

GisPy: A Tool for Measuring Gist Inference Score in Text

Pedram Hosseini¹ Christopher R. Wolfe² Mona Diab^{1,3} David A. Broniatowski¹

¹The George Washington University ²Miami University ³Meta AI

{phosseini, broniatowski}@gwu.edu, wolfecr@miamioh.edu, mdiab@fb.com

Abstract

Decision making theories such as Fuzzy-Trace Theory (FTT) suggest that individuals tend to rely on gist, or bottom-line meaning, in the text when making decisions. In this work, we delineate the process of developing GisPy, an open-source tool in Python for measuring the Gist Inference Score (GIS) in text. Evaluation of GisPy on documents in three benchmarks from the news and scientific text domains demonstrates that scores generated by our tool significantly distinguish low vs. high gist documents. Our tool is publicly available to use at: <https://github.com/phosseini/GisPy>.

1 Introduction

According to Fuzzy-Trace Theory (FTT) (Reyna, 2008, 2012), when individuals read text, they encode multiple mental representations of the text in parallel in their mind. These mental representations vary along a continuum ranging from 1) *gist* to 2) *verbatim*. While verbatim representations are related to surface-level information, gist represents the bottom-line meaning of the text, given its context. FTT sees the word gist in much the same way as everyday usage, as the essence or main part, the substance or pith of a matter. Gist representations are important to assess because they influence judgments and decision making more than verbatim representations (Reyna, 2021). Knowing gist helps us measure the capability of a document (e.g., news article, social media post, etc.) in creating a clear and actionable mental representation in readers' mind and the degree to which a document can communicate its message.

The majority of existing Natural Language Processing (NLP) tools and models focus on measuring coherence, cohesion, and readability in text (Graesser et al., 2004; Lapata et al., 2005; Lin et al., 2011; Crossley et al., 2016; Liu et al., 2020; Laban et al., 2021; Duari and Bhatnagar, 2021). It is worth mentioning that even though coherence

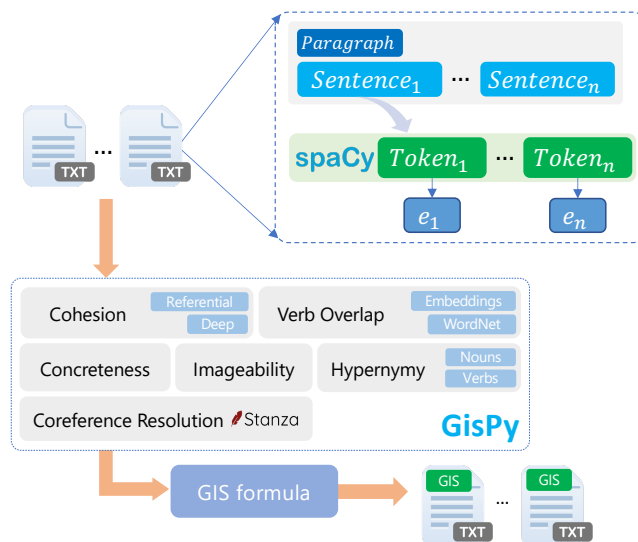


Figure 1: Overview of GisPy pipeline. e_1, \dots, e_n are contextual embedding of tokens in a sentence.

promotes gist extraction, these two are not the same. And gist can be viewed as a mechanism that allows coherence apprehension (Glanemann et al., 2016). To the best of our knowledge, there is no publicly available tool for directly measuring gist in text. Wolfe et al. (2019); Dandignac and Wolfe (2020); Wolfe et al. (2021) are the only studies that introduced a theoretically motivated method to measure Gist Inference Score (GIS) using a subset of Coh-Metrix indices. Coh-Metrix (Graesser et al., 2004) is a tool for producing linguistic and discourse representations of a text including measures of cohesion and readability. Coh-Metrix, even though useful and inspiring, has several limitations. For example, its public version does not allow batch processing of documents, is only available via a web interface, and its cohesion indices focus on local and overall cohesion (Crossley et al., 2016). In this work, inspired by Wolfe et al. (2019) and definition of a subset of indices in Coh-Metrix, we develop a new open-source tool to automatically compute GIS for a collection of text documents.

We leverage the state-of-the-art NLP tools and models such as contextual language model embeddings to further improve the quality of indices in our tool.

Our contributions can be summarized as follows:

- We introduce the first open-source and publicly available tool to measure Gist Inference Score in text.
- We unify and standardize three benchmarks for measuring gist in text and report improved baselines on these benchmarks.
- By leveraging the explainability of indices in our tool, we investigate the role of individual indices in producing GIS for low vs. high gist documents across benchmarks.

2 Methods

In this section, we explain how we implement each of the indices in GisPy and compute GIS. We start by explaining common implementation features among indices followed by specific details about each of them.

2.1 Local vs. Global Indices

We have taken different approaches in implementing indices for which we need to compute the overlap between words or sentences (e.g., semantic similarity). In particular, these indices are computed in two settings: 1) *local* and 2) *global*. In the local setting, we only take into account *consecutive/adjacent* words/sentences whereas in the global setting, we consider *all* pairs not just consecutive ones. Moreover, we compute indices one time by separating the paragraphs in text and another time by disregarding the paragraph boundaries. For clarity, we use postfixes listed in Table 1 for these variations.

Postfix	Explanation
*_l	Local ignoring paragraph boundary
*_a	Global ignoring paragraph boundary
*_lp	Local at paragraph-level
*_ap	Global at paragraph-level

Table 1: Local and global index postfixes

We assume every document is broken into paragraphs $\{P_0, P_1, \dots, P_n\}$, separated by at least one newline character, each with one or more sentences $\{S_{0,0}, S_{0,1}, \dots, S_{i,j}\}$ where each sentence has one

or more tokens $\{t_{0,0,0}, t_{0,0,1}, \dots, t_{i,j,k}\}$. As an example, for a document with two paragraphs each with two and three sentences, respectively:

$$P_0 \rightarrow \{S_{0,0}, S_{0,1}\}$$

$$P_1 \rightarrow \{S_{1,0}, S_{1,1}, S_{1,2}\}$$

Where $S_{i,j}$ is the j th sentence of paragraph i , this is how we compute local and global versions of index X –assuming X measures the similarity among sentences and similarity is computed by \oplus :

$$X_{_l} = \text{mean}(S_{0,0} \oplus S_{0,1}, S_{0,1} \oplus S_{1,0}, S_{1,0} \oplus S_{1,1}, S_{1,1} \oplus S_{1,2})$$

$$X_{_a} = \text{mean}(S_{0,0} \oplus S_{0,1}, S_{0,0} \oplus S_{1,0}, S_{0,0} \oplus S_{1,1}, S_{0,0} \oplus S_{1,2}, S_{0,1} \oplus S_{1,0}, S_{0,1} \oplus S_{1,1}, S_{0,1} \oplus S_{1,2}, S_{1,0} \oplus S_{1,1}, S_{1,0} \oplus S_{1,2}, S_{1,1} \oplus S_{1,2})$$

$$X_{_lp} = \text{mean}(S_{0,0} \oplus S_{0,1}, S_{1,0} \oplus S_{1,1}, S_{1,1} \oplus S_{1,2})$$

$$X_{_ap} = \text{mean}(S_{0,0} \oplus S_{0,1}, S_{1,0} \oplus S_{1,1}, S_{1,0} \oplus S_{1,2}, S_{1,1} \oplus S_{1,2})$$

2.2 GisPy Indices Implementation

Referential Cohesion: This index (PCREFz in Coh-Matrix¹) reflects the overlap of words and ideas across sentences and the entire text. To measure this overlap, we leverage the Sentence Transformers (Reimers and Gurevych, 2019)² to compute the embeddings of all sentences in a document using the all-mpnet-base-v2 model.³ We chose this model since it provides the best quality and has the highest average performance among all the other models introduced by Reimers and Gurevych (2019). Once we computed the embeddings, to measure the overlap across all sentences, we find the cosine similarity between embeddings of every pair of sentences one time at paragraph-level and another time with ignoring the paragraph boundaries. This process

¹To make comparison of our indices with Coh-Matrix easier, we mainly follow Coh-Matrix indices’ names when naming our indices.

²<https://github.com/UKPLab/sentence-transformers>

³Model is available on HuggingFace hub by the name: sentence-transformers/all-mpnet-base-v2



Figure 2: Gist Inference Score (GIS) formula by Wolfe et al. (2019)

results in four indices of referential cohesion including: `PCREF_1`, `PCREF_a`, `PCREF_1p`, `PCREF_ap`.

We additionally implement a new index based on coreference resolution in paragraphs in a document. In particular, using Stanford CoreNLP’s coreference tagger (Manning et al., 2014) through Stanza’s wrapper (Qi et al., 2020), we first find the number of coreference chains (`corefChain`) to the number of sentences in each paragraph. Then we compute the mean value of all paragraphs as our index and call it `CoREF`.

Deep Cohesion: This dimension reflects the degree to which a text contains causal and intentional connectives. To find the incidence of causal connectives, we first created a list of causal markers in text. In particular, using the intra- and inter-sentence causal cues introduced by Luo et al. (2016), we manually generated a list of regular expression patterns and used these patterns to find the causal connectives in a document. Then we computed the total number of causal connectives to the number of sentences in the document as deep cohesion score. We call this index `PCDC`.

Verb Overlap: Based on FTT, abstract rather than concrete verb overlap across a text might help readers construct gist situation models. Wolfe et al. (2019) use two indices from Coh-Metrix to measure the verb overlaps in text including `SMCAUSlsa` and `SMCAUSwn`. Inspired by Coh-Metrix, we make some changes to further improve these indices. In particular, instead of Latent Semantic Analysis (LSA) vectors, we leverage contextualized Pretrained Language Models (PLMs) to get token vector embeddings to later compute the cosine similarity among verbs. Our hypothesis is that since PLMs have encoded contextual knowledge of words in a text, they may be a better choice than LSA for computing the vector representation of verbs in the text. We use spaCy’s⁴ transformer-based pipeline and the `en_core_web_trf` model –which is based on *roberta-base* (Liu et al., 2019)– to compute token vector embeddings and find Part-of-speech (POS)

tags. Different forms of this index in GisPy follow the name pattern `SMCAUSE_*` where *e* stands for language model *embedding*.

To compute the WordNet verb overlap, we first find all synonym sets of verbs in a document in WordNet with POS tag `VERB`. Then for every pair of verbs, we check whether they belong to the same synonym set in WordNet or not. If yes, we assign score 1 to the verb pair, 0, otherwise. Then we compute the average of *Is* to the total number of sentences. Different implementations of this index follow the name pattern `SMCAUSwn_*`.

Word Concreteness and Imageability: To compute word concreteness and imageability (`PCCNC` and `WRDIMGc` in Coh-Metrix) we use two different resources including 1) MRC Psycholinguistic Database Version 2 (Wilson, 1988), a resource that is used by Coh-Metrix and 2) word concreteness and imageability prediction scores using a supervised method introduced by Ljubešić et al. (2018).⁵ In each document, first we search tokens in these two resources based on their POS tags. Then we compute the average concreteness and imageability scores of all tokens in the document as the final scores. This process results in four scores in total named: `PCCNC_mrc`, `WRDIMGc_mrc`, `PCCNC_megahr`, `WRDIMGc_megahr` (two scores for each resource).

Hypernymy Nouns & Verbs: This index shows the specificity of a word in a hierarchy. The idea is that words with more levels of hierarchy are less likely to help readers form gist inference than words with fewer levels (Wolfe et al., 2019). To compute this index, we first list all Nouns and Verbs in a document. Then for each word in the list, we find all synonym sets in the WordNet with the same part of speech tag (Noun or Verb). And, we compute the average *hypernym path* length of all synonym sets of a word. The reason we find all synonym sets of a word instead of only one is that every word can have more than one synonym sets with the same part of speech and there is no way to

⁴<https://spacy.io/>

⁵<https://github.com/clarinsi/megahr-crossling>

know which synonym set has the same meaning as the word in the document. As future work, it would be interesting to see how we can find the synonym set that is closest in meaning to a word in context.

2.3 Computing GIS

Since indices can be on different scales, after computing all indices and before computing GIS which is a linear combination of these indices, we normalize all indices by converting them to z-scores. Then using the formula shown in Figure 2, we compute the final GIS for every document.⁶ Documents with scores greater than zero in the positive direction have higher, and smaller scores than zero in the negative direction have lower levels of gist, respectively.

3 Experiments

To test whether GisPy can correctly group and measure the level of gist in documents, we run our tool on a collection of datasets with known gist levels –low or high. We selected three benchmarks including two introduced by Wolfe et al. (2019) and one introduced by Broniatowski et al. (2016) to test the quality of scores in our tool. We give more detail about these benchmarks in the following subsections. Before running GisPy, we also run Coh-Metrix on each dataset and compute GIS using the original Coh-Metrix indices. Our goal for doing so is to: 1) make sure we have a reliable gold standard that we can compare GisPy scores with and 2) reproduce the results from Wolfe et al. (2019). Once we computed the GIS score using GisPy, to compare low vs. high gist groups, we compare the mean of their GIS scores. Moreover, we run a Student’s t-test with the null hypothesis that there is no difference between the two groups in terms of the level of gist. The goal of running the t-test is to see whether our scores can *significantly* distinguish groups with lower and higher levels of gist.

Also, since for five indices including Referential Cohesion, Verb Overlap based on Embeddings, Verb Overlap using WordNet, Concreteness, and Imageability we have multiple implementations, we compute the final GIS based on all possible combinations of these indices (320 sets of indices

⁶To enable computation of weighted combination of indices or calculating GIS in a different way (e.g., by removing some indices,) we have defined a *weight* variable for each index that can be easily modified and multiplied by its associated index.

for each benchmark). Our goal is to find out what implementation of each index contributes better to distinguishing low vs. high gist documents. In a separate analysis, we also run two robustness tests to ensure our results are not biased by seeing all possible combinations of indices.

3.1 Benchmarks

3.1.1 News Reports vs. Editorials

This benchmark includes 50 documents in two groups including 1) News Reports and 2) Editorials. Based on Wolfe et al. (2019), compared to News Editorials that provide a more coherent narrative, Reports are more focused on facts. As a result, News Reports tend to have a lower level of gist than Editorials.

3.1.2 Journal Article Methods vs. Discussion

This benchmark includes 25 pairs of Methods and Discussion sections (total 50 text documents) from the same peer-reviewed scientific psychology journal articles. Based on Wolfe et al. (2019), while Methods section provides enough detail so that results of an article could be replicated, the Discussion section emphasizes interpretation of results. Hence, Discussion section should produce a higher gist score than Methods. This approach also controls for a number of variables such as author, journal, and topic.

3.1.3 Disneyland Measles Outbreak Data

Disneyland Measles Outbreak Data introduced by Broniatowski et al. (2016) also annotates gist. Documents in this dataset are articles (e.g., news) that are manually annotated by Amazon Mechanical Turk. There are a total of 191 articles with gist annotation among which there are *Gist-Yes*: 147, *Gist-No*: 38, and *unsure*: 6 gist labels. We leave out the *unsure* labels. Since full text of articles in this dataset were not available and each article only had a URL associated with it, we retrieved the full texts using the provided URLs. For those URLs that were no longer available, we used Wayback Machine to find the most recent image of the URL. At the end, we manually cleaned all articles and fixed the paragraph boundaries.

4 Results and Discussion

Results of running GisPy on three benchmarks are shown in Tables 2, 3, and 4. For each benchmark, we listed the top 10 combinations that most

significantly distinguish low vs. high gist documents. As can be seen, for indices that we have paragraph-level vs. non-paragraph-level implementations, in the majority of cases, paragraph-level indices achieve better results. We do not necessarily observe a strong difference between local vs. global implementations. Also, for concreteness and imageability indices, almost all the time we see better performance when we use *megahr* scores by Ljubešić et al. (2018). We leveraged *megahr* as a replacement for MRC that was originally used by Coh-Metrix.

Comparisons of individuals indices for low vs. high gist documents from the best combination on each benchmark are shown in Figures 3, 4, and 5.

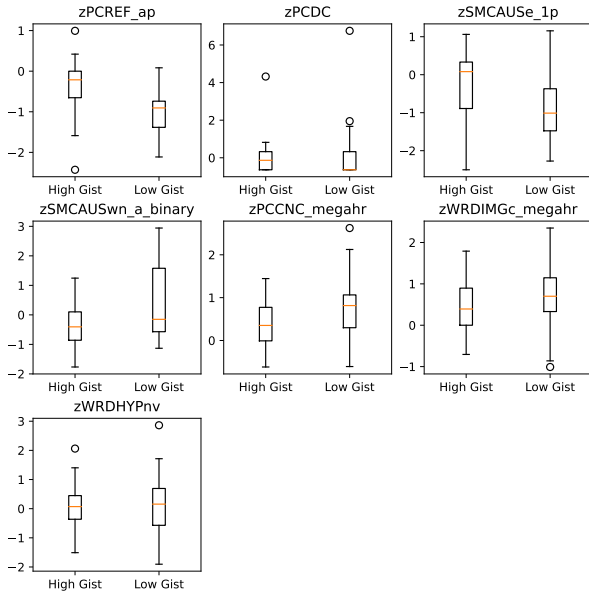


Figure 3: Indices of best GIS on *Reports (Low Gist)* vs. *Editorials (High Gist)*. All values are z-scores.

In Table 5, we also report the comparison of our best results on each benchmark with two other implementations including 1) GIS computed based on the indices of Coh-Metrix, and 2) GIS reported by Wolfe et al. (2019) (For Disney, since we are the first to create a gist benchmark and report a baseline on this dataset, there are no other baselines). As can be seen, on the Reports vs. Editorials and Methods vs. Discussion we achieved performance on par with and slightly better than Coh-Metrix. And we achieved a significantly better distinguishment of low vs. high gist documents than what was reported by Wolfe et al. (2019). And on Disney, GisPy significantly outperformed Coh-Metrix. These results show that we not only could replicate GIS indices, but in contrast to Coh-Metrix, we im-

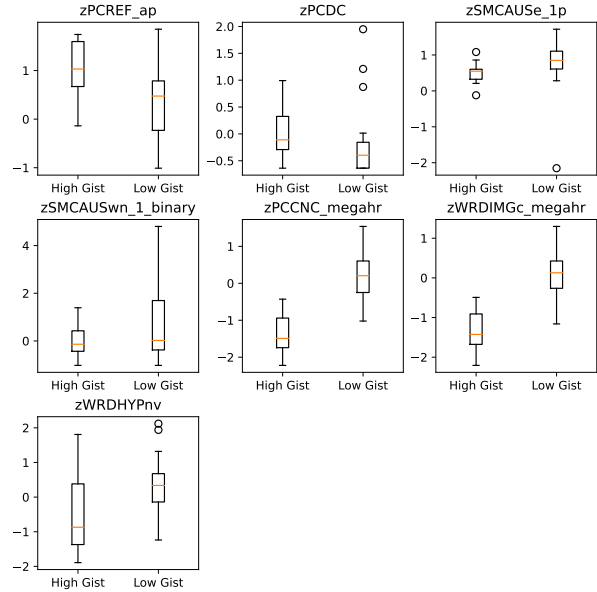


Figure 4: Indices of best GIS on *Methods (Low Gist)* vs. *Discussion (High Gist)*. All values are z-scores.

proved indices in a fully open and transparent way. We hope this implementation transparency helps further improvement of these indices.

4.1 Testing Robustness

We did further testing to see whether our results are robust and generalize across all three benchmarks from the news and scientific text genres.

Test 1: First, out of all combinations of indices, we separated those that significantly distinguished low and high gist groups in each benchmark resulting in 38, 281, 110 combinations for Report vs. Editorials, Methods vs. Discussion, and Disney benchmarks, respectively. We noticed that *all* combinations that are statistically significant in terms of t-test in Reports vs. Editorials benchmark are also statistically significant in the other two benchmarks. In other words, there are 38 different combinations of indices that significantly distinguish low and high gist documents in *all* benchmarks. This confirms the robustness of indices implementation and their generalization across the three benchmarks.

Test 2: Second, we ran an extra experiment to ensure our best GIS scores on each benchmark are also robust when we do not know all possible combinations of indices to pick the best one. In particular, for each benchmark, using three different random seeds, we randomly split texts into a train and a test set each with balanced number of low and high gist documents. Then we computed GIS for documents in the train set and chose the best

Indices Combination					Low Gist	High Gist	Distance	t-statistic	p-value
PCREF	SMCAUSE	SMCAUSwn	PCCNC	WRDIMGc					
ap	1p	a	megahr	megahr	-3.842	-1.292	2.551	3.643	$* 7 \times 10^{-4}$
1p	1p	a	megahr	megahr	-3.833	-1.365	2.467	3.535	$* 9 \times 10^{-4}$
ap	1p	a	megahr	mrc	-3.850	-1.567	2.283	3.265	$* 2 \times 10^{-3}$
1p	1p	a	megahr	mrc	-3.840	-1.640	2.200	3.152	$* 3 \times 10^{-3}$
ap	1	a	megahr	megahr	-3.018	-0.830	2.189	3.216	$* 2 \times 10^{-3}$
ap	1p	a	mrc	megahr	-3.817	-1.662	2.155	2.967	$* 5 \times 10^{-3}$
1p	1	a	megahr	megahr	-3.009	-0.903	2.106	3.088	$* 3 \times 10^{-3}$
1p	1p	a	mrc	megahr	-3.807	-1.736	2.072	2.857	$* 6 \times 10^{-3}$
ap	1	a	megahr	mrc	-3.026	-1.104	1.921	2.792	$* 8 \times 10^{-3}$
ap	1p	a	mrc	mrc	-3.824	-1.937	1.887	2.429	$* 2 \times 10^{-2}$

Table 2: Top 10 GIS scores computed for **Reports** (*Low Gist*) vs. **Editorials** (*High Gist*). **ap**: all pairs at paragraph-level, **1p**: only consecutive/adjacent pairs at paragraph-level, **a**: all pairs in entire document, **1**: only consecutive/adjacent pairs in entire document. * significant p-value ($p \leq 0.05$)

Indices Combination					Low Gist	High Gist	Distance	t-statistic	p-value
PCREF	SMCAUSE	SMCAUSwn	PCCNC	WRDIMGc					
ap	1p	1	megahr	megahr	-0.282	4.730	5.012	7.188	$* 4 \times 10^{-9}$
ap	ap	1	megahr	megahr	-0.576	4.414	4.991	6.528	$* 4 \times 10^{-8}$
ap	1p	1p	megahr	megahr	-0.180	4.701	4.881	7.829	$* 4 \times 10^{-10}$
a	1p	1	megahr	megahr	-1.203	3.678	4.881	7.424	$* 2 \times 10^{-9}$
ap	ap	1p	megahr	megahr	-0.474	4.386	4.860	6.883	$* 10^{-8}$
a	ap	1	megahr	megahr	-1.497	3.362	4.860	6.460	$* 5 \times 10^{-8}$
1p	1p	1	megahr	megahr	-0.159	4.670	4.829	6.989	$* 8 \times 10^{-9}$
1p	ap	1	megahr	megahr	-0.453	4.355	4.808	6.328	$* 8 \times 10^{-8}$
a	1p	1p	megahr	megahr	-1.101	3.649	4.750	7.820	$* 4 \times 10^{-10}$
a	ap	1p	megahr	megahr	-1.395	3.333	4.729	6.594	$* 3 \times 10^{-8}$

Table 3: Top 10 GIS scores computed for **Methods** (*Low Gist*) vs. **Discussion** (*High Gist*). **ap**: all pairs at paragraph-level, **1p**: only consecutive/adjacent pairs at paragraph-level, **a**: all pairs in entire document, **1**: only consecutive/adjacent pairs in entire document. * significant p-value ($p \leq 0.05$)

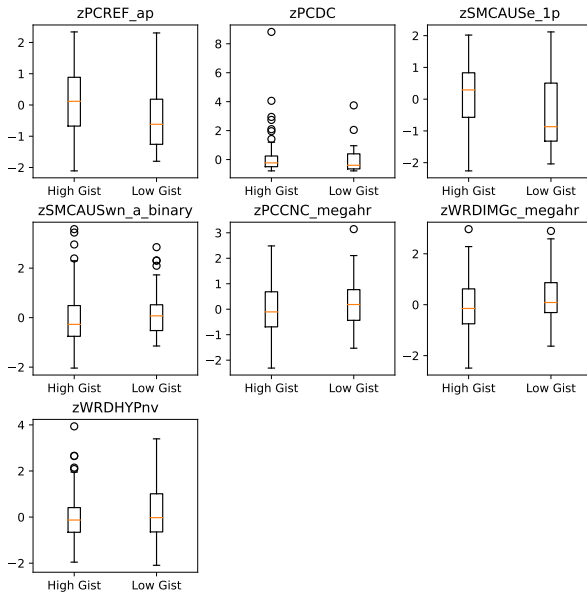


Figure 5: Indices of best GIS on *Disney Gist=No* (*Low Gist*) vs. *Gist=Yes* (*High Gist*). All values are z-scores.

combination of indices that achieved the largest GIS distance between low and high gist groups. Then using that combination we computed GIS for documents in the test set. Results are reported in Tables 6. As can be seen in the table, in all three benchmarks, the best indices combination on the train set also significantly distinguished the low and high gist documents in the test set. This further confirms that our GisPy indices are also robust when tested on unseen documents.

We also analyzed the individual indices from best combinations on the train set in robustness test 2. These combinations are listed in Table 7. We noticed that for *PCREF* and *SMCAUSE*, in %83 of the experiments, *zPCREF_ap* and *zSMCAUSE_1p* are part of the best combination. Also, for these two indices, in %89 of the times we obtained a better result using paragraph-level implementations than when we ignore paragraph boundaries. In other words, we obtain a better result by computing referential cohesion and semantic verb overlap using word embeddings at paragraph-level most

Indices Combination					Gist=No	Gist=Yes	Distance	t-statistic	p-value
PCREF	SMCAUS _e	SMCAUS _{wn}	PCCNC	WRDIMG _c					
ap	1p	a	megahr	megahr	-1.921	0.497	2.418	3.440	$* 7 \times 10^{-4}$
1p	1p	a	megahr	megahr	-1.911	0.494	2.405	3.410	$* 8 \times 10^{-4}$
ap	1p	1	megahr	megahr	-1.729	0.447	2.176	3.158	$* 2 \times 10^{-3}$
1p	1p	1	megahr	megahr	-1.719	0.444	2.164	3.131	$* 2 \times 10^{-3}$
ap	1p	a	mrc	megahr	-1.676	0.433	2.109	3.416	$* 8 \times 10^{-4}$
1p	1p	a	mrc	megahr	-1.666	0.431	2.097	3.384	$* 9 \times 10^{-4}$
ap	1p	a	megahr	mrc	-1.652	0.427	2.079	3.415	$* 8 \times 10^{-4}$
1p	1p	a	megahr	mrc	-1.642	0.424	2.066	3.381	$* 9 \times 10^{-4}$
ap	1p	ap	megahr	megahr	-1.557	0.403	1.960	3.001	$* 3 \times 10^{-3}$
ap	ap	a	megahr	megahr	-1.554	0.402	1.956	2.950	$* 4 \times 10^{-3}$

Table 4: Top 10 GIS scores computed by GisPy for **Gist=No** vs. **Gist=Yes** articles in the Disney dataset. **ap**: all pairs at paragraph-level, **1p**: only consecutive/adjacent pairs at paragraph-level, **a**: all pairs in entire document, **1**: only consecutive/adjacent pairs in entire document. * significant p-value ($p \leq 0.05$)

Benchmark	Approach	Low Gist	High Gist	Distance	t-statistic	p-value
Reports vs. Editorials	GisPy	-3.842	-1.292	2.551	3.643	$* 7 \times 10^{-4}$
	Coh-Metrix	-4.148	-1.613	2.535	3.826	$* 3 \times 10^{-4}$
	(Wolfe et al., 2019)	-0.620	-0.252	0.368	-	-
Methods vs. Discussion	GisPy	-0.282	4.730	5.012	7.188	$* 3 \times 10^{-9}$
	Coh-Metrix	-2.077	2.933	5.010	6.331	$* 7 \times 10^{-8}$
	(Wolfe et al., 2019)	-0.297	0.45	0.747	-	-
Disney	GisPy	-1.921	0.497	2.418	3.440	$* 7 \times 10^{-4}$
	Coh-Metrix	-1.148	-0.151	0.998	1.878	6×10^{-2}

Table 5: Comparison of GIS scores generated by GisPy vs. other methods for all benchmarks. * significant p-value ($p \leq 0.05$)

of the time. For *PCCNC* and *WRDIMG_c*, in all experiments with the exception of only one case only for *WRDIMG_c*, scores computed by *megahr* achieved the best performance. And finally for *SMCAUS_{wn}*, in %67 of the experiments, the *_a implementation resulted in a better distinguishment between low and high gist documents than the local (*_l) implementation. Also, in only two experiments the paragraph-level *SMCAUS_{wn}* worked better than its non-paragraph-level implementation. Additionally, we dug a little deeper to understand why there is a difference between local vs. global *SMCAUS_{wn}* across benchmarks. We noticed that the local indices only perform better in the Methods vs. Discussion dataset. So we took a closer look to understand why this is the case. Interestingly, when we computed the ratio of the number of sentences to the number of paragraphs for all benchmarks, we observed that ratios for Reports vs. Editorials and Disney benchmarks, where global indices achieve a better performance, are **1.89** and **2.04**, respectively. And for Methods vs. Discussion where local indices perform better, the ratio is **6.48**

which is significantly greater than the other two benchmarks. This may suggest that the density of paragraphs in terms of the number of sentences in each paragraph is one factor we need to keep in mind when selecting what implementation we want to choose for a benchmark. It would be interesting to run this analysis on more documents to see how our observation generalizes across different datasets.

5 Next Steps and Future Work

Despite achieving significant improvements and solid results from robustness tests on three benchmarks from two domains, there is still great room to further improve the quality of GisPy indices. In this section, we list challenges in the current implementation of GisPy and explain what we think can be a proper next step and direction in addressing them. We hope these insights inspire the community to keep working on this exciting line of research.

We did our best to bring three different benchmarks for measuring gist inference score to life by aggregating, standardizing, and making them very

Benchmark	Train			Test		
	Low Gist	High Gist	p-value	Low Gist	High Gist	p-value
Reports vs. Editorials	-3.770	-1.131	* 2×10^{-2}	-3.663	-1.413	* 5×10^{-2}
Methods vs. Discussion	-0.634	4.538	* 7×10^{-5}	-0.342	4.346	* 2×10^{-4}
Disney Gist=Yes vs. Gist=No	-1.926	0.496	* 4×10^{-2}	-1.910	0.493	* 4×10^{-2}

Table 6: GisPy GIS scores for train and test sets on all benchmarks. * significant p-value ($p \leq 0.05$)

Benchmark (S/P Ratio)	PCREF	SMCAUSE	SMCAUSwn	PCCNC	WRDIMGc
Reports vs. Editorials (1.89)	ap	1	a	megahr	megahr
	ap	1p	a	megahr	megahr
	ap	1p	a	megahr	megahr
	ap	1p	a	megahr	megahr
	ap	1p	a	megahr	mrc
	ap	1p	a	megahr	megahr
Methods vs. Discussion (6.48)	ap	1p	1	megahr	megahr
	ap	ap	1	megahr	megahr
	ap	1p	1	megahr	megahr
	a	1p	1p	megahr	megahr
	ap	ap	1p	megahr	megahr
	ap	1p	1	megahr	megahr
Disney (2.04)	ap	1p	a	megahr	megahr
	1p	1p	a	megahr	megahr
	ap	1p	a	megahr	megahr
	ap	1p	a	megahr	megahr
	ap	1p	a	megahr	megahr
	1p	1p	a	megahr	megahr

Table 7: Best combinations in robustness Test 2 on the train set for all experiments separated by benchmark.

easy to use. However, since measuring gist is a relatively newer and less investigated topic compared to readability, coherence, or cohesion, there is still a need for having higher quality benchmarks from different domains. The benchmarks we have tested our tool with are mainly from the news and scientific text domains. It would be interesting to see how our tool can be tuned on not only more documents from these domains but also other genres of text.

Also, our PCDC index, even though based on strong causal connective markers, mainly covers the explicit causal relations while not all causal relations are expressed explicitly in text. It would be interesting to think how we can enhance the quality of this index by also including implicit relations and disambiguating causal connectives that can also be non-causal (e.g., temporal markers such as *since* or *after*) or leveraging discourse parsers such as DiscoPy (Knaebel, 2021).

We initially hypothesized that utilizing coreference resolution chains (CoREF index) may also help us improve the referential cohesion index. By

looking at the most significant combinations of indices in each benchmark, we noticed that CoREF appeared in 0/38, 53/281, 1/110 combinations for Report vs. Editorials, Methods vs. Discussion, and Disney benchmarks, respectively. As a follow-up, it would be interesting to see how coreference resolution can be leveraged in a different way – individually or in combination with other implementations of referential cohesion– to further improve this index.

6 Conclusion

In this work, we introduced GisPy, a new open-source tool for measuring Gist Inference Score (GIS) in text. Evaluation of GisPy and robustness tests on three different benchmarks of low and high gist documents demonstrate that our tool can significantly distinguish documents with different levels of gist. We hope making GisPy publicly available inspires the research community to further improve indices of measuring gist inference in text.

References

- David A Broniatowski, Karen M Hilyard, and Mark Dredze. 2016. Effective vaccine communication during the disneyland measles outbreak. *Vaccine*, 34(28):3225–3228.
- Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2016. The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, 48(4):1227–1237.
- Mitchell Dandignac and Christopher R Wolfe. 2020. Gist inference scores predict gist memory for authentic patient education cancer texts. *Patient Education and Counseling*, 103(8):1562–1567.
- Swagata Duari and Vasudha Bhatnagar. 2021. Ffcd: A fast-and-frugal coherence detection method. *IEEE Access*.
- Reinhild Glanemann, Pienie Zwitserlood, Jens Bölte, and Christian Dobel. 2016. Rapid apprehension of the coherence of action scenes. *Psychonomic bulletin & review*, 23(5):1566–1575.
- Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-matrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.
- René Knaebel. 2021. *discopy: A neural system for shallow discourse parsing*. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 128–133, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.
- Philippe Laban, Luke Dai, Lucas Bandarkar, and Marti A Hearst. 2021. Can transformer models measure coherence in text: Re-thinking the shuffle test. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1058–1064.
- Mirella Lapata, Regina Barzilay, et al. 2005. Automatic evaluation of text coherence: Models and representations. In *IJCAI*, volume 5, pages 1085–1090. Cite-seer.
- Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. *Automatically evaluating text coherence using discourse relations*. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 997–1006, Portland, Oregon, USA. Association for Computational Linguistics.
- Sennan Liu, Shuang Zeng, and Sujian Li. 2020. Evaluating text coherence at sentence and paragraph levels. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1695–1703.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Nikola Ljubešić, Darja Fišer, and Anita Peti-Stantić. 2018. *Predicting concreteness and imageability of words within and across languages via word embeddings*. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 217–222, Melbourne, Australia. Association for Computational Linguistics.
- Zhiyi Luo, Yuchen Sha, Kenny Q Zhu, Seung-won Hwang, and Zhongyuan Wang. 2016. Commonsense causal reasoning between short texts. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*.
- Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.
- Nils Reimers and Iryna Gurevych. 2019. *Sentence-bert: Sentence embeddings using siamese bert-networks*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Valerie F Reyna. 2008. A theory of medical decision making and health: fuzzy trace theory. *Medical decision making*, 28(6):850–865.
- Valerie F Reyna. 2012. A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision making*.
- Valerie F Reyna. 2021. A scientific theory of gist communication and misinformation resistance, with implications for health, education, and policy. *Proceedings of the National Academy of Sciences*, 118(15):e1912441117.
- Michael Wilson. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers*, 20(1):6–10.
- Christopher R Wolfe, Mitchell Dandignac, and Valerie F Reyna. 2019. A theoretically motivated method for automatically evaluating texts for gist inferences. *Behavior research methods*, 51(6):2419–2437.
- Christopher R Wolfe, Mitchell Dandignac, Cynthia Wang, and Savannah R Lowe. 2021. Gist inference scores predict cloze comprehension “in your own words” for native, not esl readers. *Health Communication*, pages 1–8.