TOPTEXT. Topological Text Analysis via Embedded Flow Graphs

Anonymous ACL submission

Clemens Berger and Timon Boehm, May 2026

Abstract

We propose a new model for topological text analysis based on so-called flow graphs (TOPTEXT). Our model refines existing approaches insofar as its architecture has a built-in distinction between syntax (horizontal structure) and semantics (vertical structure). The horizontal structure is generated by the individual sentences of the text while the vertical structure is generated by lexical recurrence or semantic fields. The whole flow graph embeds in a two-dimensional oriented surface whose geometry is richer than the one-dimensional models traditionally used in distributional semantics. The combinatorics of the flow graph allow for several transparent and interpretable algorithms of topic modeling and focus detection. The windows of distributional semantics are replaced here with so-called boundary cycles, well known in mathematics and not depending on any choice of parameter. The semantic fields can be generated manually or via existing word embeddings. Our model is well suited for analyzing literary texts (e. g. poems), scientific texts or philosophical texts.

1 Introduction

With the digitalisation and use of quantitative methods in literature, linguistics and philosophy, several methods have been developed to identify 'themes' in a text through keywords. Particular interest has been given to the so-called 'topic modeling'. The output is usually a set of keywords intended to describe the main topics of the text. Although existing models, see e. g. Deerwester et al. (1990); Blei (2012), have been successful in applications, it seems appropriate to reconsider the basic assumptions of such approaches, first and foremost, the underlying conception of language.

The methods used in NLP models rely, in one form or another, on the hypothesis that the meaning of a word is determined by its frequency distribution. This goes back, amongst others, to Harris (1951), and roughly claims that two words have similar meaning if they occur in similar contexts. A word is often represented as a vector, whose dimension corresponds to the size of context, and whose coordinates count the number of occurrences in this context (Turney and Pantel, 2010).

One of the basic motivations for TOPTEXT was to propose a model for text analysis which remains as long as possible independent from any statistical hypothesis. For us, text analysis includes at least a determination of the main themes/rhemes of a text: a theme is what is being discussed, what is at issue, or the context in which information becomes comprehensible; a rheme is what is said about the theme, what is relevant and in focus. A text is an explication of a theme (or several themes) by means of rhemes. This functional distinction has been discussed by the Prague School, see Benes (1973); Danes (1974); Mathesius (1975) as well as in the Anglo-American sphere Halliday (1967), where it appears under the designation topic-focus or topic-comment, e. g. Chomsky (1990).

In many of the existing text analysis methods, crucial decision-making processes are not really transparent or interpretable and take place within a black box. It is often impossible to understand how the system works and what steps it performs (transparency) or how it produces certain results (interpretability) (Serrano and Smith, 2019; Lipton, 2018; Vanni and Mayaffre, 2026). At the same time, these results are only reproducible to a limited extent; consequently, stability is not guaranteed and errors occur that are difficult to explain (Arkoudas, 2023).

In general, the 'distributional view' of language is successful when dealing with everyday language, where the level of complexity is relatively low. Literary, philosophical, and scientific texts, on the other hand, are characterised by intertextual and intratextual references, self-referentiality, rhetorical figures etc., where understanding requires a

"grasp of the objective internal process logic of the works and their contextual meaning. [...] It is not the depicted meaning, but the meaning articulated through internal appropriateness, form, organisation and functionality that creates meaningful connections." (our transl.) (Zittel, 2026) The question therefore arises as to whether such structures are better accounted for by other models than by distributional semantics.

The originality of the TOPTEXT model lies in the fact that a text is represented from the outset as a two-dimensional object, in which sentences form the horizontal structure while intersentential word repetitions (or semantic fields) form the vertical structure. Syntax and semantics thus generate a two-dimensional graphical network of words. The usual representation of words as vectors is dispensed with as is the choice of a window width familiar in distributional semantics. Instead, the context of a word is entirely determined by the horizontal and vertical structure of the surrounding graph. The mathematical structure of this graph is what we call a flow graph. It is intended to be shown that the flow graph of a text carries more 'hermeneutic information' than existing models for text analysis.

It is possible to incorporate additional information in form of semantic fields into the flow graph. These semantic fields can be specified manually, or generated automatically. In the latter case, we use word embeddings to generate a word similarity matrix, and some elementary clustering methods to determine the semantic fields. Even in this case, transparency and interpretability remain largely preserved, since the generated semantic fields can be supervised, and corrected if necessary.

The advantages of the TOPTEXT model lie particularly in the analysis of texts where formal aspects are highly relevant to semantics, i. e. in philosophical, scientific, and lyrical texts. We hope that our model puts the digital access to literary, philosophical and academic texts on a new footing. A prototype was presented in Boehm (2023) and for all examples so far analyzed the result has been very satisfactory.

2 Architecture

The flow graph of a text contains, as mentioned before, two classes of edges, the horizontal edges describing the syntactical order of the words in the individual sentences, and the vertical edges representing lexical recurrences or semantic fields. The construction of the flow graph implies that the edges around a term of the text are cyclically ordered. This induces the existence of so-called boundary-cycles, as well as an embedding of the flow graph into a two-dimensional surface. The latter induces so-called homology cycles. Our hypothesis is that boundary cycles contain thematic information, homology cycles rhematic information. The occurrences of words along these paths can then be examined using conventional statistical methods. In this way, we replace linear one-dimensional statistics with a two-dimensional model of semantic circulation, accounting for long-range semantic connections in the text. Meaning is not bound to local proximity but arises from participation in a global recurrence structure.

1. Sentence cycles Sentences possess a natural syntagmatic order, which we model as closed edge paths and hereafter refer to as sentence cycles. Two consecutive words (nodes) as well as the last and the first are connected by an edge (see Fig. 1). The sentence cycle takes account of the intuition that a sentence is ordered and can be read several times. Thus, the TOPTEXT model also incorporates aspects of human cognition and memory. Every term of a text is thus located in a specific sentence cycle.

2. Relational cycles Relationships between terms are also represented by cycles. In the simplest case, all identical words (homonyms) that appear in different sentences are linked by a relational cycle. In more complex cases, words that have similar semantic content are likewise linked by a relational cycle. Here, the cycle structure is generated by the natural order of the sentences inside the text. Word repetitions thus generate a second vertical dimension. The relational cycles take account of the intuition that, when reading a text, we retain word repetitions and similar semantic references and incorporate them into our text interpretation. In principle, intertextual associations, connotations or other world knowledge can also be incorporated into relational cycles.

The word similarity classes are represented by text-specific semantic fields. These can be specified either manually or generated automatically. In the latter case, we accept a possible loss of transparency and interpretability, in order to gain, from a hermeneutic perspective, a preliminary understanding that might contribute to the coherence of the interpretation. In academic texts, lexical recurrence generally suffices to produce a connected

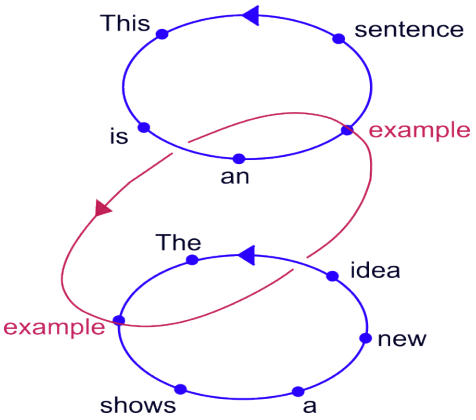

Figure 1: Two sentence cycles (blue) and one relational cycle (red), which links the word "example" appearing in both cycles. Together, all sentence and relational cycles in a text form what we call flow graph. This graph is in general non-planar.

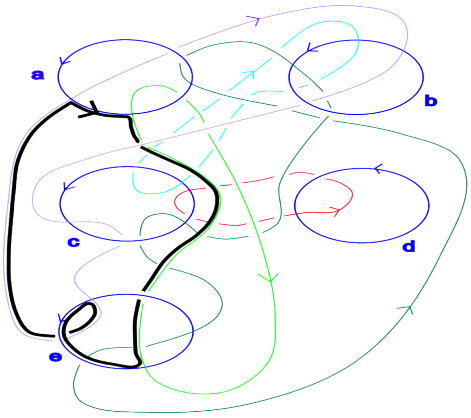

Figure 3: A schematic representation of five sentence cycles (a, b, c, d, e in blue), linked by relational cycles (red and other colours). One of the boundary cycles is shown in black.

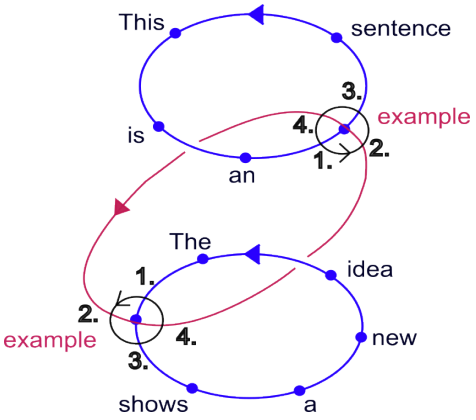

Figure 2: The intersections of sentence and relational cycles are cyclically ordered.

flow graph; in poems, however, this is not the case, and semantic fields provide essential hermeneutic extra-information.

Sentence cycles and relational cycles together form what we call the flow graph of the text (see Fig. 1). This structure ties in with Algirdas Greimas structural semantics: "Language is not a system of signs, but an association of structures of meaning" Greimas (1971), p. 15; cf. also Rastier (2001).

3. Boundary cycle The flow graph is endowed with a specific structure known in mathematics: The intersection points between sentence and relational cycles are cyclically ordered according to the following rule: 1. the sentence edge entering the node, 2. the relational edge entering the node, 3. the sentence edge leaving the node, 4. the relational edge leaving the node.

This additional structure makes it possible to embed the flow graph into a two-dimensional oriented surface (Labourie, 2013), the topology of which provides important clues for the hermeneutics of the text. In particular, edge paths can be formed in the flow graph in which sentence and relational edges are used alternately. Such an edge path is necessarily closed and is called a boundary cycle because it bounds a two-dimensional cell of the aforementioned oriented surface. Every edge of the flow graph is traversed exactly twice, either by two different boundary cycles or by the same boundary cycle. Thus, all terms of the text are incorporated into boundary cycles. The surface is divided into cells, each bounded by a uniquely determined boundary cycle of the flow graph. One can visualise an edge traversed twice as a ribbon. For this reason, the flow graph with cyclically oriented nodes is referred to in mathematics as a ribbon graph.

One of the key features of the TOPTEXT architecture is therefore the embedding of the flow graph into an oriented surface. This represents an innovation compared to other models, which are essentially one-dimensional. The two-dimensional surface contains the text as a graph. Each term corresponds to a node. A term is a word located within a sentence, i. e. a word that carries both an index indicating its sentence membership and an index indicating its position within the sentence.

4. Homology cycles The embedding surface is a compact two-dimensional topological space. It is characterised by a topological invariant, the Euler characteristic $\chi = V - E + F = 2 - 2g$. In this formula, V, E, F denote the number of vertices, edges and faces (equal to the number of boundary

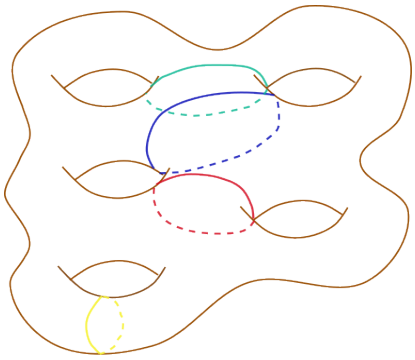

Figure 4: An embedding space of genus $g = 5$. The figure shows a schematic representation of some homology cycles.

cycles); g is the genus of the surface. For a torus, $g = 1$; for a double torus, $g = 2$; for a pretzel, $g = 3$. The surface is further characterised by the so-called first homology group H_1, which is generated by $2g$ homology cycles. Visually, one can imagine that the homology cycles 'entangle' the embedding surface (cf. Fig. 4).

3 Hermeneutics

The first hypothesis of TOPTEXT is that terms contained in the same boundary cycle of a flow graph form a thematic unit of the underlying text. This is prima facie plausible since each boundary cycle surrounds a cell of the surface and all cells together partition the surface in a unique manner. Moreover, the cells can be contracted so that the words within the same boundary cycle can be brought into arbitrary proximity inside the embedding surface. This hypothesis has to be tested empirically by comparing these thematic units with human expectations regarding interpretation.

There are longer and shorter boundary cycles, and the latter can be understood not only as thematic units but also (and often better) as motives.

As second hypothesis we claim that the homology cycles correspond to the rhematic units of the text. This hypothesis can be justified by the fact that homology cycles 'entangle' the embedding surface (cf. Fig. 4) and, in this sense, may represent progressions between certain thematic units. In this paper though, we shall focus on thematic units.

4 Implementation

To ensure efficient processing and to handle longer texts, our model has been implemented in Python. The current code (approx. 800 lines) uses five Python packages: networkx, matplotlib (for flow graphs), sympy (for linear algebra), gensim.models and sklearn.cluster (for word embeddings and semantic clustering).

For a given text, the program computes: flow graph, dual graph, boundary cycles, homology cycles, surface genus, and various correlation matrices representing the word distribution with respect to the boundary cycles. All cycles can be plotted graphically.

Our lemmatisation follows established standards: the terms are replaced with their normal forms resp. lexemes (verbs in infinitive, nouns in nominative singular). Punctuation marks, definite and indefinite articles, prepositions, pronouns and auxiliary verbs are eliminated. Among the conjunctions only the temporal, correlative and causal ones are retained.

The determination of semantic fields is via word embeddings (we use fasttext) and DBSCAN. The algorithm is such that the semantic fields are 'smallest possible' in order to get a connected flow graph. So, if the flow graph is already connected, the semantic fields are singletons, and semantic circulation reduces to lexical recurrence.

In the dual graph of a flow graph, the boundary cycles are represented by nodes and any two nodes are joined by a weighted edge. The weight is a measure of the proximity of the corresponding boundary cycles. The dual graph, weighted in this way, makes it possible to assemble boundary cycles (i. e. thematic units) into themes: these themes correspond to the connected components of the reduced dual graph, where only those edges are kept whose weight exceeds a certain threshold (in practice a trimmed average of all weights).

To assign to a boundary cycle (i. e. a thematic unit) a list of keywords describing it, we propose three methods: The first just extracts the most frequent words within the boundary cycle. The second (resp. third) method involves computing what we call the T-centrality (resp. C-centrality) of a word. The T-centrality of a word is the number of terms in the text corresponding to this word. The C-centrality of a word is the sum of the C-centralities of all terms corresponding to this word, where the C-centrality of a term is the probability that an arbitrary term of the text is connected to the given term via a boundary cycle. For each thematic unit (or theme), the words with highest T-centrality (resp. C-centrality) are extracted. In practice the three methods produce similar results.

Instead of extracting keywords we also propose

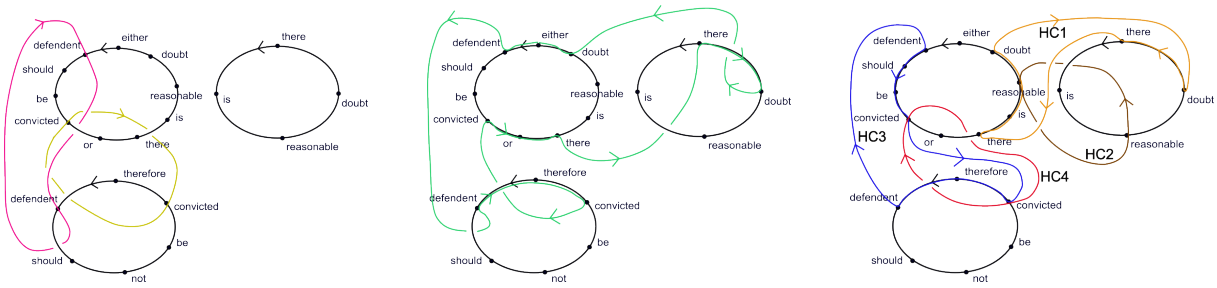

Figure 5: The four types of cycles for a text example consisting of three sentences. For the sake of clarity, only 2 of the 8 relational cycles (left-hand image) and 1 of the 14 boundary cycles (middle image) are shown. The homology cycles are shown in full (right-hand image).

Feynman_orig.txt

We begin with light. When Newton started looking at light, the first thing he found was that white light is a mixture of colors. He separated white light with a prism into various colors, but when he put light of one color - red, for instance - through another prism, he found it could not be separated further. So Newton found that white light is a mixture of different colors, each of which is pure in the sense that it can't be separated further.

(In fact, a particular color of light can be split one more time in a different way, according to its so-called "polarization." This aspect of light is not vital to understanding the character of quantum electrodynamics, so for the sake of simplicity I will leave it out - at the expense of not giving you an absolutely complete description of the theory. This slight simplification will not remove, in any way, any real understanding of what I will be talking about. Still, I must be careful to mention all of the things I leave out.)

When I say "light" in these lectures, I don't mean simply the light we can see, from red to blue. It turns out that visible light is just a part of a long scale that's analogous to a musical scale in which there are notes higher than you can hear and other notes lower than you can hear. The scale of light can be described by numbers - called the frequency - and as the numbers get higher, the light goes from red to blue to violet to ultraviolet. We can't see ultraviolet light, but it can affect photographic plates. It's still light - only the number is different. (We shouldn't be so provincial: what we can detect directly with our own instrument, the eye, isn't the only thing in the world!) If we continue simply to change the number, we go out into X-rays, gamma rays, and so on. If we change the number in the other direction, we go from blue to red to infrared (heat) waves, then television waves, and radio waves. For me, all of that is "light." I'm going to use just red light for most of my examples, but the theory of quantum electrodynamics extends over the entire range I have described, and is the theory behind all these various phenomena.

Newton thought that light was made up of particles - he called them "corpuscles" - and he was right (but the reasoning that he used to come to that decision was erroneous). We know that light is made of particles because we can take a very sensitive instrument that makes clicks when light shines on it, and if the light gets dimmer, the clicks remain just as loud - there are just fewer of them. Thus light is something like raindrops - each little lump of light is called a photon - and if the light is all one color, all the "raindrops" are the same size.

Figure 6: Feynman's original text before preprocessing.

Feynman_4.txt

begin light
Newton start look light first thing find white light mixture color
separate white light prism various color but put light one color red forinstance through another prism find not separate further
Newton find white light mixture different color each pure sense not separate further
infact particular color light split different way according polarization
aspect light not vital understanding character quantum electrodynamics sake simplicity leave-out expense not give you absolute complete description theory
slight simplification not remove real understanding talk about
still careful mention all thing leave out
say light lectures not mean light see red blue
turn out visible light part long scale analogous musical scale note higher you hear other note lower you hear
scale light describe number called frequency number get higher light go red blue violet ultraviolet
not see ultraviolet light but affect photographic plate
still light only number different
not provincial detect direct own instrument eye not only thing world
continue change number go out X-rays gamma-rays
change number other direction go blue red infrared heat wave television wave radio wave all light
use red light most example but theory quantum electrodynamics extend over entire range describe theory behind all various phenomena
Newton think light make up particle call corpuscle right but reasoning use come decision erroneous
know light make particle because take very sensitive instrument make click light shine light get dimmer click remain just loud fewer
thus light something like raindrop each little lump light called photon light all one color all raindrops same size

Figure 7: Input text for the Python code after preprocessing according to our lemmatisation rules. The auxiliary verbs "can, be, should, will" and "must", and the pronouns "I, we, he, it, this" and "that" have been eliminated. Contrasting words such as "but" and "not" have been kept.

to locate sentences in the text 'describing' a given thematic unit (resp. theme). We extract those sentences that are most frequently intersected by a boundary cycle (resp. the boundary cycles) corresponding to a thematic unit (resp. theme).

5 Examples

The selection of our examples is guided by the idea that they should treat different genres, but also carry complexity and literary quality. Their referential structures are therefore likely to be better described by our two-dimensional model than by the one-dimensional models of distributional semantics.

Our first example is an excerpt from Richard Feynman's popular science book on quantum electrodynamics (Feynman, 1985), p. 13-14.

The output of the TOPTEXT algorithm provides the following list of themes into which the text is divided:

theme 1 in boundary components [1, 2, 4, 6, 7, 9, 12] in sentences [2] keywords : 'light' T-central words : [('light', 73), ('not', 30), ('color', 22), ('red', 18), ('all', 18)] C-central words : [('light', 23.76), ('not', 11.59), ('all', 6.6), ('color', 6.12), ('thing', 5.47)]

theme 2 in boundary components [3] in sentences [1, 3] keywords : 'find', 'white' T-central words : [('find', 2), ('white', 2)] C-central words : [('find', 0.75), ('white', 0.75)]

theme 3 in boundary components [5] in sentences [2, 3] keywords : 'light', 'white' T-central

Figure 8: Flow graph for the Feynman text, cf. Fig. 1. The sentence cycles are here depicted linearly for the sake of clarity; the last word in each line is thought to be linked to the first. The blue edges are sentence edges, while the red edges are relational cycles. The graph is connected.

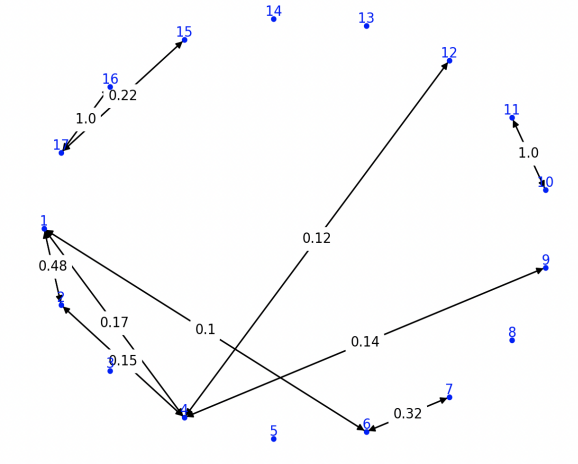

Figure 9: In the dual graph for the Feynman text, each node represents a boundary cycle (17 in total). The edge weights in the dual graph indicate how many edges in the flow graph are shared. For instance, boundary cycle 1 and boundary cycle 2 share almost half of their edges (0.48) and can therefore be regarded as neighbouring regions in the embedding surface.

words : [('light', 2), ('white', 2)] C-central words : [('white', 0.75), ('light', 0.59)]

theme 4 in boundary components [8] in sentences [3, 4, 5, 10, 12, 20] keywords : 'light' T-central words : [('light', 3), ('not', 2), ('different', 2), ('each', 2), ('number', 2)] C-central words : [('light', 1.05), ('called', 0.91), ('not', 0.75), ('number', 0.75), ('each', 0.59)]

theme 5 in boundary components [10, 11] in sentences [5, 17] keywords : 'quantum', 'electrodynamics' T-central words : [('quantum', 4), ('electrodynamics', 4)] C-central words : [('quantum', 1.18), ('electrodynamics', 1.18)]

theme 6 in boundary components [13] in sentences [8, 10] keywords : 'red', 'blue' T-central words : [('red', 2), ('blue', 2)] C-central words : [('red', 0.59), ('blue', 0.59)]

theme 7 in boundary components [14] in sentences [14, 15] keywords : 'change', 'number' T-central words : [('number', 2), ('change', 2)] C-central words : [('number', 0.59), ('change', 0.59)]

theme 8 in boundary components [16, 17, 15] in sentences [18, 19] keywords : 'make' T-central words : [('make', 6), ('particle', 4), ('light', 2), ('up', 2)] C-central words : [('particle', 1.82), ('light', 0.75), ('make', 0.63), ('up', 0.01)]

For instance, the output "theme 3 in boundary components [5] in sentences [2, 3]" means: The third of the eight extracted themes is "white light", which is the most frequent term in boundary component 5. It is best expressed or articulated in sentences 2 and 3: "He separated white light with a prism into various colors, but when he put light of one color – red, for instance – through another prism, he found it could not be separated further. Thus Newton found that white light is a mixture of different colors, each of which is pure in the sense that it cannot be separated further." Here, keywords, T-central words and C-central words coincide up to ordering.

If one were to ask human readers 'what this text is about' or how they would summarise it or structure it thematically, a plausible answer would be: i) The text is about light. ii) Newton found that white light is a mixture of different colours; iii) Light does not refer solely to visible light such as red or blue. Changing the frequency yields different wavelengths; iv) All this is described in quantum electrodynamics; v) Light therefore consists of particles, as Newton had already surmised.

One can now easily assign the above themes to this answer: i) The text is about light (theme 1). ii) Newton found (theme 2) that white light (theme 3) is a mixture of different colours; iii) Light does not only refer to visible light such as red or blue (theme 6). If one changes the frequency, one arrives

Lear_orig.txt

```
Meantime, we shall express our darker purpose.
Give me the map there. Know that we have divided
In three our kingdom; and 'tis our fast intent
To shake all cares and business from our age,
Conferring them on younger strengths, while we
Unburthen'd crawl toward death. Our son of Cornwall,
And you, our no less loving son of Albany,
We have this hour a constant will to publish
Our daughters' several dowers, that future strife
May be prevented now. The Princes, France and Burgundy,
Great rivals in our youngest daughter's love,
Long in our court have made their amorous sojourn,
And here are to be answer'd. Tell me, my daughters,
Since now we will divest us both of rule,
Interest of territory, cares of state
Which of you shall we say doth love us most?
Where nature doth with merit challenge. Goneril,
Our eldest-born, speak first.
```

Figure 10: Shakespeare's original text before preprocessing.

Lear_10.txt

```
meantime we shall express dark purpose
give me map there know we divide
three kingdom fast intent
shake all care business age
conferring them young strength while we
unburdened crawl toward death son Cornwall
you no less loving son Albany
we have this hour constant will publish
daughter several dower future strife
may prevented now prince France Burgundy
great rival young daughter love
long court make amorous sojourn
here answered tell me daughter
since now we will divest us both rule
interest territory care state
which you shall we say do love us most
where nature do merit challenge Goneril
eldestborn speak first
```

Figure 11: Input text for the Python code after preprocessing according to our lemmatisation rules.

at different wavelengths (theme 7); iv) All this is described in quantum electrodynamics (theme 5); v) Light therefore consists, as Newton had already assumed, of particles (theme 8). In this context, theme 4 is redundant in relation to the main theme, as evidenced by the fact that, like theme 1, it is expressed through many sentences in the text.

Notably, this result is achieved purely through the text itself, without the addition of semantic fields or any external additions such as world knowledge. This is interesting for the envisaged application to literary, philosophical, and scientific texts where formal elements are important for meaning. Furthermore, the vocabulary used in these types of text differs from that found in everyday and ordinary texts: literary texts contain specific, carefully chosen, and rare words. In philosophical and academic texts it is important that the terminology remains consistent. Both aspects are precisely taken care of by the relational cycles.

After this scientific text, our second example is a poetic text from Shakespeare's play King Lear. Here, the sentence cycles are verses.

Based on the grouping of boundary cycles in the dual graph shown in Fig. 13, the text is divided into four thematic components, whose output is as follows:

theme 1 in boundary components [1, 2, 3] in sentences [15] keywords : 'we' T-central words : [('we', 20), ('daughter', 10), ('shall', 8), ('care', 8), ('you', 8)] C-central words : [('we', 6.5), ('daughter', 4.93), ('care', 3.94), ('you', 3.94), ('now', 3.94)]

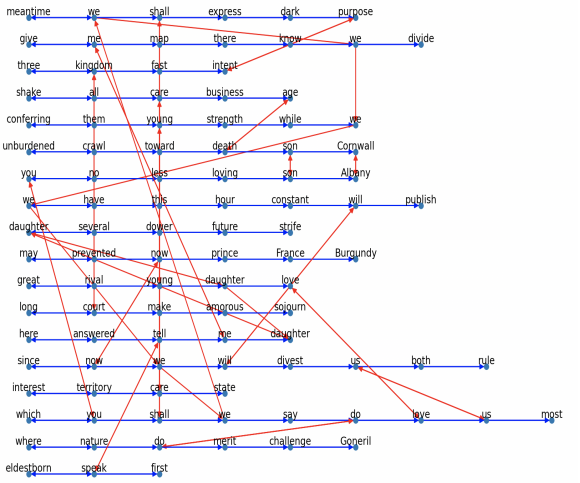

Figure 12: Flow graph for the Shakespeare text.

theme 2 in boundary components [4] in sentences [1, 4, 10, 12] keywords : 'daughter', 'me', 'we', 'young' T-central words : [('we', 2), ('me', 2), ('young', 2), ('daughter', 2), ('map', 1)] C-central words : [('we', 0.99), ('me', 0.99), ('young', 0.99), ('daughter', 0.99), ('strength', 0.49)]

theme 3 in boundary components [5, 6] in sentences [5, 6] keywords : 'son' T-central words : [('son', 4), ('Cornwall', 2), ('Albany', 2)] C-central words : [('son', 1.97), ('Cornwall', 0.99), ('Albany', 0.99)]

theme 4 in boundary components [7] in sentences [7, 13] keywords : 'we', 'will' T-central words : [('we', 2), ('will', 2), ('have', 1), ('this', 1), ('hour', 1)] C-central words : [('we', 0.99), ('will', 0.99), ('have', 0.49), ('this', 0.49), ('hour', 0.49)]

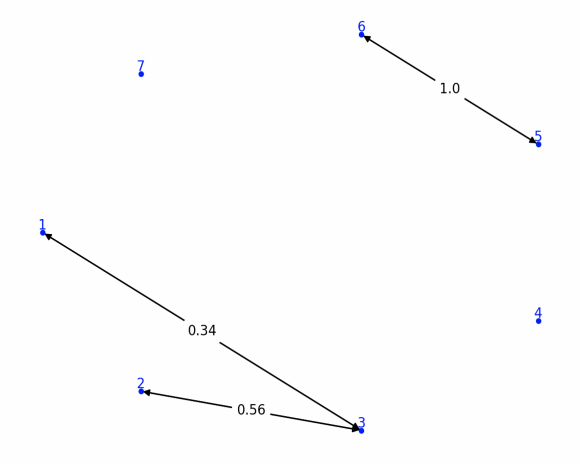

Figure 13: In the dual graph for the Shakespeare text, each node represents a boundary cycle. In this example, there are 7 boundary cycles in total, with two connected components: boundary cycles 1, 2 and 3, as well as 5 and 6, due to the number of shared edges exceeding the average.

This thematic structure is also highly plausible: the main character is Lear, who speaks of himself in the plural majestatis, using "we", and makes a declaration of intent regarding his succession ("we will"), in which he calls upon his 'sons' Cornwall and Albany, who are already married to Lear's first two daughters, the third of whom is now also to be married. This paraphrase becomes clearer when the themes, as indicated above, are formulated or expressed in respective sentences.

Note that the extracted keywords are not merely the most frequent words in the text, but occur with roughly the same frequency as others. The result was achieved by incorporating manually defined semantic fields, namely: {purpose, intent}, {age, death}, {kingdom, court}, {Cornwall, Albany}, {tell, speak} making the text more coherent in line with what we said above.

6 Outlook

At this stage, the criteria for a successful text analysis are simply consistency with human expectation. In principle, this issue could be objectified by increasing the number of interpreters and the number of texts in an empirical survey. The text selection described above, comprising texts that are as varied and diverse as possible, is also beneficial for this purpose. One difficulty in defining criteria lies in the fact that the TOPTEXT model seeks to understand texts not merely in general terms, but in their particularity and immanent to the text itself. It is therefore not possible, as with distant reading, to calculate averages across many texts and then compare them with general thematic expectations.

We are well aware that the analysis presented leaves many open questions such as: what is the influence of the lemmatisation on the results? What can be gained in terms of accuracy by including semantic fields? How well does the model capture different thematic structures (concentrated, scattered, linearly structured, circularly structured, hierarchical)? How does our model compare to other methods? Regarding the first question, we have carried out extensive tests using varying numbers and types of stop words. The model proved to be suffiently stable, most of the topics remaining unchanged. Regarding other models we must first work out a method for a fair comparison as our model is, for instance, per se not trained on any data unless semantic fields are included.

Striking features of the TOPTEXT model are: Full transparency and interpretability. In line with the requirement for explainable AI, our algorithm clearly delivers comprehensible results. Prior training on a reference corpus is not required. Where external semantic fields are incorporated, the processes are controllable and scannable. This provides the basis for a transparent hermeneutic analysis of texts. Due to the simple architecture, the results are obtained in an efficient and cost-effective way.

References

Konstantine Arkoudas. 2023. Chatgpt is no stochastic parrot. but it also claims that 1 is greater than 1. *Philosophy & Technology*, 36(3)(54).

Eduard Benes. 1973. *Thema - Rhema - Gliederung und Textlinguistik*, pages 42–62.

David M. Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Timon Boehm. 2023. Zur erschliessung semantischer strukturen mit algebraisch-topologischen konzepten. vorstellung eines neuen textmodells anhand von also sprach zarathustra iv. *Nietzscheforschung*, 30,1:289–299.

Noam Chomsky. 1990. Topic ... comment: On formalization and formal linguistics. *Natural Language and Linguistic Theory*, 8(1):143–147.

Frantisek Danes. 1974. *Functional Sentence Perspective and the Organization of the Text*. Academia, Prague.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Richard P. Feynman. 1985. *QED: The Strange Theory of Light and Matter*. Princeton University Press, Princeton, NJ.

Algirdas Julien Greimas. 1971. *Strukturale Semantik. Methodologische Untersuchungen*. Vieweg, Braunschweig.

Michael A. K. Halliday. 1967. Notes on transitivity and theme in english. *Journal of Linguistics*, 3(1):37–81.

Zellig S. Harris. 1951. *Structural Linguistics*. University of Chicago Press, Chicago.

Francois Labourie. 2013. *Lectures on Representations of Surface Groups*. European Mathematical Society Publishing House, Zurich.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43.

Vilém Mathesius. 1975. *A Functional Analysis of Present Day English on a General Linguistic Basis*. Mouton, The Hague.

F. Rastier. 2001. *Arts et sciences du texte*. PUF, Paris.

Sophia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

L. Vanni and D. Mayaffre. 2026. Explore political discourse with transformers: Emergent paradigmatic and syntagmatic representations. In *Proceedings of the 15th Language Resources and Evaluation Conference (LREC 2026)*, Palma de Mallorca, Spain.

Claus Zittel. 2026. *Romansymphonik. Adornos ‚Theorie des Romans'*, volume 2 of *Leseszenen, hg. von Irina Hron*. Universitaetsverlag Winter, Heidelberg.