Topographic analysis of semantic space

Jonas Jørgensen Telle, Tuva Løvås Gregersen and August Heggenhougen Leikvam

Data availability

The code and dataset used in this analysis is stored in a private GitHub repository which may be shared on request.

I. Introduction

Study of text – how it is produced and how it is read – has a long history across many different academic disciplines. Recent advances in Natural Language Processing and AI have made it possible to analyze text in new ways. One of the seminal insights of this line of work is that words represented as vectors, i.e. word embeddings, capture meaningful semantic properties, such as similarity as perceived by humans (Mikolov et al., 2013). Representing a text as vectors makes it possible to manipulate it mathematically (Odden et al., 2024) and to quantify topographic characteristics such as how fast it travels through the semantic space and how much semantic ground it covers. The idea that a text moves in an abstract semantic space is not new; Kurt Vonnegut stated in his master thesis from the 1940s that "*stories have shapes that can be drawn on graph paper"* (Vonnegut, cited in Toubia et al., 2021).

In our project this summer, we have used the concepts of semantic speed, semantic volume and circuitousness introduced by Toubia et al. (2021) to analyze texts. These terms can be interpreted as how fast a text moves between different semantic concepts, how much semantic ground the text covers and how optimally the text travels through different semantic concepts, respectively (Toubia et al., 2021).

Hypotheses and goal of the study

We wanted to study semantic properties of research articles, to see whether we could find differences in terms of semantic speed, semantic volume and circuitousness between i) scientific journals of different disciplines and ii) scientific journals of the same discipline of different rating. Our hypotheses were:

H1*: We predict statistically significant differences in topographic semantic properties between journals of different academic disciplines.*

H2: *We predict statistically significant differences in topographic semantic properties between journals from the same discipline, but with different rating.*

With the time available, it was not feasible to do a large study of multiple journals and disciplines. Thus, our project should be viewed as a pilot, potentially laying the groundwork for future research by creating general programs with which one can easily do topographic analysis of many journals. Still, we have performed a simple comparison of three different journals to demonstrate how one could go about testing these hypotheses. We thought that comparing different disciplines would be a simple check on the methodology, as we expected reasonably large differences here. Comparing journals of different rating is especially interesting, since this could both give new insights into what makes a scientific article attractive to top journals and – if topographic metrics are a good predictor of rating – constitute another metric with which to rate future publications.

The goal of this kind of text analysis is foremost to give quantitative metrics to a field else limited to the qualitative. Often, the production of scientific papers is conceptualized as an activity undertaken to write objectively about something "true", and the writing of a text itself is given little attention (Boghrati et al., 2023). However, Boghrati et al. (2023) makes the case that style or form of academic texts can influence how easy they are to read and remember, and potentially change how we evaluate the content or findings of these texts. Looking at quantified semantic properties of research articles could therefore give information about both the writer of the text, as explored in Berger and Toubia (2024) where topographic properties in a college application essay were predictive of the writer's future academic success, and about how the text will be read, understood and evaluated, as explored in "*How quantifying the shape of stories predicts their success*" by Toubia et al. (2021).

II. Theory

Text embeddings & the distributional hypothesis

In natural language processing, embedding is a method of representing text mathematically which captures *meaning*, which differentiates it from earlier Bag of Words-methods based on word frequencies (Mansurova, 2024). Embedding methods are trained on the context of words, meaning that they do not treat words as independent dimensions, but rather transform words (or text) into vectors in a high-dimensional continuous semantic space (Odden et al., 2024). In this semantic space, words or text with similar meaning are located close to each other.

Embedding models are based on the distributional hypothesis (Kutuzov, 2020:13). The distributional hypothesis is a linguistic theory with a set of assumptions about the nature of language and meaning (Sahlgren, 2008), which could be summarized by the quote: "*You shall know a word by the company it keep*s" (Firth, 1957, cited in Kutuzov, 2020:13). The model is empirical; it is based entirely on language data (Sahlgren, 2008). It postulates that statistical word co-occurrence distributions, given a large enough amount of natural language data, can capture meaning (Kutuzov, 2020:13). It seems to be a useful hypothesis for solving practical problems, and works well in Natural Language Processing (Kutuzov, 2020:25-28).

Our work is fundamentally based on embedding technology and the distributional hypothesis. In the following, we outline the central concepts that we use in this project: semantic speed, semantic volume and circuitousness.

Semantic speed

Some texts seem to move more quickly from one thing to another. A text which moves fast might spend less time on semantically related topics, while a text that moves more slowly does the opposite (Toubia et al., 2021). An object that covers a greater physical distance in less time can be said to be moving faster, and the same can be said of a text covering a greater semantic distance in less words (Berger & Toubia, 2023). One way to quantify the speed of a text is to measure the distance between the embeddings of consecutive chunks of text (Toubia et al., 2021).

There are two different lines of interpreting semantic speed, analogous to the key functions of language: It could reflect something about the i) *producer* of the text, or it could say something about the ii) *impact* it has on reader or consumer of it (Boghrati et al., 2023). It is difficult to claim anything about the writer of the text based on how fast it moves between concepts. For the reader, one could argue that a text which moves faster might require more cognitive effort to be read, but it can also make the text more engaging and exiting (Toubia et al., 2021). The fact that a text that requires more cognitive effort to read can be more engaging makes intuitive sense and is in line with the "effort paradox" in cognitive science; even though humans tend to avoid effort instinctively, there is evidence to support that effort does not necessarily reduce value, and that people can ascribe value to effort in itself (Inzlicht et al., 2018).

Semantic volume

Semantic volume can be conceptualized as how much ground a text covers. Volume is a global feature of a text and looks at, regardless of sequence, how much semantic space the text covers (Berger & Toubia, 2024).

Regarding an interpretation of volume, there is perhaps more to be said about the writer of the text. Covering more ground in a text, controlling for length, can involve generating ideas that combine different concepts, which can be linked to creativity (Berger & Toubia, 2024). Creativity is often conceptualized as involving associative processes, which leads to the generation of new ideas. One longstanding theory of creativity is "*the associative theory of creativity*" which claims that what characterizes creative people is that they have a higher capacity for *association*; a greater ability to make connections between concepts stored in their memory that are not obviously related (Beaty & Kenett, 2023). It is also possible to hypothesize that people who cover more ground have a deeper understanding of the concepts they are discussing (Berger & Toubia, 2024). For the reader, volume offers the same tradeoff as speed regarding cognitive effort; a large volume or ground covered offers the reader an opportunity to connect a wider range of topics, but may increase cognitive load (Toubia et al., 2021).

Circuitousness

Some texts visit the same concepts or themes again and again. To measure this feature, one can calculate the circuitousness of the text. Texts that are more circuitous will tend to move back and forth between the same semantic concepts more than texts that are less circuitous (Berger & Toubia, 2024). Berger and Toubia (2024) describe circuitousness as "*the extent to which the actual latent semantic path differs from the shortest path that starts and ends at the same point, and visits all the same points in between"*.

Is it difficult to say anything about the writer of a text if the text is highly circuitous. However, a highly circuitous text might allow the reader to create deeper connections between the themes or concepts and make it easier to integrate information (Toubia et al., 2021). Structural affect theory furthermore suggests that the ordering of a narrative can have different emotional effects on the reader (Piper & Toubia, 2023).

Earlier findings

Since Toubia et al. (2021) published the first paper outlining the concepts of semantic volume, semantic speed and circuitousness, several papers have explored whether these measures can have predictive power. In the following, we outline some of these findings.

In the 2021 paper, Toubia et al. evaluated 4,000 movies, 12,000 TV-show episodes and 29,000 academic papers and investigated the connection between semantic speed, volume and circuitousness, and success. They found that movies and TV-show episodes with higher semantic speed were rated more favorably. Furthermore, TV-show episodes with lower semantic volume were rated higher. For academic papers, the opposite was true for speed; papers with higher semantic speed were cited less. Academic papers with higher semantic volume and circuitousness, however, were cited more.

In another study, Dos Santos & Berger (2022) analyzed 40,000 movie scripts to see if semantic progression, operationalized as semantic speed between adjoining portions of narrative, influenced cultural success. The study sought to disentangle not just whether high semantic speed is good or bad, but whether effects of semantic speed have anything to do with what part of a text is being considered. The results of this study indicate that movies with higher semantic speed early on are evaluated less positively, and movies with higher semantic speed in the end are evaluated more positively (Dos Santos & Berger, 2022).

Piper & Toubia (2023) have also studied the semantic topographic properties of a sample of 2,348 books, finding that on average, fiction books have lower volume and are more circuitous than non-fiction books. In the same study, they compared adult fiction books with youth fiction books, and found that youth fiction typically had lower circuitousness. For all the books in the dataset, the results of this study indicate that narratives that travel further through semantic space (i.e. high minimum required speed), but in an efficient way (i.e. low speed), are more successful. For works of fiction, readers seem to demonstrate a preference for lower volume. (Piper & Toubia, 2023).

Berger & Toubia (2024) looked at 40,000 college application essays and found that the students whose essays had higher semantic volume and lower semantic speed had higher grade point averages during their college career, even when controlling for other relevant factors.

Human perceptions of semantic speed, semantic volume and circuitousness

Having outlined how the concepts of semantic speed, semantic volume and circuitousness can be understood and some previous findings relating to them, one central question is whether they capture something about texts that are meaningful or perceptible to humans.

Toubia et al. (2021) looked at whether human perceptions of speed, volume and circuitousness correspond to the measures they have used. Human rankings of semantic distance corresponded in 69/100 tests to machine-calculated rankings. In volume-judgement tasks human judgement corresponded with the automated volume measure 75% of the time. For circuitousness, human and machine also agreed 75% of the time (Toubia et al., 2021). Based on these tests, we can tentatively assume that semantic speed, volume and circuitousness capture something about a text that humans also perceive. This can further be argued with reference to the studies that find correlations between semantic topographic properties and measures of success, for example how well the text is received by readers.

III. Method

Dataset

This project is based on data from academic journals published in a timeframe of about ten years (2013-2024). We chose to compare three different academic journals from Wiley: *Noûs,* one of the top-ranked philosophy journals, *Journal of Applied Philosophy (JAP)*, a lower-ranked philosophy journal, and *Journal of Research in Science Teaching (JRSE),* a journal on science teaching. We found the rankings of the journals on the SCImago Journal Rank website. The reason we selected these three journals is that we wanted to compare semantic properties in i) journals from the same discipline but with different rankings and ii) journals from different disciplines. The reason we chose philosophy and science education journals specifically is that we wanted texts with few illustrations and figures, and we believed that journals from these disciplines would meet these requirements. The journal in science education was suggested by a researcher at CCSE. We wanted to compare journals with different rankings from the same discipline, partly because Toubia et al. (2021) found that articles with less speed and higher volume and circuitousness were cited more frequently. Thus, we thought it could be interesting to investigate whether articles from a highly ranked journal had these semantic properties on a group level. We chose only journals from Wiley due to i) different regulations and restrictions for bulk-downloading academic journals across publishers and ii) Wiley has an available and easily accessible API.

Collecting data

The first step in our process was to collect article metadata from the journals we selected from Crossref and create a data frame with this information. After creating the data frame, we removed issues published before 2013 in the journals. The reason we chose this timeframe was mainly due to issues with accessing pdfs from *JAP* before 2013 due to an apparent change in DOI-formatting for this journal, so that early DOIs in the Crossref metadata lead nowhere. We then checked the data for duplicates, using DOI and titles, and removed these. For *Noûs* we also removed all files listed as type “journal issue” (around 100) rather than “journal article" in our data frame. In *JRSE* and *JAP,* all files were listed as type “journal-article”. After doing this initial pruning, we collected pdfs from the data frame using Wiley’s API.

Preprocessing of text files

After collecting the pdfs, we converted them into txt files with OCR. Before embedding, we tried to remove parts of the texts with a different semantic function than the articles themselves, since such elements could influence our measurements. The preprocessing we did for all the journals included removing reference lists, notes, page numbers, headers/footers, special characters, short titles (such as "Introduction") and excessive spacing. We also removed all texts with less than 1000 words. Our reasoning behind this was that i) we embed with 250-word chunks, which would mean that these texts would be embedded to three points or less, and ii) we wanted to remove texts that were not research articles, such as editorials, errata, etc., which tend to be quite short. However, we only found four texts with less than 1000 words, all in *JRSE*. In *JAP* and *Noûs,* we did not find editorials or similar texts when manually going through a sample of our data.

We also did some journal-specific preprocessing. In the *Noûs* dataset, we removed lines with copyright signs and lines with email addresses. In the *JAP* dataset, we had some book reviews that we wanted to remove, and we removed all texts that had "book review" as one line and texts with "book review + number" in one line (book reviews had this as headers). In *JRSE* we removed editorials that had not been cut through the wordcount-approach. We did this by dropping texts from our dataset that had "editorial" as a header.

After preprocessing, we had 510 articles from *Noûs*, 578 articles from *Journal of Applied Philosoph*y and 697 articles from *Journal of Research in Science Teaching*. Since our dataset is quite large, we did not consider differences in the size of our groups to be a problem.

Embedding & chunk size

We used the "mixedbread-ai/mxbai-embed-large-v1" transformer from huggingface to embed the selected articles, embedding each text in chunks of 250 words, which is about the length of one paragraph. One potential limitation of our project was using only one text embedding model. However, Toubia et al. (2021) tested different models, and found the measures of semantic speed, volume, and circuitousness to be quite stable.

Initially, we wanted to do sentence-wise embedding of 25 words, but we found this to be quite unstable. We therefore investigated an unprocessed *Noûs*-article embedded in 25-word

chunks to find the biggest semantic leaps in the text. We expected this to be within the notes, the reference list or perhaps some figure or in a sentence split down the middle by a header/footer. Shockingly, the biggest distance was between the following chunks (from the same paragraph and with no need of preprocessing!): i) *"inspect, a glass of clear liquid); and [7] that the present king of France is not a dolphin (given the coherency of the view that"* and ii) *"presupposition failure strips a proposition of any truth value and that [7] is an instance of presupposition failure)? [5] is particularly worrisome. For with it"* (Barnett, 2000). Clearly, logicians use weird examples when illustrating a point. Choosing a larger chunk size would make our embeddings, and subsequent measurements, less sensitive to such weird examples. We also hypothesized that this would make our measurements less sensitive to other text elements such as tables and figures.

Having decided on a paragraph as the scale of our chunks, the choice of 250 words followed the work of Toubia, et al. (2021). Given a set scale (sentence/paragraph/section/etc.), there is no correct answer for the chunk size, and Toubia et al. (2021) compared their results chunking with 125 words and 375 words and found largely the same results in their analyses as with 250.

Measures of semantic speed, semantic volume and circuitousness

After embedding our dataset, we measured speed, volume and circuitousness for each academic article. We strongly recommend using a GPU to speed up the embedding; we used CUDA to facilitate this.

Semantic speed

To measure the semantic speed of a text, we measured the average semantic distance between consecutive chunks. Toubia et al. (2021) argues that it is reasonable to use Euclidean distance to measure semantic speed in the embedding space, amongst other reasons because distance is typically measured that way, and because it corresponds to the distance when plotting the path taken by the text in the word embedding space. Toubia et al. (2021) tested using cosine similarity instead and found the estimations of semantic speed to be similar across metrics.

Semantic volume

To calculate the volume of each text, we used the method outlined by Toubia et al. (2021): We found the volume of the minimum-volume enclosing ellipsoid containing all the points of the text. To find the minimum-volume ellipsoid, we solved the optimization problem detailed in Moshtagh (2005).

Since we used a chunk size of 250 words, our texts would always be transformed to less than 1024 points (the number of dimensions in our embedding model). When the number of points in a vector space is less than the number of dimensions, it is possible to cover all the points in the space with a "flat" ellipse. To get a non-zero volume, we therefore found the minimum volume enclosing ellipsoid in the subspace spanned by the embedding vectors. According to the supplementary material of Toubia et al. (2021), they did the same.

Since different texts may have a differing number of points, we normalized over dimensions by taking the geometric mean of the axes of the ellipsoid. This is representative of the volume since the factor of enlargement of the unit hypersphere is the inverse of the square root of the determinant of the center form ellipsoid matrix (or, equivalently, the product of the inverses of the square roots of the eigenvalues, which are indeed the lengths of the axes) (Toubia et al., 2021). Still, this seemed not to be quite a successful normalization in the extremes where the number of points is close either to zero or the number of dimensions.

Circuitousness

The calculation of circuitousness was done by finding the ratio of the speed of the actual latent semantic path and the minimum required speed: *circuitousness = speed / minimum required speed*. The minimum required speed can be found by solving a version of the travelling salesman problem, where the start and end points are fixed (Toubia et al., 2021). We used Google OR-tools to do this (Google OR-tools, 2023).

Statistical tests

Our initial idea was to use independent sample t-tests to compare semantic speed, volume and circuitousness between the journals. The independent sample t-test is used to check whether the

means of two samples are different (Kent State University Libraries, 2024). The first thing we did was to check that our data met the assumptions necessary to perform the t-tests. Ideally, we wanted to do a parametric test, as this is generally considered better *if* data aligns with the assumptions of the t-test.

One of the key assumptions for performing the t-test is that both of the datasets we want to compare are normally distributed (Kim & Park, 2019). We check for normality visually using histograms and Q-Q plots and formally with the Shapiro Wilk test. The histogram and the Q-Q plot additionally provides information about how many outliers we have. In the Shapiro Wilk test the null-hypothesis is that the sample tested originates from a normally distributed population. Thus, if the p-value is less than the alpha level (0.05), the null hypothesis is rejected and there is evidence that the data is not normally distributed ("Shapiro-Wilk test", 2024). For our purposes we want a p-value on the Shapiro Wilk test to be >0.05, so that we can proceed with the t-test.

Another assumption is that there is equal variance between the groups we want to compare (Kim & Park, 2019). We test for this using Levene's test. In Levene's test the null hypothesis is that the population variances are equal, and if the alpha level is >0.05 we have evidence for equality of variances ("Levene's test", 2024).

In the cases where our data is consistent with the assumptions of the t-test, we did a normal t-test. If our data did not meet the assumption of equality of variances we did a Welch t-test, and if our data did not meet the assumption of normality, we did a Mann-Whitney U test. To account for the risk of getting Type I errors (false positives) when doing multiple t-tests, we do a Bonferroni correction of the significance level. We try to interpret the statistical findings considering what we know about the different journals but do so in a highly tentative manner. In reporting our results, we continue to use abbreviations for the names of the journals: *Journal of Applied Philosophy* (JAP) and *Journal of Research in Science teaching* (JRST).

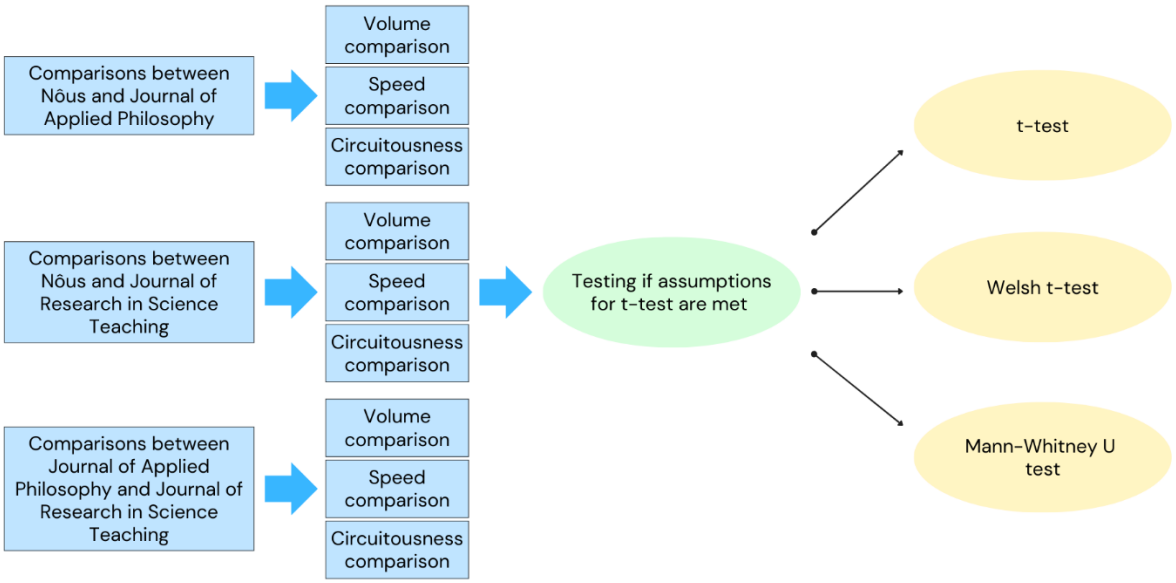

Figure 1: *Illustration of the work process.*

IV. Results

Tests of normality

	Volume	Speed	Circuitousness
Noûs	0.254	0.0718	0.00323
JAP	0.706	0.132	0.0683
JRST	0.567	0.0927	< 0.001

Table 1: *Shapiro-Wilk test of normality (p-values).*

As can be seen from the table above, most of the data is normally distributed according to the Shapiro-Wilk test ($p > 0.05$). The trend is that the volume samples have the highest p-values, the speed samples have lower p-values, and the circuitousness samples have the lowest p-values. We also plotted histograms and Q-Q plots for each sample, to help us decide whether a normal t-test would be appropriate to compare the samples. The plots are shown below.

There are two samples that are not normally distributed according to the Shapiro-Wilk test: circuitousness for both *Noûs* and *JRST*. We therefore chose to do the Mann-Whitney U test for the comparisons including the circuitousness of *JRST*. This choice is supported by the plots of circuitousness of *JRST*, which clearly shows that the sample is not normally distributed. From the plots of circuitousness of *Noûs*, we can see that there is one clear outlier. When removing this outlier, the Shapiro-Wilk test gave a p-value of 0.205. Since the sample is very large, we decided to remove this outlier for the *Noûs-JAP* comparison with regards to circuitousness, so that we could do the t-test. In all the other cases, we also proceeded with a t-test or Welsh t-test, depending on equality of the variances.

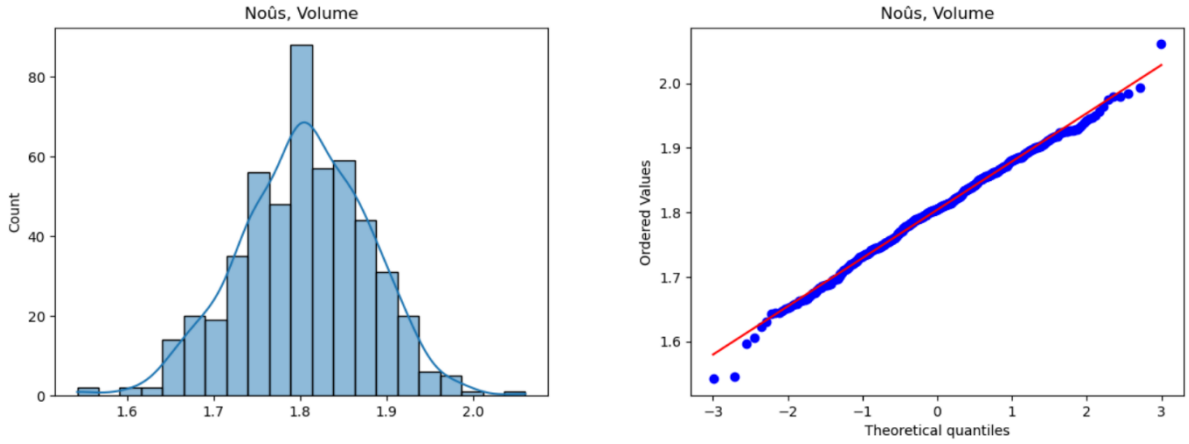

Figure 2A: *Histogram and Q-Q plot of volume in Noûs.*

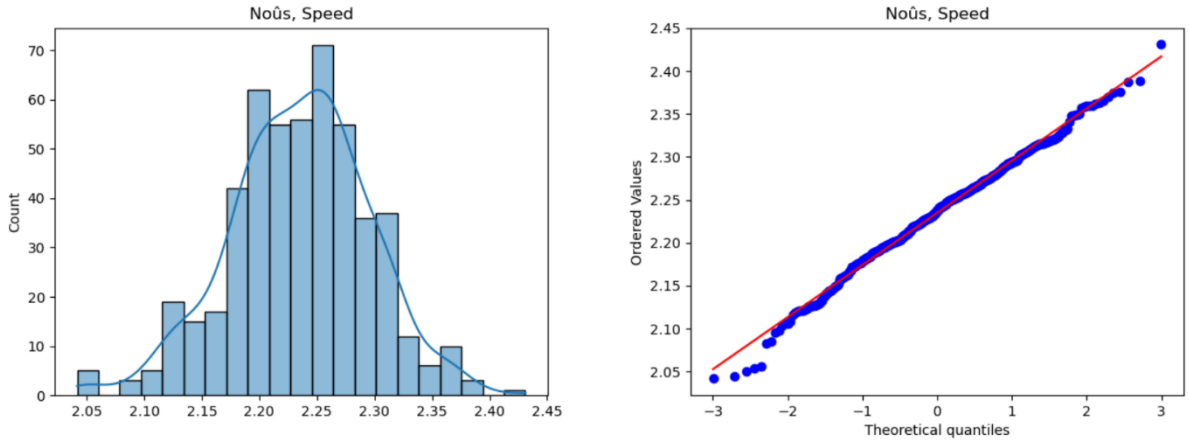

Figure 2B: *Histogram and Q-Q plot of speed in Noûs.*

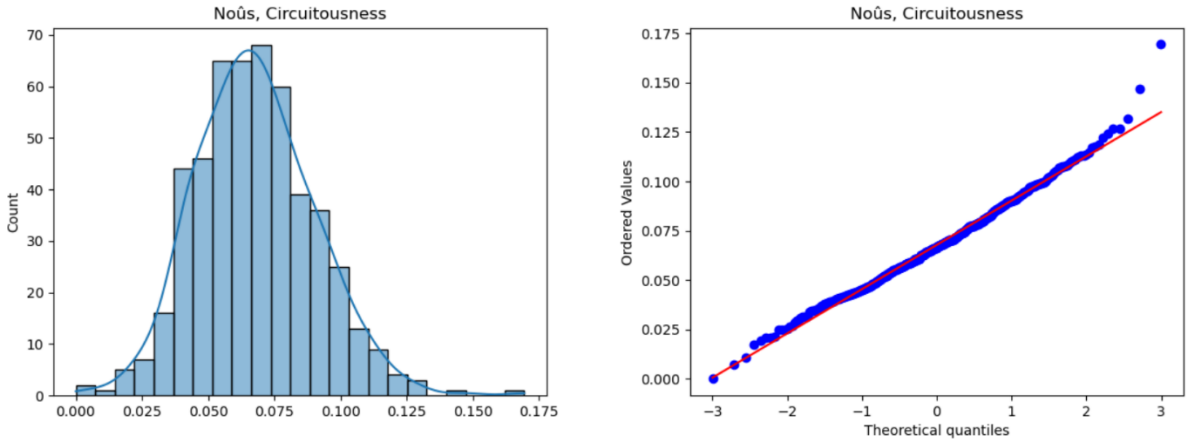

Figure 2C: *Histogram and Q-Q plot of circuitousness in Noûs.*

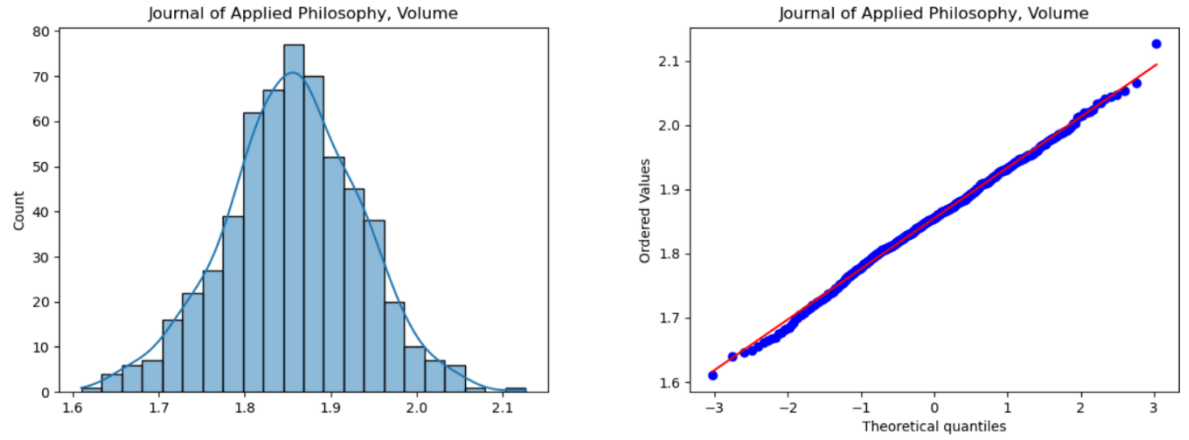

Figure 3A: *Histogram and Q-Q plot of volume in JAP.*

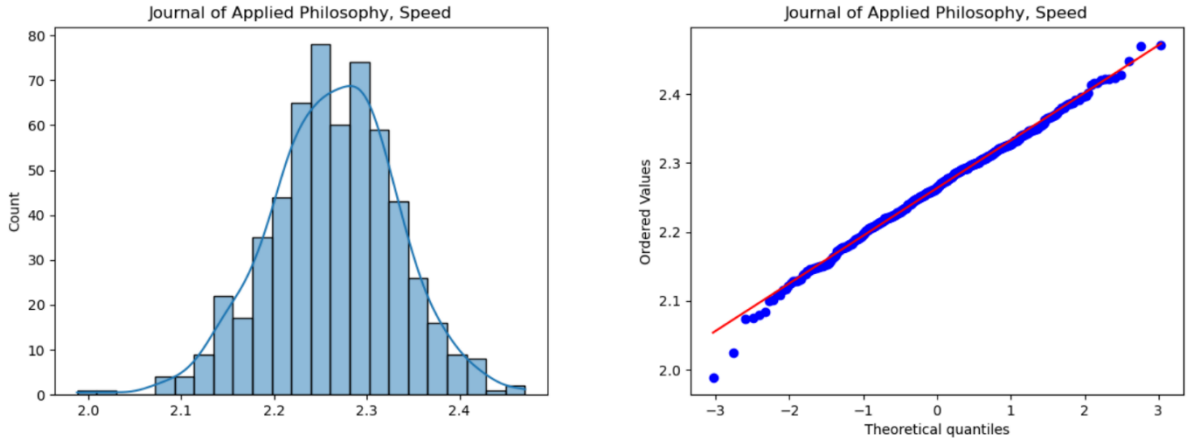

Figure 3B: *Histogram and Q-Q plot of speed in JAP.*

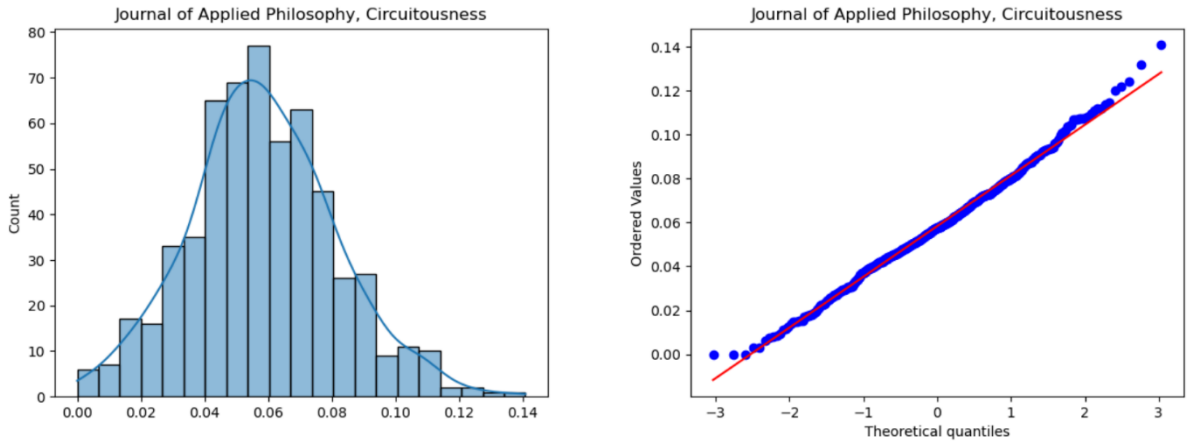

Figure 3C: *Hisogram and Q-Q plot of circuitousness in JAP.*

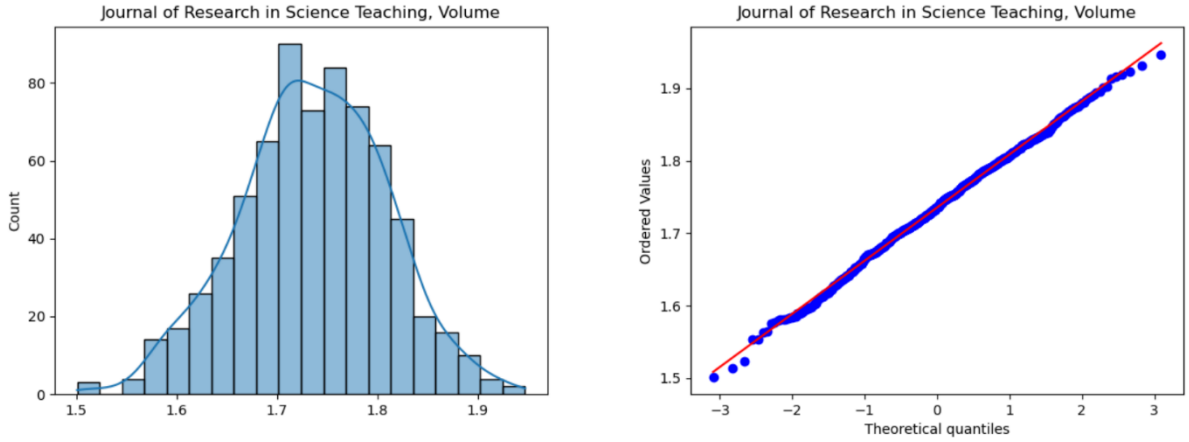

Figure 4A: *Histogram and Q-Q plot of volume in JRST.*

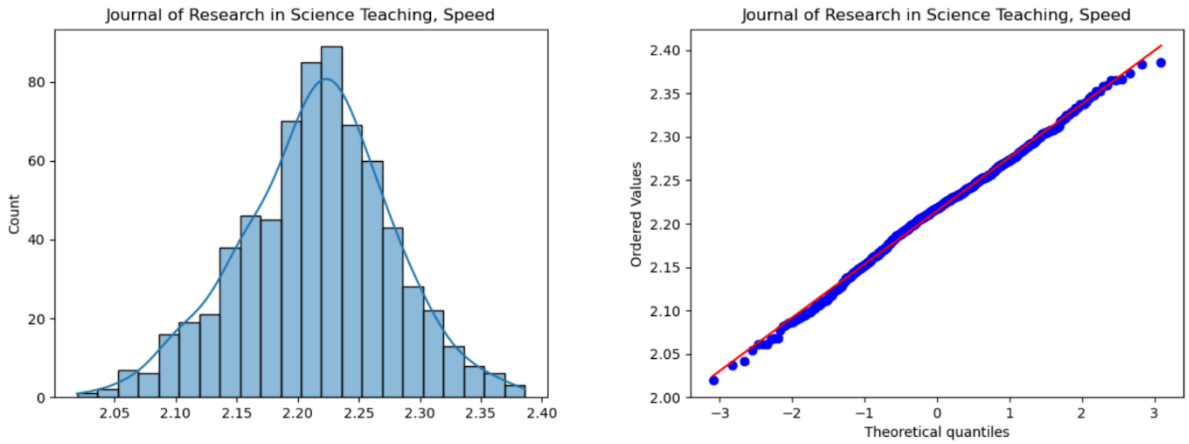

Figure 4B: *Histogram and Q-Q plot of speed in JRST.*

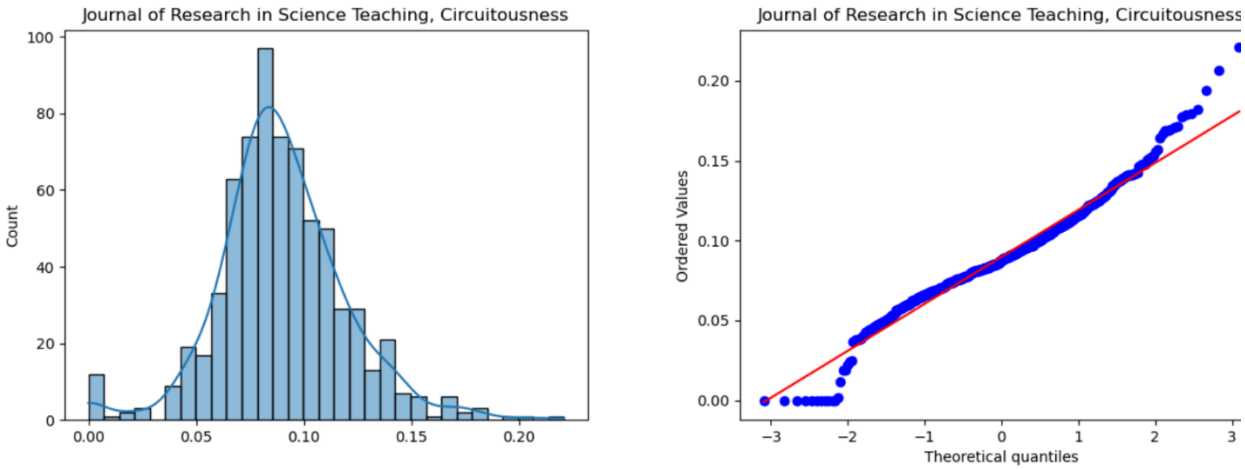

Figure 4C: *Histogram and Q-Q plot of Circuitousness in JRST.*

Test of equality of variances

	Volume	Speed	Circuitousness
Noûs - JAP	0,307	0,00758	0,422
Noûs - JRST	0,562	0,891	(Mann-Whitney U)
JAP- JRST	0,591	0,00592	(Mann-Whitney U)

Table 2. *Levene's test for equality of variances (p-values).*

Another step in determining whether it was appropriate to use t-tests is to look for equality of variances. As can be seen in table 2, for the data in volume comparisons, speed comparison of *Noûs-JRST* and circuitousness comparison of *Noûs-JAPK*, we could assume equality of variances, using Levene's test (p>0,05) and proceed with using t-tests. For speed comparison of *Noûs-JAP* and *JAP-JRST* we could not assume equality of variances for speed and opted for using a Welsh t-test. For comparison between *Noûs-JRST* and *JAP-JRST* for circuitousness we used the Mann-Whitney U test, so we did not check equality of variances.

Comparison between *Noûs* and *Journal of Applied Philosophy*

		Volume	Speed	Circuitousness
T-test	Statistic	-10,9		6,70

	P-value	<0,001		<0,001
Welsh t-test	Statistic		-7,32	
	P-value		<0,001	
Cohens d		-0,664	-0,441	0,407

Table 3. *Comparison of Noûs and JAP (removed outlier from circuitousness Noûs).*

There was a significant difference between *Noûs* and *JAP* in all measured topographic qualities. Of note is that all differences were highly significant <0,001, and that the p-values so low that a Bonferroni correction of the significance level would not have made a difference, given a standard alfa level of 0.05.

There was a significant difference in volume between *Noûs* (M=1.80, SD= 0.0746) and *JAP* (M=1.85, SD=0.0786), t(1086)=-10.9, p<0,001. The effect size, as measured by Cohen's d was d=-0.664, indicating a medium effect. On average *JAP* had bigger volumes than *Noûs*. There are, however, not dramatic differences in volume between journals, given that all measured volumes for both journals are measured in the range of 1.6-2.1. There were also significant differences in speed between *Noûs* (M=2.23, SD=0.0607) and *JAP* (M=2.26, SD=0.0690). We did a Welsh t-test and found t(1086)=-7.32, p<0.001. The effect size was d=-0.441. On average, *JAP* had higher speed as well, but the effect was smaller than the difference between volumes. Of note is that we chose to do a t-test and assumed a normal distribution for *Noûs*, even though the Shapiro Wilk p-value is quite close to the significance level, and there are some outliers in the Q-Q plot for speed. To compare circuitousness between the journals, we decided to remove one outlier from the *Noûs* dataset (this outlier is present in the other comparisons) so that we could do a normal t-test. The results were *Noûs* (M=0.0678, SD=0.0225) and *JAP* (M=0.0583, SD=0.0230), t(1085)=6.78, p<0.001 and the effect size was d=0.412. *Noûs* texts were on average slightly more circuitous than *JAP* texts, which we can hypothesize has something to do with *Noûs* having a more analytic style with logical notation and repetition of the same arguments with slight changes for example.

Comparison between *Journal of Applied Philosophy* and *Journal of Research in Science Teaching*

		Volume	Speed	Circuitousness
T-test	Statistic	27.1		
	P-value	<0.001		
Welsh t-test	Statistic		13.0	
	P-value		<0.001	
Mann-Whitney U	U-statistic			80 741
	P-value			<0.001
Cohen's d		1.52	0.737	
Rank-biserial correlation				-0.635

Table 4. *Comparison of JAP and JRST.*

When comparing JAP and *JRST*, we also found significant differences across all topographic qualities measured. There was a significant difference in volume between *JAP* (M=1.85 SD=0.0786) and *JRST* (M=1.74 SD=0.0733); t(1273)=28.0, p<0.001. Effect size, as measured by Cohen's d, was 1.6. This is one of the most robust and large differences we have observed between journals, and this should be kept in mind for the discussion. There was also a significant difference in speed, which we tested using the Welsh t-test, as we could not assume equality of variances for speed between *JAP* and *JRST*. *JAP* (M=2.26, SD=0.0690) and *JRST* (M=2.22, SD=0.0615); t(1273)=13.2, p<0.001. Effect size, as measured by Cohen's d, was d=0.737. Speed in *JAP* articles were generally higher in *JAP* than in *JRST*. When comparing circuitousness, we did not have a normal distribution in *JRST* due to quite a few outliers. We therefore opted to use the non-parametric equivalent of the t-test, the Mann-Whitney U test. The result of this test indicated significant differences between *JAP* (n=578) and *JRST* (n=697) in circuitousness, with U= 80 741, p<0.001. Effect size, measured by rank-biserial correlation was r_b=-0.635, indicating that *JRST* in general had higher circuitousness than *JAP*.

Comparison between *Noûs* and *Journal of Research in Science Teaching*

		Volume	Speed	Circuitousness
T-test	Statistic	15.0	5.40	
	P-value	<0.001	<0.001	
Welsh t-test	Statistic			
	P-value			
Mann-Whitney U	U-statistic			99160
	P-value			<0.001
Cohen's d		0.926	0.328	
Rank-biserial correlation				-0.481

Table 5. *Comparison of Noûs and JRST.*

Overall, we found significant differences in topographic qualities between *Noûs* and *JRST.* Comparing volume between *Noûs* (M=1.80, SD=0.0746) an *JRST* (M=1.74, SD=0.0733), t(1205)= 15.9, p<0,00, with an effect size measured by Cohen's d (d=0.926). Once again, the largest effect size was in volume. There were also significant differences in speed *Noûs* (M=2.235, SD=0.0607) and *JRST* (M=2.215, SD=0.0614) and t(1205)=5.62, p<0.001, with an effect size of d=0.328. We chose to do a t-test for speed, even though speed both in *Noûs* and *JRST* gave a low Shapiro Wilk p-value as reported in Table 1 and got the lowest effect size of our speed comparisons.

To test for differences in circuitousness between *Noûs* and *JRST,* we did a Mann-Whitney U test as we did not have a normal distribution. We did not remove the outlier from *Noûs* for this analysis, as we had to do a non-parametric test anyway. The result indicated significant differences between *Noûs* (n=510) and *JRST* (n=697) in circuitousness, with U=99160, p<0.001. A rank-biserial correlation of r_b=-0.481 indicated that *JRST* in general had higher circuitousness than *Noûs*.

V. Discussion

While we know that our log transformed data, as presented in the plots earlier, is in the range of 1.50-2.13 for volume, 1.99-2.47 for speed and 0-0.221 for circuitousness, we do not know the span of these values among academic articles in general. Thus, it is difficult to interpret how large these differences are outside the context of our project, and a control comparison of similarly ranked journals from the same discipline would be necessary to draw any conclusions.

Do our findings support our hypotheses?

Our results support our hypothesis (H1) that there are statistically significant differences in topographic semantic properties between journals of different academic disciplines, but they say nothing about whether this difference is due to the disciplines. Interestingly, the difference in volume across disciplines was larger than the difference across journal rankings – the largest effect sizes among all our findings were in fact for comparisons between volume of *Noûs* and *JSRT(*d=0.93X*)*, and *JAP* and *JSRT* (d=1.52), showing larger volumes in philosophical articles than in science education articles, as one might have expected. The findings regarding volume *might* support our intuition that the topographic differences between academic disciplines are generally greater than differences within some discipline.

Still, while our tests for differences in speed and circuitousness also confirm a statistically significant difference between academic journals, we did not find that differences between journals of different disciplines was systematically larger than differences between journals of the same discipline. Furthermore, issues relating to journal selection together with the small sample size present a major problem when trying to state any general results, as will become clear in the following section on limitations.

Our second hypothesis (H2) was that we predict differences in topographic properties between journals of the same discipline, but with different rankings, and indeed we found statistically significant differences. Earlier research indicates that high volume, low speed and high circuitousness correlates positively with citations. *Noûs,* which is very highly rated, indeed has both lower speed and higher circuitousness than *JAP,* but has *lower* volume.

Still, we would give very little weight to this result, since all differences could be entirely due to the vastly different styles of the respective sub-disciplines of the journals. When reading

some sample articles from each journal, the results relating to volume and circuitousness seem rather intuitive (and should likely not be attributed to a difference in quality): A general article in *Noûs* is confined to a specific question of logic (small volume) giving multiple similar examples (high circuitousness), while articles in JAP often touch on grander philosophical issues and follow more of a narrative/linear structure. While this is a major weakness in trying to assert the correctness of our hypothesis, it could speak to the strength of topographic analysis, once again indicating that the metrics correspond to human perceptions.

Overall, we get *highly* significant results across all comparisons; in fact, our largest p-value was $2.31 \cdot 10^{-8}$. One reason for the low values could be our relatively big dataset of around 500-700 articles per journal. If the semantic properties of different journals truly originate from different distributions, the p-value will get closer to zero when the number of articles increases. Still, it might indicate some statistical problems, as will be discussed in the following.

Limitations

Statistics

We found statistically significant differences between topographic qualities in all our comparisons, indicating that there are meaningful differences in semantic properties between journals. All the p-values from the t-tests, Welsh t-tests and Mann-Whitney U tests were *very* low, ranging from $4.82 \cdot 10^{-135}$ to $2.31 \cdot 10^{-8}$. P-values as low as the first one mentioned are for all purposes equal to zero and are presumably not values one would normally get from the statistical tests we have used. Since we had limited prior experience in statistics and we had limited time left when we started doing these analyses, one should keep the possibility open that we have made errors in the calculations of these values. Another limitation is that we might have chosen the wrong statistical tests altogether. We did discuss whether it would be more appropriate to use an analysis of variance, as it might be inappropriate to use t-tests when comparing three or more means (Skaik, 2015).

Sample size and journal selection

The generalizability of our findings is limited, due to the small sample size of only three journals. Adding insult to injury, our choice of journals was suboptimal for answering the hypotheses, since our two philosophy journals are from very dissimilar subdisciplines. In short,

while there is clear evidence that different journals give rise to different topographic properties (given that our p-values are roughly accurate), we cannot say whether the differences can be attributed to either discipline or ranking. This is left for future research with greater sample sizes.

It would have been beneficial to be more familiar with the journals and disciplines in question; we have no prior knowledge of any of the journals and have only skimmed a small number of articles in each one. This limits our ability to theorize around the causes of our findings and might have caused us to miss some potential sources of noise in the data which could have been removed in preprocessing.

Indeed, going through a sample of the processed material, we did find some potential sources of error. Logical notation in *Noûs* turned for the most part into gibberish, and there were more tables and figures than we expected in *JRSE*. A minority of these turned into gibberish, and a handful turned into readable, but not meaningful, text. It would perhaps have been better to pick journals with a minimal number of tables, numbers and figures. However, we did a rudimentary test to see if figures (specifically those turned into gibberish by the OCR) significantly changed our topographic metrics in a text from *JRSE* and found that they did not. Thus, we deemed it not worth it to remove this from our material, since it would be time consuming, and we expected our embedding chunk size of 250 words to buffer against the effects of these parts of the texts. If we had embedded in smaller chunks, for example sentence by sentence (25 words), this might have influenced our measurements more, and have warranted further processing or at least a trimming of the data before analysis. One could, for instance, check each word against a dictionary – if it's not there, it might not be of much use to the transformer anyway!

One could argue that the figures in *JRSE*, perhaps in contrast to the examples and logical notation in *Noûs,* are mostly supplementary to the articles and could be scrapped in preprocessing (or, if our rudimentary analysis holds up, be ignored) without too great a loss for the purpose of analysis. Topography can then be applied rather generally to disciplines where *only the main body of text is essential to the content of the article*. Math and physics are examples of disciplines especially problematic. It is unclear to us whether philosophy of logic passes this constraint; if picking again, we would perhaps have avoided *Noûs*.

Future research

The obvious next step is to take many journals of different ranking dating back a couple of years (not too far back, since ratings change with time) to do a linear regression analyzing the correlation between topographic metrics and journal ranking. This can be done relatively easily with the programs in the journal_analysis-folder in our GitHub repository – see the README for more information. As the program for downloading pdfs stands now, one is limited to Wiley journals, but any publisher providing a (functional) API could presumably be made to work without much hassle. Finally, the Wiley API has a limit of 60 articles per ten minutes, making the process quite slow; together with the glacial OCR, the programs take quite a while, but they can be run unsupervised. Embedding is hardly a problem with a GPU.

It would be nice to compare far more academic texts to further explore if there are any significant topographic differences between disciplines. Especially, it would be interesting to choose academic disciplines perceived as very different, such as comparing philosophy to biology. One could even map different disciplines along an axis based on, for example, their semantic speed and check whether the ordering corresponds to our intuitions about which disciplines are similar.

One should also remain open to finding new topographic metrics. Semantic speed is the average distance traversed per chunk; perhaps some acceleration-metric could provide further insight into the semantic movement of the text? In longer texts with a smaller chunk size (e.g. 25 words), using a more local measure of speed, such as taking a separate measure for each 500 words, would give the *development* of the speed throughout the narrative: It would be interesting to take books/movies/etc. with a given narrative arc determined by humans and match this arc to such a speed plot to find the relationship between tension in the narrative and semantic speed. This would expand on the work of Dos Santos & Berger (2022).

Finally, it might be worthwhile to integrate fields such as cognitive psychology into topographic research. In this respect, two central and interesting questions remain to be answered: *i) What can topographic qualities tell us about the producer of a text?* and *ii) What can topographic qualities tell us about how the text will be received and read?* With respect to i), the correlation between creativity and semantic speed, volume and circuitousness should be explored: Do creative minds manage to span a vast volume without sacrificing coherence by resorting to large semantic speeds? Do creative people write more or less circuitously than others? Relating to ii), one could explore readability by analyzing the properties of textbooks at

different levels. Knowing which properties make a textbook successful could give insight into the psychology of learning and prove helpful to authors – perhaps one could make a tool which searches out the parts of the book which has the highest semantic speed (which might be harder to follow) so that the author can pay special attention to these parts in revisions.

References

Beaty, R. E., & Kenett, Y. N. (2023). Associative thinking at the core of creativity. *Trends in cognitive sciences*, *27*(7), 671-683. https://doi.org/10.1016/j.tics.2023.04.004

Berger, J., Kim, D.Y., Meyer, R. (2021). What Makes Content Engaging? How Emotional Dynamics Shape Success, *Journal of Consumer Research*, *48*(2), 235–250. https://doi.org/10.1093/jcr/ucab010

Berger, J., & Toubia, O. (2024). The topography of thought. *PNAS Nexus*, *3*(5), 161-163. https://doi.org/10.1093/pnasnexus/pgae163

Boghrati, R., Berger, J., & Packard, G. (2023). Style, content, and the success of ideas. *Journal of Consumer Psychology*, *33*(4), 688–700. https://doi.org/10.1002/jcpy.1346

Boyd, R. L., & Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, *40*(1), 21-41.

Inzlicht, M., Shenhav, A., & Olivola, C. Y. (2018). The effort paradox: Effort is both costly and valued. *Trends in cognitive sciences*, *22*(4), 337-349.

Dos Santos, H.L., & Berger, J. (2022). The speed of stories: Semantic progression and narrative success. *Journal of Experimental Psychology: General, 151*(8), 1833–1842. https://doi.org/10.1037/xge0001171

Google OR-Tools. (2023). Traveling Salesperson Problem. [Python code]. https://developers.google.com/optimization/routing/tsp

Kent State University Libraries, (2024, July 10). *SPSS Tutorials: Independent samples t-test*. Kent State University Libraries. https://libguides.library.kent.edu/SPSS/IndependentTTest

Kim TK, Park JH. (2019). More about the basic assumptions of t-test: normality and sample size. *Korean Journal of Anesthesiology, 72*(4), 331-335. doi:10.4097/kja.d.18.00292.

Kutuzov, A. (2020). *Distributional word embeddings in modeling diachronic semantic change*. [Doktorgradsavhandling]. Universitetet i Oslo.

Levene's test. (2024, April 3). In *Wikipedia*, https://en.wikipedia.org/w/index.php?title=Levene%27s_test&oldid=1217056073

Mansurova, M. (2024, February 13). Text embeddings: Comprehensive Guide. *Towards Data Science (Medium)*. https://towardsdatascience.com/text-embeddings-comprehensive-guide-afd97fce8fb5

Mikolov, T., Yih, W. T., & Zweig, G. (2013, June). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 746-751).

Moshtagh, N. (2005). Minimum volume enclosing ellipsoid. *Convex optimization*, *111*(January), 1-9.

Odden, B.O.T., Mjaaland, T.J., Kreutzer, F.M., Malthe-Sørenssen, A. (2024). Using Text Embeddings for Deductive Qualitative Research at Scale in Physics Education. *ArXiv.org*. https://doi.org/10.48550/arxiv.2402.18087

Piper, A., & Toubia, O. (2023). A quantitative study of non-linearity in storytelling. *Poetics*, *98*, 101793. https://doi.org/10.1016/j.poetic.2023.101793

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of linguistics*, *20*, 33-53.

Skaik Y. (2015). The bread and butter of statistical analysis "t-test": Uses and misuses. *Pakistan Journal of Medical Sciences, 31*(6), 1558–1559. https://doi.org/10.12669/pjms.316.8984

Shapiro-Wilk test. (2024, April 12). In *Wikipedia,* https://en.wikipedia.org/w/index.php?title=Shapiro%E2%80%93Wilk_test&oldid=1218522924

Tor Ole B Odden, Tyseng, H., Jonas Timmann Mjaaland, Markus Fleten Kreutzer, & Malthe-Sørenssen, A. (2024). Using Text Embeddings for Deductive Qualitative Research at Scale in Physics Education. *arXiv.org*. https://doi.org/10.48550/arxiv.2402.18087

Toubia, O., Berger, J., & Eliashberg, J. (2021). How quantifying the shape of stories predicts their success. *Proceedings of the National Academy of Sciences*, *118*(26). https://doi.org/10.1073/pnas.2011695118