattern of accents was perforce altered, but we attempted to make the alteration less noticeable than the number change, in contrast to the NPA condition in which, while there were strictly 11 syllables/verse, the accents followed more unusual patterns."

To realize the memory task, muted non-words were selected among those consistent across conditions, i.e., that they did not change from one condition to another, allowing then a balanced comparison of effects. They were generally in the third verse, so as not to overload working memory. This position was unfortunately not always possible, given the original Dante and Ariosto passages, and in those cases, we chose the muted non-words close to the end of the second verse. From the language point of view, we would also like to specify that, while it is true that both poets wrote in now outdated Italian, different from that spoken by (and sometimes hard to fully understand for) our participants, it would have been normally possible for them to comprehend the general meaning or at least the gist of the original passages. To get rid of this effect on memory, we decided to use pseudowords. In doing this, we tried to maintain the original prosody as much as possible. We cannot exclude the possibility that some pseudowords could be more memorable than others. In principle, however, that should not affect much the task outcome, because of the balanced design. This is in any case a comment we would like to consider in a follow-up of the study, which looks at statistical aspects of the prosody of these and other poets. As a concluding remark, we are aware that this study leaves several questions open, which we believe to be normal, when addressing a relatively unexplored issue with novel, if quite simple, methodology. Later studies are welcome to adopt more conventional hypothesis-driven paradigms, taking us a step further in clarifying the cognitive mechanisms which allowed humankind to transmit verbal information even before the advent of the writing system. Luckily, the conclusions from our study are by no means final; beyond technical improvements, there lies a vast ocean of cultural diversity, with independent poetic traditions requiring different points of view, and novel experimental schemata, in order to understand the beauty of our cognitive schemata.

Competing Interests: No competing interests were disclosed.

Reader Comment 16 Jul 2021

Johann-Mattis List, Max Planck Institute for Evolutionary Anthropology, Leizpig, Germany

Thanks for this very interesting study. As a linguist interested in the evolution of poetry across languages and cultures, the work you are doing is of crucial importance. What I miss from the current study, however, are more explicit explanations on the data which you have shared (detailed

description of column names, which information is used where in the article, etc.), and also that you share more detailed information on the software that was used for plotting. For example, you mention the use of the SUBTLEX-IT data for assessing word frequencies, but I had to look quite a bit when I was trying to find where in the data files you had this information provided. In order to avoid that readers interested in the details of your methods have to second-guess what part of the data relates to what part of the article, it is always recommended to be very verbose about the data, ideally providing a README file that provides all necessary information, specifically explaining what one can find in which column. As a scientist who has been struggling a lot with studies in which code is not being shared fully, I'd also recommend to share your plotting code for the individual data plots, also in order to allow young scholars to learn from your expertise. Thanks again for this very interesting study. I am curious to see the reviews.

Competing Interests: No competing interests were disclosed.