The Challenges of Multilingualism in the Search for Ancient Wisdom: A Case Study of VERITRACE's Text Matching Tool

Jeffrey C. Wolf*[1], Nicolò Cantoni[1], Eszter Kovács[1], Demetrios Paraschos[1], and Cornelis J. Schilt[1]

[1] *Vrije Universiteit Brussel, Pleinlaan 2 B-1050 Brussels, Belgium*

Abstract

The ERC-funded VERITRACE project is applying the latest digital tools, including machine learning algorithms, on a large corpus of early modern texts in order to trace the influence of ancient wisdom writings on the development of early modern natural philosophy. Innovative capabilities of the project include text matching, where a query text is used as a search 'query' across a much larger comparison corpus. This poses challenges when query and comparison corpora are multilingual. This paper will explore these issues using VERITRACE's Text Matching tool.

Keywords

digital humanities, distant reading, text matching, machine translation, multilingual texts, similarity

1. Introduction

The European Research Council is funding the ambitious five-year project (2023-2028) *Traces de la Verité: The reappropriation of ancient wisdom in early modern natural philosophy*, aka VERITRACE (ERC-StG Project VERITRACE, 101076836) [1]. Led by Professor Cornelis J. Schilt at the Vrije Universiteit Brussel (VUB), VERITRACE aims to uncover the influence of prominent ancient wisdom writings on natural philosophical discourse in the early modern period.[1]

VERITRACE applies sophisticated digital analysis techniques, including machine learning algorithms, to a large corpus of early modern texts, tracing referenced and more subtle uses of the *Corpus Hermeticum* (including the *Asclepius*), the *Chaldean* and *Sibylline Oracles*, and the *Orphic Hymns*. Moreover, it analyses how these texts were being used (employing Latent Semantic Analysis, among other tools), and with what sentiment they were discussed (using Sentiment Analysis) by their proponents and antagonists, and how these debates were influenced by key episodes in the transmission history of these texts. VERITRACE will provide the first-ever comprehensive analysis of ancient wisdom's role in shaping early modern natural philosophy, and it will do so by making use of new methodologies never employed at this scale to interpret the early modern history of science.

Humanities-Centred AI (CHAI), 4th Workshop at the 47th German Conference on Artificial Intelligence, September 23, 2024, Würzburg, Germany

* Corresponding author.

Jeffrey.Charles.Wolf@vub.be (J. Wolf); Nicolo.Cantoni@vub.be (N. Cantoni); Eszter.Kovacs@vub.be (E. Kovács); Demetrios.Paraschos@vub.be (D. Paraschos); Cornelis.Johannes.Schilt@vub.be (C. Schilt)

0009-0007-9879-4476 (J. Wolf); 0009-0008-4438-4778 (N. Cantoni); 0000-0002-5900-8301 (E. Kovács); 0009-0004-6785-8033 (D. Paraschos); 0000-0001-7826-8355 (C. Schilt)

[1] Much of the text in this introductory section has been adapted from the VERITRACE ERC-funding proposal [2]. A variation of this introductory material, along with portions of the case study of Everard's English translation of the *Divine Pymander*, will be printed in a forthcoming special issue of the journal *Society and Politics* [3].

VERITRACE draws on printed books, the most ubiquitous intellectual materials found in this period. Although early modern debates followed other modes of discourse, such as oral discussions and the circulation of manuscripts and letters, these often pertained to small circles of select readers. Books, on the other hand, were everywhere. Indeed, even if we only focus on works in Latin, English, German, French, Dutch, and Italian, the number of books that have come down from the early modern period is staggering, presenting lots of challenges – and opportunities.

Some of these challenges uniquely characterise VERITRACE as a digital humanities project. These features include:

1. **Multilingual Complexity:** The project grapples with at least 6 different languages, both modern and classical ones. Many digital humanities projects only contend with one or two languages, especially English. Since many natural language processing (NLP) techniques were initially developed for English-speaking users, the multilinguistic nature of VERITRACE raises its own set of challenges. In fact, even when available tools are comprehensive in terms of modern languages, they sometimes exclude, or have limited support for, classical ones like Latin.[2]
2. **Longue durée:** Spanning nearly two centuries (1540-1728), VERITRACE must account for significant shifts in historical context and linguistic meaning. An interpretation that applies to a smaller slice of the data cannot be assumed to apply to the whole, given change in historical context and linguistic meaning over time.
3. **Big Data Management:** With a corpus comprising hundreds of thousands of texts, the project requires sophisticated data handling and analysis techniques, beyond simple search processes. It will be resource intensive and sometimes require different tools and solutions, given the size of the data collections with which we are working.
4. **Complex Data Integration:** Because our data comes from different sources held in separate institutions collected over long periods of time, there are inherent challenges to integrating and harmonising the data. This necessitates paying careful attention to data cleaning and data transformation, along with careful documentation, so that we have a solid basis for subsequent analysis.

VERITRACE must also grapple with the familiar challenges of any distant reading project: uncertain accuracy of the underlying digital texts (OCR quality), the parameter-dependent nature of various NLP techniques, and so forth.

1.1. Distant Reading

Traditional methods of tracing textual influence would require an enormous research team or a drastically reduced scope. But this is where digital techniques come in, most notably from the field of distant reading, which have been developed specifically to query large corpora. These techniques, closely related to natural language processing, allow for the analysis of large text corpora, identifying patterns and uncovering both prominent and neglected works, the latter termed 'the great unread' by Margaret Cohen [4, 5]. Famously, early modern writers would

[2] For example, Latin is not available as a default trained pipeline package in the open-source natural language processing library *spaCy* (https://spacy.io/usage/models), although Patrick J. Burns has created *LatinCy* to fill this gap (https://spacy.io/universe/project/latincy). The *Natural Language Toolkit* (*NLTK:* https://www.nltk.org) likewise has limited support for Latin (e.g. for tokenisation), though the *Classical Language Toolkit* (*CLTK*: http://cltk.org) – which does support Latin – has been developed to supplement this. Another example: *OpenSearch* has no built-in language analyser for Latin (https://opensearch.org/docs/latest/analyzers/language-analyzers/).

rarely include references to their source material, which provides one of the key challenges for the project.

Recent advancements in book digitisation have greatly expanded the potential of distant reading approaches. Improved OCR technology now yields meaningful results even with suboptimal text recognition [6, 7]. Online repositories, like those of the Bibliothèque nationale de France, provide standardised data for content extraction, facilitating large-scale analysis [8, 9].

VERITRACE's chosen Distant Reading Corpus (DRC) consists of several hundred thousand works from important European library collections, written in Latin, French, German, Dutch, English, and Italian, including:

- *Early English Books Online (EEBO) (ProQuest),* which in its EEBO-TCP format developed by the Text Creation Partnership[3] contains about 60,000 English and Latin texts published between 1540 and 1700 (hereafter **'EEBO'** unless we refer specifically to our custom version of EEBO, which we call 'VEEBO')
- *Gallica* (Bibliothèque nationale de France) contains almost 125,000 books published between 1540 and 1728 in a variety of languages including French, Italian, Dutch, and Latin (hereafter '**Gallica**')
- The *Digitale Sammlungen* of the Bavarian State Library, which contain more than 340,000 books published between 1540 and 1728, including in Latin, German, French, Greek, Italian, and Dutch, among others (hereafter '**BSB**')

This carefully selected corpus enables VERITRACE to make substantive claims about the existence and evolution of the *prisca sapientia* tradition, e.g. how prevalent it was and the level of interest in it over time. Did curiosity in the *Corpus Hermeticum*, for instance, decline after the first quarter of the seventeenth-century, or not? By interrogating a truly representative sample of books, we can make reasonable claims about levels of interest and prevalence. A rigorously statistical frame of mind underpins our approach, and we believe the sources we have chosen can be the basis for constructing a representative sample size of books published in Europe between 1540 and 1728.

2. Monolingual Text Matching

The above has been a general introduction to VERITRACE and how it will harness digital techniques in the pursuit of its intellectual goals. We turn now to a specific tool in development, which we call *Text Matching*. In the following discussion, we see some of the promise, as well as the challenges, inherent in using such a tool, especially with multilingual corpora.

To make the following observations more concrete, we will work with a specific research question to explore our Text Matching tool: *what was the influence (however vaguely defined) of the first English translation of the* Divine Pymander *(1650) upon the subsequent generation of thinkers, who published texts in English between 1650 and 1680?*

We approach this in terms of the influence of a specific **query text** upon a much larger **comparison corpus.** Sometimes, especially in the Figures below, we refer, intuitively, to a *source text* and a *target corpus*, but the query text and comparison corpus terminology is more

[3] https://textcreationpartnership.org/about-the-tcp/

generalisable and to be preferred.[4] The query text is the first translation into English of a work from the *Corpus Hermeticum*; namely, John Everard's *The divine Pymander of Hermes Mercurius Trismegistus,* published in 1650 [10], and the comparison corpus (a subset of our larger Distant Reading Corpus) consists of all the English-language texts contained in EEBO published between 1650 and 1680: 18,633 individual texts in total.

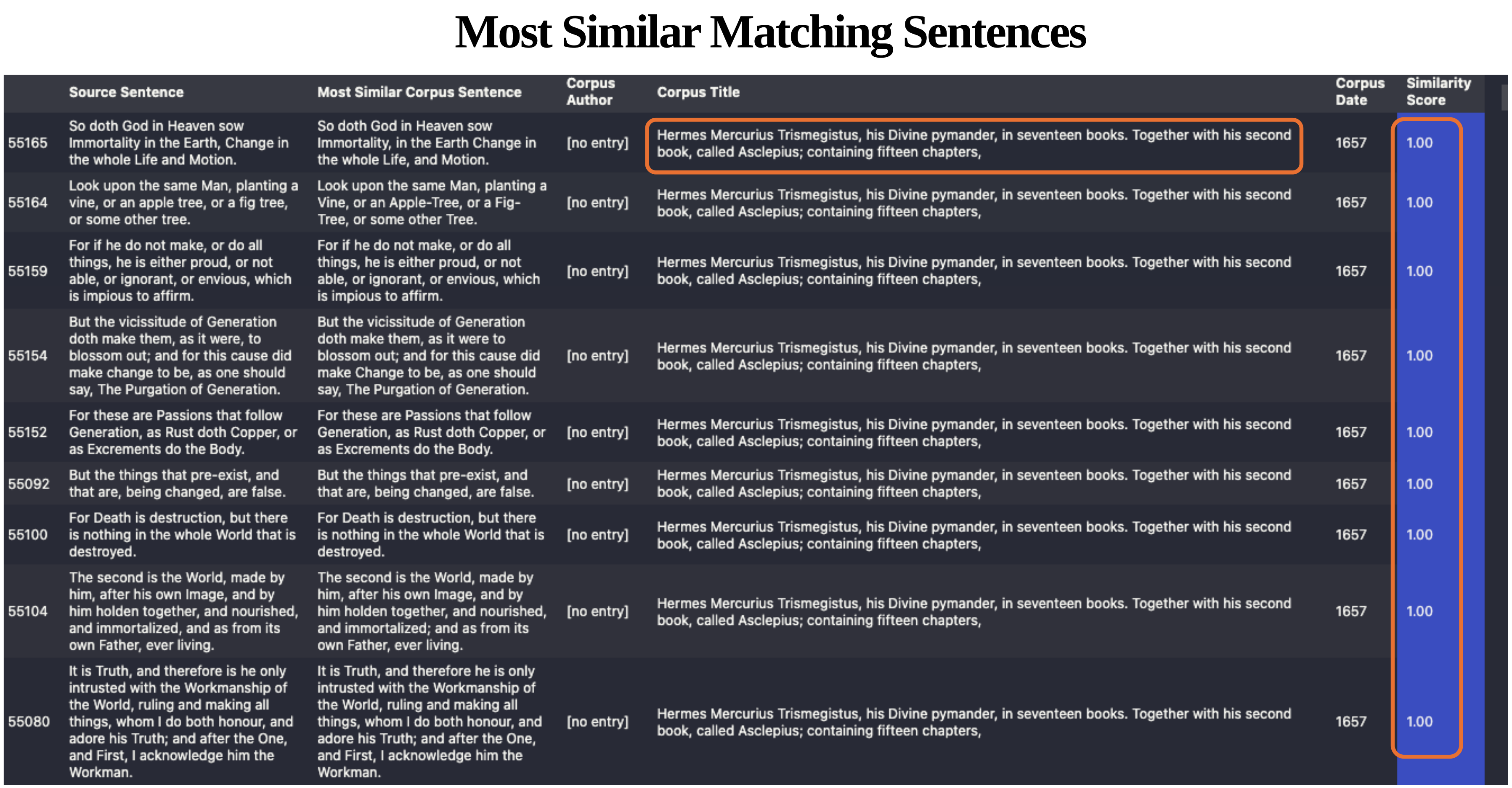
VERITRACE provides the user the ability to conduct simple and more complex keyword searches of the comparison corpus, including keyword search, fuzzy searches, wildcard searches, and exact phrase searches, among others. So traditional keyword search is part of the VERITRACE toolkit, but our Text Matching tool moves beyond this, for we want the ability to match entire passages from our query text with similar ones from the comparison corpus. Similar lexical phrasing, regardless of the precise words used, should be discoverable. In other words, we are building a kind of early modern plagiarism detector.[5]

Text Matching, as we call this, can likewise be seen as a more ambitious kind of search. It does not take a single keyword or phrase as input but instead the entire query text itself – all its sentences individually and collectively. Then we attempt to find the most similar matching sentences and passages (groups of sentences) in the comparison corpus and rank them based on similarity to the sentences found in the query text.

Before further discussion, let us see this in action. First, we want to identify the most similar matching sentences between query text and comparison corpus (see next page):

[4] We are being cautious about the terms in use here. In this case, the query text is indeed the *source* text, and the comparison corpus is the *target* corpus – we are asking what influence the source text had on the subsequent target corpus. We assume some kind of cause and historical effect. But the Text Matching tool works in the other direction as well. Perhaps one has a set of all of Isaac Newton's works, and one wants to know if they had an influence on a single text of a later thinker's work. In that case, it might make more sense to use the target corpus (text) as the source text, and Newton's collective works as the target corpus, as it better aligns with traditional information retrieval. But because we precompute all the vectors beforehand, computationally it will not make much difference if we choose a one-to-many instead of a many-to-one comparison. We retain bidirectionality. Therefore, we favour the more neutral 'query text' and 'comparison corpus' terminology.

[5] At this stage of the VERITRACE project, this is meant more metaphorically than in the strict sense that we are consciously using standard methods of plagiarism detection (though there is some overlap). For a description of some research in plagiarism detection – including the use of translated texts – see [11], especially 49.4.



Most Similar Matching Sentences

	Source Sentence	Most Similar Corpus Sentence	Corpus Author	Corpus Title	Corpus Date	Similarity Score
55165	So doth God in Heaven sow Immortality in the Earth, Change in the whole Life and Motion.	So doth God in Heaven sow Immortality, in the Earth Change in the whole Life, and Motion.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00
55164	Look upon the same Man, planting a vine, or an apple tree, or a fig tree, or some other tree.	Look upon the same Man, planting a Vine, or an Apple-Tree, or a Fig-Tree, or some other Tree.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00
55159	For if he do not make, or do all things, he is either proud, or not able, or ignorant, or envious, which is impious to affirm.	For if he do not make, or do all things, he is either proud, or not able, or ignorant, or envious, which is impious to affirm.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00
55154	But the vicissitude of Generation doth make them, as it were, to blossom out; and for this cause did make change to be, as one should say, The Purgation of Generation.	But the vicissitude of Generation doth make them, as it were to blossom out; and for this cause did make Change to be, as one should say, The Purgation of Generation.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00
55152	For these are Passions that follow Generation, as Rust doth Copper, or as Excrements do the Body.	For these are Passions that follow Generation, as Rust doth Copper, or as Excrements do the Body.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00
55092	But the things that pre-exist, and that are, being changed, are false.	But the things that pre-exist, and that are, being changed, are false.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00
55100	For Death is destruction, but there is nothing in the whole World that is destroyed.	For Death is destruction, but there is nothing in the whole World that is destroyed.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00
55104	The second is the World, made by him, after his own Image, and by him holden together, and nourished, and immortalized, and as from its own Father, ever living.	The second is the World, made by him, after his own Image, and by him holden together, and nourished, and immortalized; and as from its own Father, ever living.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00
55080	It is Truth, and therefore is he only intrusted with the Workmanship of the World, ruling and making all things, whom I do both honour, and adore his Truth; and after the One, and First, I acknowledge him the Workman.	It is Truth, and therefore he is only intrusted with the Workmanship of the World, ruling and making all things, whom I do both honour, and adore his Truth; and after the One, and First, I acknowledge him the Workman.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00

Figure 1. The most similar matching sentences between query text and comparison corpus.

Please note a few details about this result (see **Figure 1**). First, the 'Similarity Score' column at the far right contains score values, which should be interpreted as equivalent to 'relevance scores' (relative to each other). The values provide a ranking of the most relevant match results of the query text matched against the comparison corpus. The similarity score ranges between 0 (no similarity) to 1 (exact similarity, e.g. identical), with higher values indicating greater similarity between the query and comparison texts being matched.

Another detail is that, within our comparison corpus, there appears to be another edition of the *Divine Pymander*, published 7 years later [12]. It is almost identical to our query text, which is why sentences from the 1657 edition of the *Divine Pymander* are the top matches to sentences in the query text (the 1650 edition). This is exactly what we would expect, if the editions are virtually identical to each other (with similarity scores of '1.00'). It provides a 'common sense' check on the validity of our Text Matching tool.

2.1. Under the Hood: TF-IDF and Cosine Similarity

Because the concept of a similarity score is so central to what we are doing, it is worth pausing to look 'under the hood' to see how we are computing it. This is important to understand because we will be using similarity scores extensively in the examples that follow.

Here is the relevant part of the NLP pipeline: we begin our text normalisation with some basic preprocessing of the 'raw' text from our data sources. Our 'raw' text is, in practice, generally derived from files obtained from our data sources. These include xml, html, or hocr files, depending on the initial data source, which we then parse to extract the text we will use for downstream analysis. The extracted text itself is often saved in individual txt files for ease-of-use. The parsing and extraction process is full of pitfalls (not to mention long compute times, given the number of files), and there are many considerations and nuances that we must consider, in order to capture the most accurate version of the digitised transcription of the original printed text (this is not the place to discuss these intricacies further).

We then segment the query text and the comparison corpus texts into sentences (with a default minimum word count of 5, which is adjustable) and groups of sentences (with a default 'chunk size' of 3 sentences, also adjustable). There are multiple decisions here too: we prefer to segment the pre-tokenised raw text into sentences rather than segment sentences from the tokenised text. This adds some time and inefficiency to the pipeline, but we believe it is more accurate because errors in tokenisation could compound into segmentation.

With the sentences in hand, we then tokenise them using the *SpaCy* library.[6] There are many ways to optimise the tokenization process at this stage. We have added some custom rules to handle early modern English, as opposed to the 21st-century English that *SpaCy* assumes in its trained English pipeline. The results seem sufficient for now, but we continue to iterate this part of the pipeline. Also, we prefer to use *LatinCy* for tokenising Latin texts, and, in general, each language requires its own customisations.

Once tokenisation is done, we vectorise the sentences and sentence chunks, using the *scikit-learn* machine learning library, extracting their features using TF-IDF (Term Frequency-Inverse Document Frequency).[7] TF-IDF is a well-known and popular algorithm in vector semantics that

[6] https://spacy.io

[7] https://scikit-learn.org/stable/. In particular, we use the TfidfVectorizer, which converts "a collection of raw documents to a matrix of TF-IDF features."

allows us to convert text into sparse numerical vectors by assigning weights to document terms (words) based on their frequency in a document, offset by assigning a higher weight to terms that only occur in a few documents in a larger corpus.[8] This is to say, words that are more unique to, but also occur frequently within a document (in comparison to a larger corpus) are given a higher weight under the TF-IDF paradigm. Note that this treats the meaning of a word simplistically as a function of the number of nearby words [13]. Word order is essentially ignored in this semantics.

Using TF-IDF as our vector semantics model here is a choice, and it needs defending.[9] After all, there are more advanced approaches readily available, including using dense vector word embeddings like word2vec or GloVe, or using the latest transformer models to generate dense embeddings (e.g. using BERT, GPT, and its descendants).[10] As the discussion below will show, the demands of our multilingual corpus are pushing us in that direction – especially towards transformer-based embeddings.[11] But it is worth seeing both why that is the case and also why using TF-IDF is still worth retaining, even if we can obtain more semantically accurate ones with newer models.[12] For TF-IDF still provides a nice balance between simplicity and effectiveness, and, as Jurafsky and Martin note, it is "a great baseline, the simple thing to try first."[13] And so we do.

Vectorisation using TF-IDF is not sufficient to produce similarity scores, of course. We must compute those, and here again, there is more than one option, though using the cosine of the angle between vectors as a measure of similarity is the predominant paradigm in the 'vector space model for scoring' – it is formally referred to as *cosine similarity* ([14], Section 6.3, p. 120). Because TF-IDF represents each textual unit as a numerical vector – a point in vector space (this is the well-known *vector space model*) – we can compute similarity by computing the cosine of

[8] This is an oversimplification of sorts, and there are different ways to compute the TF-IDF weight, and many variations of it. See [13], especially pp. 108-114.

[9] There is often a confusing array of terminology in NLP, in contrasting use both among academics and DH developers. Because we find the discussion of vector semantics in [13] to be admirably clear and because it is one of the leading textbooks in the field, we prefer to use their terminology where possible. Two other texts have been particularly helpful in situating our approach: the classic textbook on information retrieval by Manning, et al [14] – for, our 'text matching' is a kind of information retrieval, with query and ranked results retrieval. Finally, a recent reference work on many of these topics is [15], though the quality of the chapters is variable and not all of them have been recently updated despite the 2022 publication date, so some caution is warranted. But it does have a very helpful Glossary (pp. 1243-1290), and [16] and [17] are most relevant to our discussion. This is usefully compared to the older NLP handbook by Indurkhya and Damerau [18]. For a more applied approach with illuminating case studies, see [9]. Despite the older code, [19] is also still worth consulting.

[10] Many of the latest models, Transformer-based or otherwise, are readily available on the open-source machine learning platform, Hugging Face (https://huggingface.co). For word2vec, see https://www.tensorflow.org/text/tutorials/word2vec. For GloVe, see https://nlp.stanford.edu/projects/glove/.

[11] Attempts have been made to use TF-IDF on multilingual corpora by modifying it to use subword tokenization (STF-IDF) [20]. An open-source model 'Text2Text' has been created to implement this. We have not evaluated the claims or the model.

[12] Jurafsky and Martin claim that dense vector embeddings universally generate more accurate results than sparse vector models, including TF-IDF: "It turns out that dense vectors work better in every NLP task than sparse vectors" (117). It is not entirely clear why this is the case, though a plausible theory, they suggest, is that the smaller parameter space (e.g. 300-dimensional dense vectors vs. 50,000 sparse ones) better avoids overfitting and enhances generalisation. It also better represents synonymy. See [13], p. 117.

[13] See [13], p. 113. Note that we are using 'TF-IDF' here as a stand-in for both its traditional application and also more sophisticated variations of it, like the Okapi BM25 algorithm, which we are likely to prefer for our lexical similarity scoring. It is still within the family of TF-IDF models, however. For Okapi BM25, see [14], especially 11.4.3.

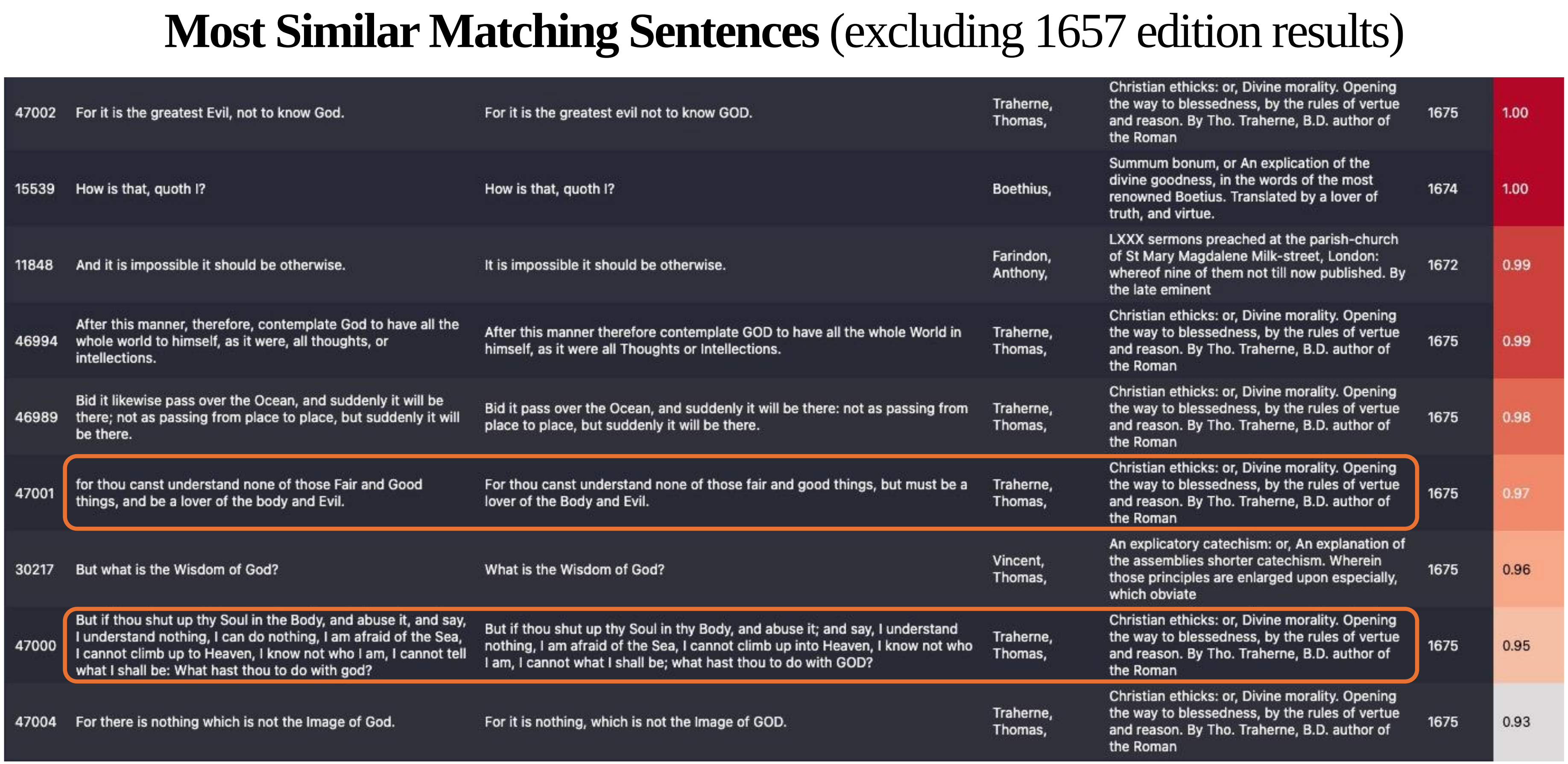
the angle between the two vectors.[14] A smaller cosine angle means the vectors are more similar. The similarity score is just the cosine similarity metric, explained above. Note that, when using TF-IDF, it will always be between 0 and 1 (even though the cosine of an angle between two vectors can vary between -1 and 1) because term frequency values cannot be negative ([13], p. 111).[15]

2.2. Text Matching: Initial Results

Let's return to the Text Matching results. For illustrative purposes, it is not very interesting to see a list of the results from the 1657 *Divine Pymander* (refer again to **Figure 1**) – given the unsurprising, near-identical nature of the texts - so what happens if we exclude these (see next page)?

[14] See Ch. 6 in [14]. Jurafsky and Martin (see their historical notes on pp. 129-131) attribute the original vector space model to Salton [21] in the realm of information retrieval, though Osgood had already suggested in 1957 that the meaning of a word could be represented as a point in a multidimensional 'semantic' space [22].

[15] There are a handful of cosine terms that are easy to mix-up: *cosine*, *cosine of an angle*, *cosine similarity* and *cosine distance*. *Cosine* simply refers to the well-known trigonometric function of the same name, normally defined as the ratio of the lengths of the side of a right triangle adjacent to the angle and the hypotenuse (https://mathworld.wolfram.com/Cosine.html). This also explains what is meant by the 'cosine of an angle' – it is just the evaluation of the cosine function for a specific angle. *Cosine similarity* is a measure of the similarity between two vectors of the same dimensionality that is derived from the cosine of the angle between those two vectors (see the helpful discussion in section 6.4 of [13]). When two vectors are more similar, the cosine of the angle between them is smaller. Therefore, cosine of an angle and cosine similarity are interrelated. Cosine distance is also related because it is simply 1 – cosine similarity (https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/cosdist.htm). But we will not use cosine distance directly in this discussion.



Most Similar Matching Sentences (excluding 1657 edition results)

47002	For it is the greatest Evil, not to know God.	For it is the greatest evil not to know GOD.	Traherne, Thomas,	Christian ethicks: or, Divine morality. Opening the way to blessedness, by the rules of vertue and reason. By Tho. Traherne, B.D. author of the Roman	1675	1.00
15539	How is that, quoth I?	How is that, quoth I?	Boethius,	Summum bonum, or An explication of the divine goodness, in the words of the most renowned Boetius. Translated by a lover of truth, and virtue.	1674	1.00
11848	And it is impossible it should be otherwise.	It is impossible it should be otherwise.	Farindon, Anthony,	LXXX sermons preached at the parish-church of St Mary Magdalene Milk-street, London: whereof nine of them not till now published. By the late eminent	1672	0.99
46994	After this manner, therefore, contemplate God to have all the whole world to himself, as it were, all thoughts, or intellections.	After this manner therefore contemplate GOD to have all the whole World in himself, as it were all Thoughts or Intellections.	Traherne, Thomas,	Christian ethicks: or, Divine morality. Opening the way to blessedness, by the rules of vertue and reason. By Tho. Traherne, B.D. author of the Roman	1675	0.99
46989	Bid it likewise pass over the Ocean, and suddenly it will be there; not as passing from place to place, but suddenly it will be there.	Bid it pass over the Ocean, and suddenly it will be there: not as passing from place to place, but suddenly it will be there.	Traherne, Thomas,	Christian ethicks: or, Divine morality. Opening the way to blessedness, by the rules of vertue and reason. By Tho. Traherne, B.D. author of the Roman	1675	0.98
47001	for thou canst understand none of those Fair and Good things, and be a lover of the body and Evil.	For thou canst understand none of those fair and good things, but must be a lover of the Body and Evil.	Traherne, Thomas,	Christian ethicks: or, Divine morality. Opening the way to blessedness, by the rules of vertue and reason. By Tho. Traherne, B.D. author of the Roman	1675	0.97
30217	But what is the Wisdom of God?	What is the Wisdom of God?	Vincent, Thomas,	An explicatory catechism: or, An explanation of the assemblies shorter catechism. Wherein those principles are enlarged upon especially, which obviate	1675	0.96
47000	But if thou shut up thy Soul in the Body, and abuse it, and say, I understand nothing, I can do nothing, I am afraid of the Sea, I cannot climb up to Heaven, I know not who I am, I cannot tell what I shall be: What hast thou to do with god?	But if thou shut up thy Soul in thy Body, and abuse it; and say, I understand nothing, I am afraid of the Sea, I cannot climb up into Heaven, I know not who I am, I cannot what I shall be; what hast thou to do with GOD?	Traherne, Thomas,	Christian ethicks: or, Divine morality. Opening the way to blessedness, by the rules of vertue and reason. By Tho. Traherne, B.D. author of the Roman	1675	0.95
47004	For there is nothing which is not the Image of God.	For it is nothing, which is not the Image of GOD.	Traherne, Thomas,	Christian ethicks: or, Divine morality. Opening the way to blessedness, by the rules of vertue and reason. By Tho. Traherne, B.D. author of the Roman	1675	0.93

Figure 2. The most similar matching sentences between query text and comparison corpus, excluding results from the 1657 edition of the *Divine Pymander*.

Not all the sentences that match are particularly meaningful (e.g. see the second result in **Figure 2** above). But we also find that many of the most similar sentences come from Thomas Traherne's *Christian ethicks* (1675) [23], and it appears that he is copying directly from the query text (or the 1657 edition), though he sometimes alters the language in minor ways (see the two results we have highlighted).[16] So this is certainly more interesting and worthy of follow-up: is Traherne acknowledging the *Divine Pymander* as his source, or pretending he has written this himself?[17] The VERITRACE Text Matching tool provides a line of investigation.[18]

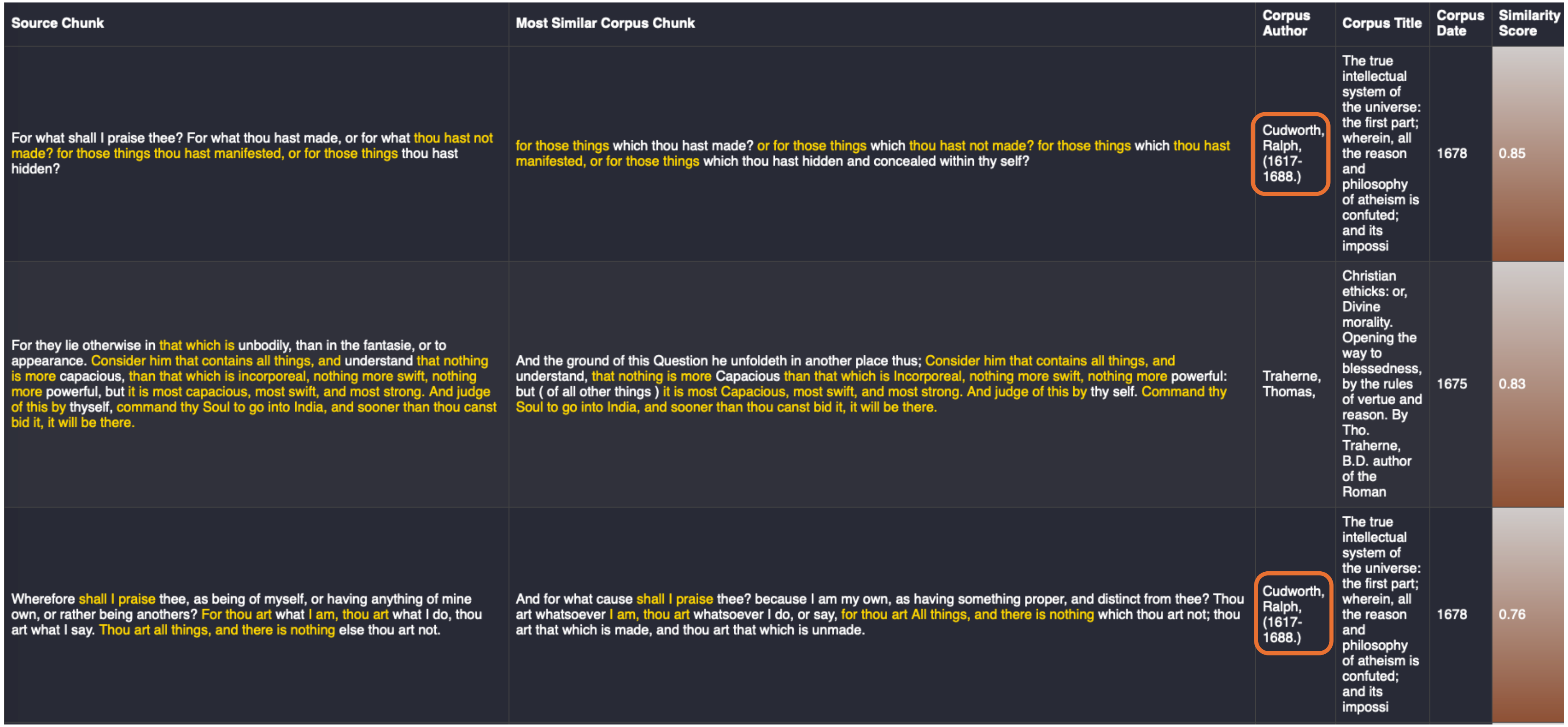
Now, sentence matching is a start, but what we would really like is to find entire passages – groups of sentences (sometimes referred to as 'chunks') – that match between our query text and the comparison corpus. And indeed, this is the next step. Once again, we first exclude the matching passages from the 1657 *Divine Pymander* and then display the following results (see next page):

[16] The connection between Traherne and the *Corpus Hermeticum* is not a novel observation, though the Text Matching tool is a new way to observe it. Already by the 1960s, historians were connecting the two [24].

[17] It is also worth emphasising that just because we find matching sentences or passages between two sentences or two sentence chunks does not prove that the author of the comparison text used the source text. As we have been reminded by the existence of both a 1650 and 1657 edition of the *Divine Pymander*, any number of similar editions could have been used. But more generally, it could be that both texts used a third source. We cannot rule that out, which is why the Text Matching Tool provides grounds for further inquiry but should not be considered determinative evidence in itself. It is a tool of inquiry – not proof.

[18] In this instance, Traherne is in fact summarising what Hermes Trismegistus says in his 'Poemander'. He is not trying to claim the thoughts as his own. See p. 443, which introduces this discussion: "*Trismegistus* counteth thus, *First GOD, secondly the World, thirdly Man: the World for Man, and Man for GOD. Of the Soul that which is sensible is Mortal, but that which is reasonable Immortal…* This in his *Poemander*." The matching sentence found in the second highlighted result in **Figure 2** ("But if thou shut up thy Soul in thy Body…") is printed on p. 447 [23].

Matching Sentence Groups (excluding 1657 edition results)



Source Chunk	Most Similar Corpus Chunk	Corpus Author	Corpus Title	Corpus Date	Similarity Score
For what shall I praise thee? For what thou hast made, or for what thou hast not made? for those things thou hast manifested, or for those things thou hast hidden?	for those things which thou hast made? or for those things which thou hast not made? for those things which thou hast manifested, or for those things which thou hast hidden and concealed within thy self?	Cudworth, Ralph, (1617-1688.)	The true intellectual system of the universe: the first part; wherein, all the reason and philosophy of atheism is confuted; and its impossi	1678	0.85
For they lie otherwise in that which is unbodily, than in the fantasie, or to appearance. Consider him that contains all things, and understand that nothing is more capacious, than that which is incorporeal, nothing more swift, nothing more powerful, but it is most capacious, most swift, and most strong. And judge of this by thyself, command thy Soul to go into India, and sooner than thou canst bid it, it will be there.	And the ground of this Question he unfoldeth in another place thus; Consider him that contains all things, and understand, that nothing is more Capacious than that which is Incorporeal, nothing more swift, nothing more powerful: but (of all other things) it is most Capacious, most swift, and most strong. And judge of this by thy self. Command thy Soul to go into India, and sooner than thou canst bid it, it will be there.	Traherne, Thomas,	Christian ethicks: or, Divine morality. Opening the way to blessedness, by the rules of vertue and reason. By Tho. Traherne, B.D. author of the Roman	1675	0.83
Wherefore shall I praise thee, as being of myself, or having anything of mine own, or rather being anothers? For thou art what I am, thou art what I do, thou art what I say. Thou art all things, and there is nothing else thou art not.	And for what cause shall I praise thee? because I am my own, as having something proper, and distinct from thee? Thou art whatsoever I am, thou art whatsoever I do, or say, for thou art All things, and there is nothing which thou art not; thou art that which is made, and thou art that which is unmade.	Cudworth, Ralph, (1617-1688.)	The true intellectual system of the universe: the first part; wherein, all the reason and philosophy of atheism is confuted; and its impossi	1678	0.76

Figure 3. The most similar matching passages (groups of sentences) between query text and comparison corpus, excluding results from the 1657 edition of the *Divine Pymander*.

The most similar passage between the query text and the comparison corpus comes from Ralph Cudworth's *The True Intellectual System of the Universe* (1678) [25] – see **Figure 3**. Scholars have known about the influence of the *Corpus Hermeticum* on the so-called Cambridge Platonists like Ralph Cudworth and Henry More for some time, and here is lexical evidence to support this [26].

Our Text Matching tool highlights all matching words from the passages in a yellow colour, so one can see how they overlap or differ. Notice that the passages in question are not exact matches – instead, they have minor differences in language and meaning, yet the corpus passages appear to be drawn from the query text. Subject to further confirmation, we appear to be observing shades of influence. This is what we hoped to see, and there are a variety of intellectual questions that could be pursued here, with just this small sample of results.

3. Text Matching: From Mono- to Multilingual

Thus far, we have seen how the VERITRACE Text Matching tool functions when the query text and the comparison corpus share the same language. But the VERITRACE project is inherently multilingual, encompassing texts in 6 different languages. We want the ability to handle query texts and comparison texts in any of them.

Suppose, for example, we want as our query text a Latin book instead of an English one – that we want to search our English-language corpus with this Latin text. How could this work with our existing Text Matching tool? Because of the difference in languages – and hence vocabulary – it would seem impossible to find matching sentences and groups of sentences between the query and comparison texts, at least using TF-IDF.

3.1. A Simple Multilingual Corpus

In order to investigate this, let us examine a simple 3-text multilingual corpus:

TEXT 1

I. Corpus omne perseverare in statu suo quiescendi vel movendi uniformiter in directum, nisi quatenus illud a viribus impressis cogitur statum suum mutare. II. Mutationem motus proportionalem esse vi motrici impressæ, et fieri secundum lineam rectam qua vis illa imprimitur. III. Actioni contrariam semper et æqualem esse reactionem: sive corporum duorum actiones in se mutuo semper esse æquales et in partes contrarias dirigi.

TEXT 2

The quick brown fox jumps over the lazy dog. This sentence is commonly used as a typing exercise because it contains every letter of the English alphabet. It has nothing to do with Isaac Newton.

TEXT 3

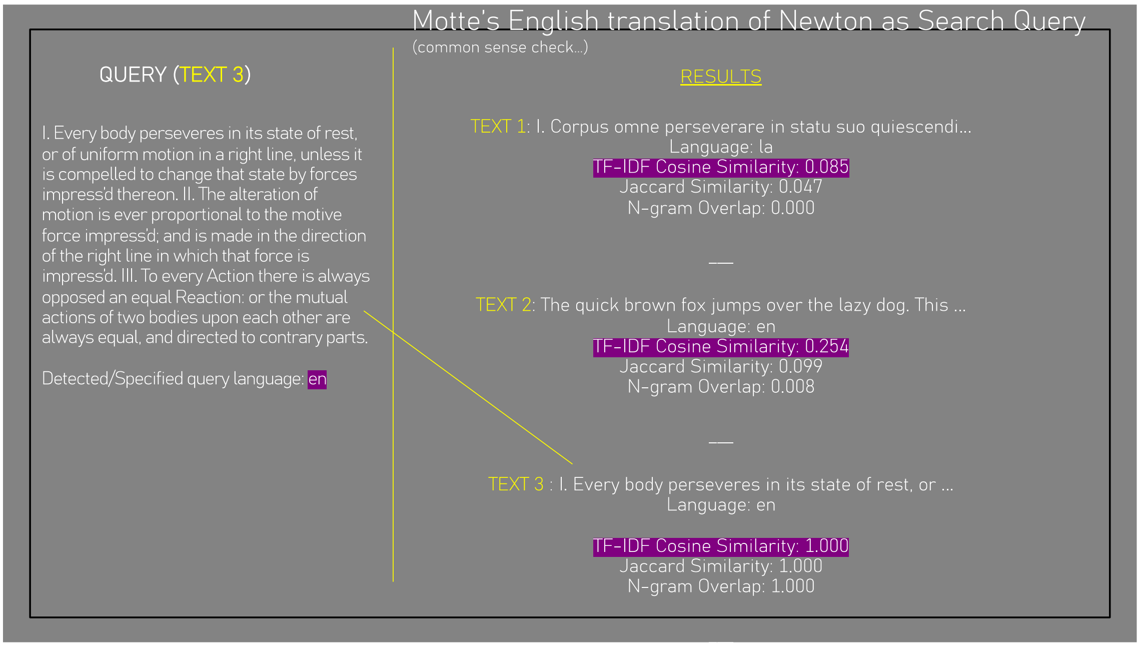
I. Every body perseveres in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impress'd thereon. II. The alteration of motion is ever proportional to the motive force impress'd; and is made in the direction of the right line in which that force is impress'd. III. To every Action there is always opposed an equal Reaction: or the mutual actions of two bodies upon each other are always equal, and directed to contrary parts.

TEXT 1, in Latin, includes Isaac Newton's three laws of motion (printed artificially together) as found in the third edition of his *Philosophiae naturalis principia mathematica* (1726) [27]. TEXT 3, in English, is Andrew Motte's 1729 translation of those very same Latin passages [28]. The assumption here is that the Motte translation is, more or less, a typical English translation of a Latin text that one finds in the early modern period and therefore ought to be representative of some of the multilingual editions we have in the VERITRACE corpus. And, finally, TEXT 2, in English, is a dummy passage, often used as a baseline in natural language processing to check our assumptions about the effect of various NLP transformations, with a few extra lines added by VERITRACE, so it resembles a passage (a group of sentences) instead of a single sentence.

3.2. Some Queries

Let us try some queries, but before we examine the results, please take note of the three different measures of similarity we are using for illustrative purposes here (see **Figure 4** below). Because we use **TF-IDF-based cosine similarity** (i.e. cosine similarity between vectors created through TF-IDF-based vectorisation) in our Text Matching, we have highlighted that metric as the one to focus on. But each metric measures slightly different kinds of similarity. **Jaccard similarity**, our second metric, is a term-based similarity measure, evaluated "as the number of shared terms over the number of all unique terms in both strings" [29, p. 14]. Finally, our **N-gram Overlap** metric is simply a variation on Jaccard Similarity, but it uses the Jaccard metric on n-grams (bigrams of words, by default) rather than single characters, so it has a wider span. Longest Common Substring (LCS) and Levenshtein distance measures would have been natural here too, but for the sake of simplicity, we have chosen just three. Now, to the queries.

First, as a 'gut check' of our assumptions, we will use TEXT 3, the Motte English translation, as our search query across the 3-text corpus. Here are the results:



Motte's English translation of Newton as Search Query

(common sense check...)

QUERY (TEXT 3)

I. Every body perseveres in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impress'd thereon. II. The alteration of motion is ever proportional to the motive force impress'd; and is made in the direction of the right line in which that force is impress'd. III. To every Action there is always opposed an equal Reaction: or the mutual actions of two bodies upon each other are always equal, and directed to contrary parts.

Detected/Specified query language: en

RESULTS

TEXT 1: I. Corpus omne perseverare in statu suo quiescendi...
Language: la
TF-IDF Cosine Similarity: 0.085
Jaccard Similarity: 0.047
N-gram Overlap: 0.000

TEXT 2: The quick brown fox jumps over the lazy dog. This ...
Language: en
TF-IDF Cosine Similarity: 0.254
Jaccard Similarity: 0.099
N-gram Overlap: 0.008

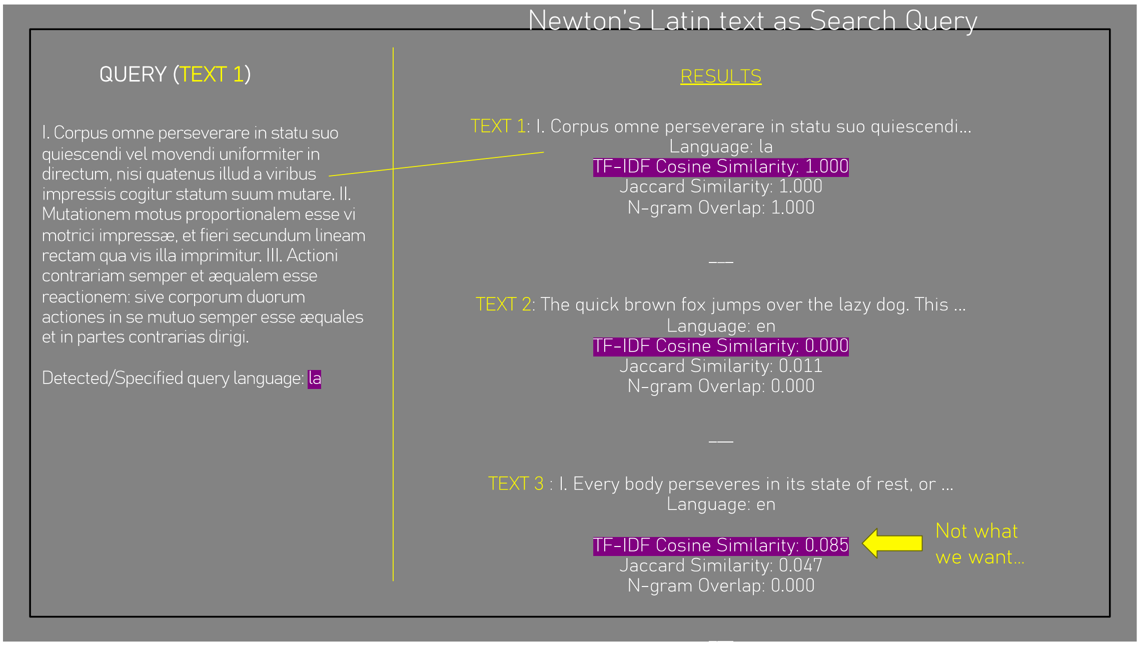
TEXT 3 : I. Every body perseveres in its state of rest, or ...
Language: en
TF-IDF Cosine Similarity: 1.000
Jaccard Similarity: 1.000
N-gram Overlap: 1.000

Figure 4. Motte's English translation of Newton's Latin (TEXT 3) is used as a search query across the 3-text multilingual corpus. The results agree with common sense.

First, as we expect, when TEXT 3 is the search query, it should match identically with TEXT 3 in the corpus – as it does.

We do not expect, nor do we find, much similarity with TEXT 2, the dummy text (only 0.254). And TEXT 1, in Latin, is also very dissimilar to our query text, given the differences in language and vocabulary. Again, this is what we expect, and it looks like our small case study confirms common sense assumptions so far.

What we want to explore, however, is using a Latin text as our query to search across our comparison corpus. Let's do that now:



Newton's Latin text as Search Query

QUERY (TEXT 1)

I. Corpus omne perseverare in statu suo quiescendi vel movendi uniformiter in directum, nisi quatenus illud a viribus impressis cogitur statum suum mutare. II. Mutationem motus proportionalem esse vi motrici impressæ, et fieri secundum lineam rectam qua vis illa imprimitur. III. Actioni contrariam semper et æqualem esse reactionem: sive corporum duorum actiones in se mutuo semper esse æquales et in partes contrarias dirigi.

Detected/Specified query language: la

RESULTS

TEXT 1: I. Corpus omne perseverare in statu suo quiescendi...
Language: la
TF-IDF Cosine Similarity: 1.000
Jaccard Similarity: 1.000
N-gram Overlap: 1.000

TEXT 2: The quick brown fox jumps over the lazy dog. This ...
Language: en
TF-IDF Cosine Similarity: 0.000
Jaccard Similarity: 0.011
N-gram Overlap: 0.000

TEXT 3 : I. Every body perseveres in its state of rest, or ...
Language: en

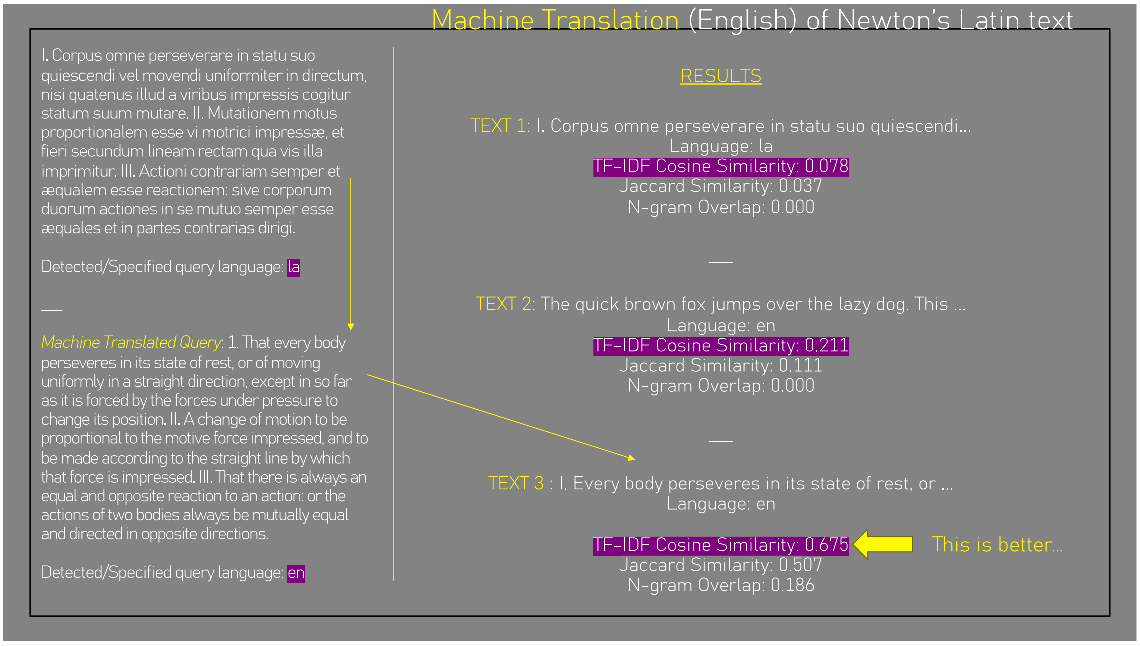
TF-IDF Cosine Similarity: 0.085
Jaccard Similarity: 0.047
N-gram Overlap: 0.000

Not what we want...

Figure 5. Newton's Latin text (TEXT 1) is used as the search query across the multilingual corpus.

When TEXT 1 (Newton's Latin) is used as our search query, it matches identically with that same text in our corpus (see **Figure 5**). And, for our dummy text (TEXT 2), there is almost no similarity, as we would expect. But the match with TEXT 3 is troublesome – there is almost no similarity here (only 0.085), even though it is an early modern English translation of the Latin query. This is not surprising given the language differences, but it is not what we want. Instead, we would prefer a much closer similarity between a text and its translation in another language, given the semantic similarity between the two. The basic problem for this task is that, up to now, we have only been comparing lexical similarity (syntactical similarity between characters or strings) – not semantic similarity (similarity between broader contextual units of meaning, regardless of specific characters, strings or words used). And, while that works well enough for monolingual text matching, it seemingly will not work across the language barrier.

So why not align the languages? We can translate the Latin search query into English and then compare it to the comparison corpus. Given the size of the VERITRACE corpus, we must do this automatically using machine translation; manual, human translation would be too time consuming. Even a few years ago, the quality of the translation simply would not have been good enough to attempt this, but it has improved rapidly since then and continues to do so. Whether it is good enough for our purposes – that is a subject for investigation. Let's try it:



Machine Translation (English) of Newton's Latin text

I. Corpus omne perseverare in statu suo quiescendi vel movendi uniformiter in directum, nisi quatenus illud a viribus impressis cogitur statum suum mutare. II. Mutationem motus proportionalem esse vi motrici impressæ, et fieri secundum lineam rectam qua vis illa imprimitur. III. Actioni contrariam semper et æqualem esse reactionem: sive corporum duorum actiones in se mutuo semper esse æquales et in partes contrarias dirigi.

Detected/Specified query language: la

—

Machine Translated Query: 1. That every body perseveres in its state of rest, or of moving uniformly in a straight direction, except in so far as it is forced by the forces under pressure to change its position. II. A change of motion to be proportional to the motive force impressed, and to be made according to the straight line by which that force is impressed. III. That there is always an equal and opposite reaction to an action: or the actions of two bodies always be mutually equal and directed in opposite directions.

Detected/Specified query language: en

RESULTS

TEXT 1: I. Corpus omne perseverare in statu suo quiescendi…
Language: la
TF-IDF Cosine Similarity: 0.078
Jaccard Similarity: 0.037
N-gram Overlap: 0.000

—

TEXT 2: The quick brown fox jumps over the lazy dog. This …
Language: en
TF-IDF Cosine Similarity: 0.211
Jaccard Similarity: 0.111
N-gram Overlap: 0.000

—

TEXT 3 : I. Every body perseveres in its state of rest, or …
Language: en

TF-IDF Cosine Similarity: 0.675 This is better…
Jaccard Similarity: 0.507
N-gram Overlap: 0.186

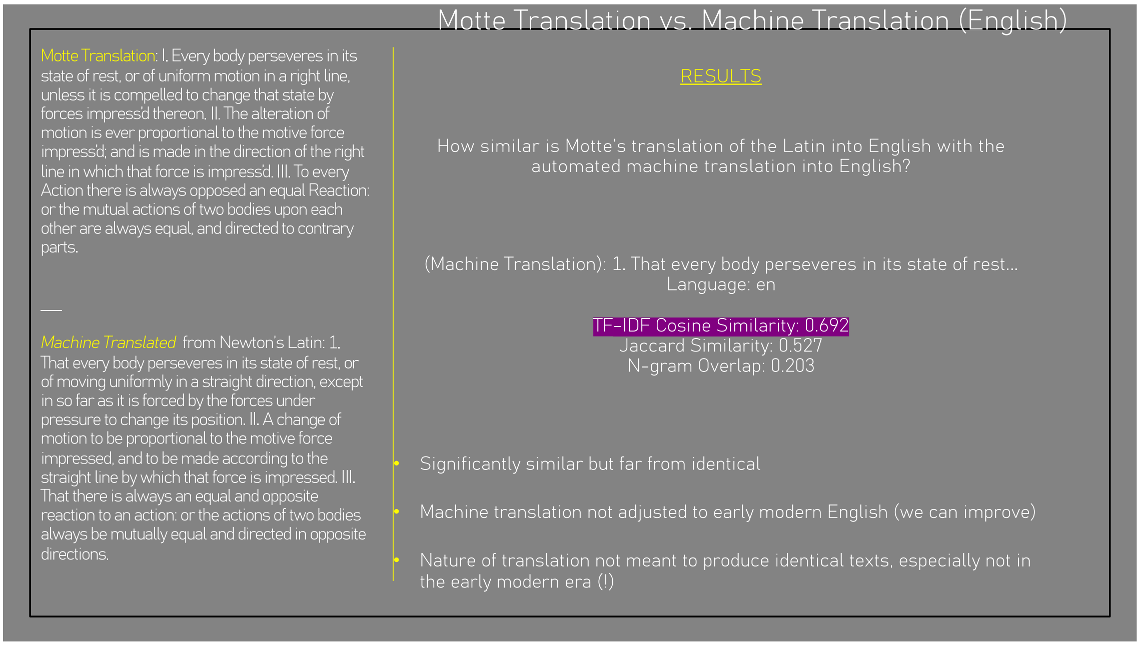
Figure 6. Newton's Latin text (TEXT 1) is machine translated into a Machine Translated Query (now in English), which is then used as the search query across the multilingual corpus.

We take Newton's Latin text (TEXT 1) and apply machine translation to it, and then send this machine-translated query (now in English) across our comparison corpus (see **Figure 6**). For automated machine translation, we used Google Translate through the 'deep-translator' library.[19] While we chose Google Translate here, we have not yet done a systematic analysis of translation tools that would best meet our needs (further optimisation is possible). In any case, notice that now, even though we began with the Newton Latin text (TEXT 1), it is no longer considered similar to TEXT 1 in the corpus because we are using its English translation. That is to be expected. The same with the similarity to the dummy text (TEXT 2). But see what happened to the comparison to TEXT 3 – the Motte translation is now considered significantly similar (0.675) to the machine-translated version of Newton's original Latin text (TEXT 1). This is quite a bit better and should allow us, in theory, to match texts that are semantically similar, even when they are in different languages.

There are some important limitations we should be aware of. The quality of the translation clearly has a big impact on the effectiveness of Text Matching across languages. And we must remember that when we create the machine translation, at least as it is set up above, we are creating a translation into 21st-century English, which can differ substantially from the early modern English found in our corpus. These differences negatively impact the effectiveness of our Text Matching. This is indicated by the low N-gram overlap between the texts. This suggests that, despite the TF-IDF cosine similarity, individual sequences of 2 consecutive words (bigrams) between the two translations are quite dissimilar. The machine translation, in other words, does not use many of the same sequences of words as the Motte translation.

For this example, we can measure the similarity between the two translations (Motte's and the machine translation):

[19] https://pypi.org/project/deep-translator/#translation-for-humans



Motte Translation vs. Machine Translation (English)

Motte Translation: I. Every body perseveres in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impress'd thereon. II. The alteration of motion is ever proportional to the motive force impress'd; and is made in the direction of the right line in which that force is impress'd. III. To every Action there is always opposed an equal Reaction: or the mutual actions of two bodies upon each other are always equal, and directed to contrary parts.

—

Machine Translated from Newton's Latin: 1. That every body perseveres in its state of rest, or of moving uniformly in a straight direction, except in so far as it is forced by the forces under pressure to change its position. II. A change of motion to be proportional to the motive force impressed, and to be made according to the straight line by which that force is impressed. III. That there is always an equal and opposite reaction to an action: or the actions of two bodies always be mutually equal and directed in opposite directions.

RESULTS

How similar is Motte's translation of the Latin into English with the automated machine translation into English?

(Machine Translation): 1. That every body perseveres in its state of rest...
Language: en

TF-IDF Cosine Similarity: 0.692
Jaccard Similarity: 0.527
N-gram Overlap: 0.203

- Significantly similar but far from identical
- Machine translation not adjusted to early modern English (we can improve)
- Nature of translation not meant to produce identical texts, especially not in the early modern era (!)

Figure 7. A similarity comparison between the machine translation of Newton's Latin text and Motte's early modern English translation (TEXT 3)

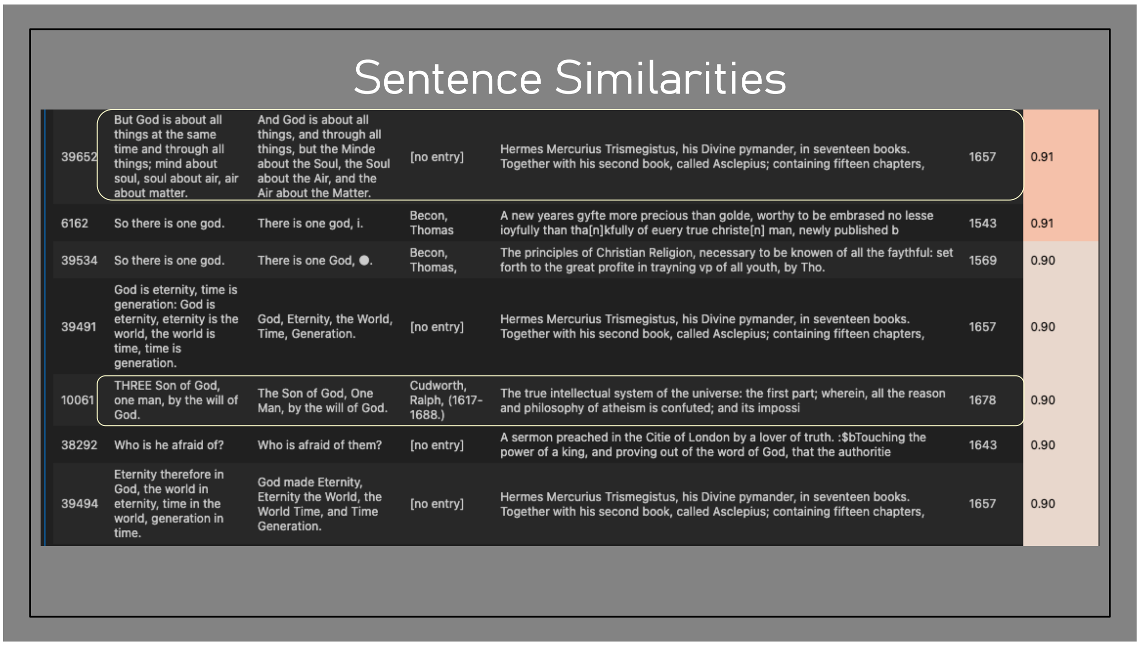
The machine translation into 21st-century English is significantly similar to the 18th-century Motte translation (0.692) but far from identical (see **Figure 7**). Indeed, even while the TF-IDF keywords are fairly similar, individual sequences of words (N-grams) are rather less so. Whether this is due more to the limitations of the machine translation of the original Latin or the lexical differences between contemporary and early modern English is unclear, but in either case, it reduces the effectiveness of the text matching. Also, we should keep in mind that the nature of translation itself – especially translation in the early modern era – is not intended to produce lexically equivalent texts but semantically similar ones, to varying degrees, sometimes rather loosely. Thus, we should not demand that the similarities between the machine translation and the original Motte translation be identical. Nonetheless, are they similar enough to produce meaningful Text Matching results? Based on our sample corpus, and the similarity scores above, we cannot know for sure. We have dramatically increased the similarity scores by introducing machine translation, which is surely in the right direction, but the only way to know if they are 'good enough' is to attempt to match some texts from our actual corpus – to test the Text Matching tool 'in the wild', with more than one language.

3.3. Multilingual Text Matching 'in the Wild'

To test this, we can re-run our original research query. Instead of using Everard's English translation of the *Divine Pymander* as our query text, we will use what may have been his original Latin source text: Marsilio Ficino's 1471 *De potestate et sapientia Dei* [30].[20] This should give us a sense of whether our results using machine translation are 'good enough'. So, now our search 'query' will be an entire text in Latin, which we then translate into English using automated

[20] While we have chosen Ficino's 1471 edition in this example, some recent work [31] points to Patrizi's 1591 translation instead [32]. In fact, the VERITRACE Close Reading Corpus, once finalised, should be able to establish this with determinative evidence because it will allow us to compare, in minute detail, all the editions used. Nonetheless, for our purposes here, the Ficino source text should be sufficient, even if we must revise our assumption later.

machine translation, before searching across our comparison corpus, which remains the one we used with our original Everard example – our corpus of primarily 17th-century English-language texts. If this approach is sound, we should get fairly similar results to what we found when we used Everard as our query text. We should not demand the very same results, however, given the nature of translation and some of the limitations of the approach we explained above. But do we at least obtain 'fairly similar' ones?



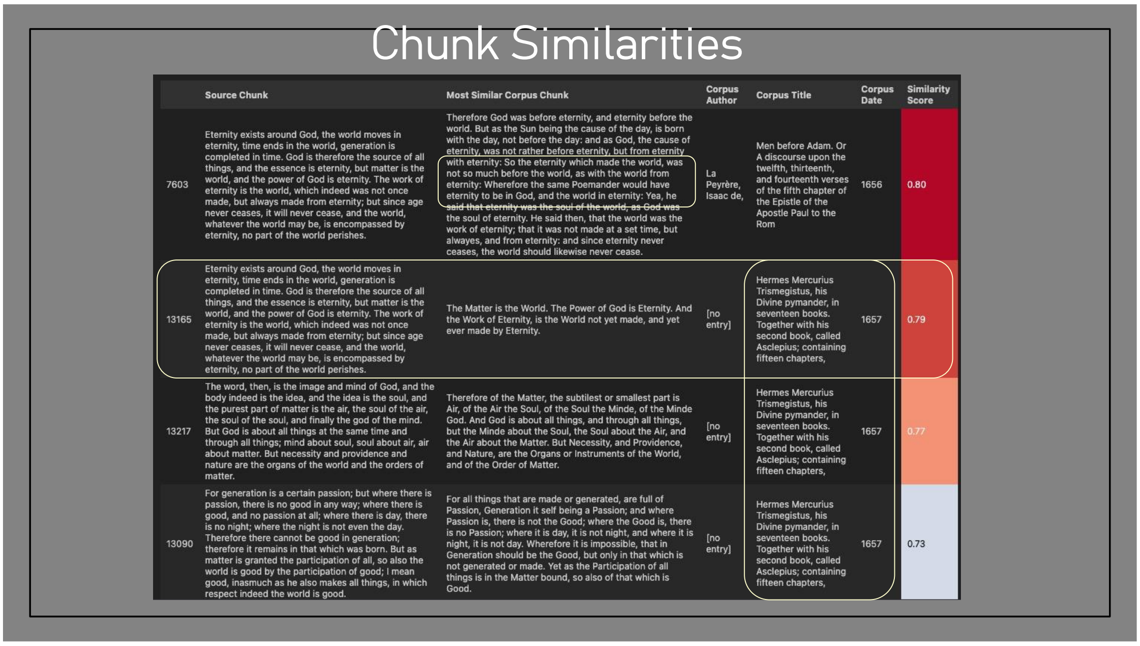
Sentence Similarities

39652	But God is about all things at the same time and through all things; mind about soul, soul about air, air about matter.	And God is about all things, and through all things, but the Minde about the Soul, the Soul about the Air, and the Air about the Matter.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	0.91
6162	So there is one god.	There is one god, i.	Becon, Thomas	A new yeares gyfte more precious than golde, worthy to be embrased no lesse ioyfully than tha[n]kfully of euery true christe[n] man, newly published b	1543	0.91
39534	So there is one god.	There is one God, ●.	Becon, Thomas,	The principles of Christian Religion, necessary to be knowen of all the faythful: set forth to the great profite in trayning vp of all youth, by Tho.	1569	0.90
39491	God is eternity, time is generation: God is eternity, eternity is the world, the world is time, time is generation.	God, Eternity, the World, Time, Generation.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	0.90
10061	THREE Son of God, one man, by the will of God.	The Son of God, One Man, by the will of God.	Cudworth, Ralph, (1617-1688.)	The true intellectual system of the universe: the first part; wherein, all the reason and philosophy of atheism is confuted; and its impossi	1678	0.90
38292	Who is he afraid of?	Who is afraid of them?	[no entry]	A sermon preached in the Citie of London by a lover of truth. :$bTouching the power of a king, and proving out of the word of God, that the authoritie	1643	0.90
39494	Eternity therefore in God, the world in eternity, time in the world, generation in time.	God made Eternity, Eternity the World, the World Time, and Time Generation.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	0.90

Figure 8. The most similar sentences matched between the machine-translated Ficino query text and the comparison corpus of English-language texts.

In **Figure 8** we find the most similar sentences to the machine-translated Ficino query text. It is a relief to see that we find some familiar results: the 1657 *Divine Pymander* has the most similar sentence (a long, meaningful one) and Ralph Cudworth appears as well. But, at the same time, we are not getting the exact same results that we did when Everard was the query text. The results are similar but not identical.

Now what about the most similar sentence groups (chunks) (see next page)?



Chunk Similarities

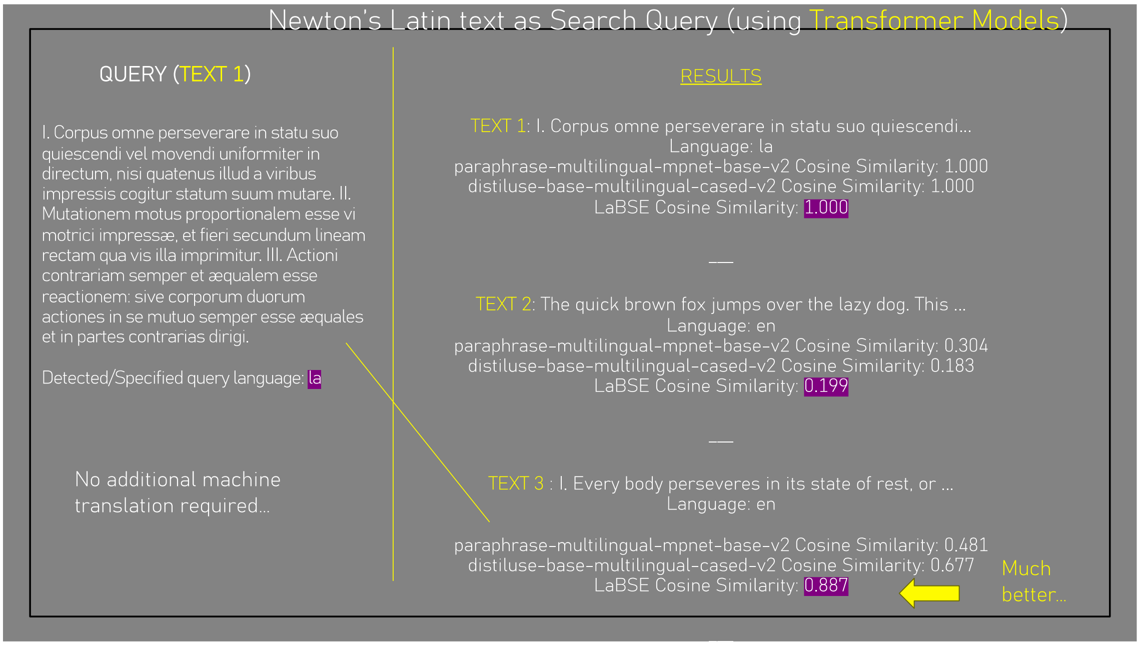
	Source Chunk	Most Similar Corpus Chunk	Corpus Author	Corpus Title	Corpus Date	Similarity Score
7603	Eternity exists around God, the world moves in eternity, time ends in the world, generation is completed in time. God is therefore the source of all things, and the essence is eternity, but matter is the world, and the power of God is eternity. The work of eternity is the world, which indeed was not once made, but always made from eternity; but since age never ceases, it will never cease, and the world, whatever the world may be, is encompassed by eternity, no part of the world perishes.	Therefore God was before eternity, and eternity before the world. But as the Sun being the cause of the day, is born with the day, not before the day: and as God, the cause of eternity, was not rather before eternity, but from eternity with eternity: So the eternity which made the world, was not so much before the world, as with the world from eternity: Wherefore the same Poemander would have eternity to be in God, and the world in eternity: Yea, he said that eternity was the soul of the world, as God was the soul of eternity. He said then, that the world was the work of eternity; that it was not made at a set time, but alwayes, and from eternity: and since eternity never ceases, the world should likewise never cease.	La Peyrère, Isaac de,	Men before Adam. Or A discourse upon the twelfth, thirteenth, and fourteenth verses of the fifth chapter of the Epistle of the Apostle Paul to the Rom	1656	0.80
13165	Eternity exists around God, the world moves in eternity, time ends in the world, generation is completed in time. God is therefore the source of all things, and the essence is eternity, but matter is the world, and the power of God is eternity. The work of eternity is the world, which indeed was not once made, but always made from eternity; but since age never ceases, it will never cease, and the world, whatever the world may be, is encompassed by eternity, no part of the world perishes.	The Matter is the World. The Power of God is Eternity. And the Work of Eternity, is the World not yet made, and yet ever made by Eternity.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	0.79
13217	The word, then, is the image and mind of God, and the body indeed is the idea, and the idea is the soul, and the purest part of matter is the air, the soul of the air, the soul of the soul, and finally the god of the mind. But God is about all things at the same time and through all things; mind about soul, soul about air, air about matter. But necessity and providence and nature are the organs of the world and the orders of matter.	Therefore of the Matter, the subtilest or smallest part is Air, of the Air the Soul, of the Soul the Minde, of the Minde God. And God is about all things, and through all things, but the Minde about the Soul, the Soul about the Air, and the Air about the Matter. But Necessity, and Providence, and Nature, are the Organs or Instruments of the World, and of the Order of Matter.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	0.77
13090	For generation is a certain passion; but where there is passion, there is no good in any way; where there is good, and no passion at all; where there is day, there is no night; where the night is not even the day. Therefore there cannot be good in generation; therefore it remains in that which was born. But as matter is granted the participation of all, so also the world is good by the participation of good; I mean good, inasmuch as he also makes all things, in which respect indeed the world is good.	For all things that are made or generated, are full of Passion, Generation it self being a Passion; and where Passion is, there is not the Good; where the Good is, there is no Passion; where it is day, it is not night, and where it is night, it is not day. Wherefore it is impossible, that in Generation should be the Good, but only in that which is not generated or made. Yet as the Participation of all things is in the Matter bound, so also of that which is Good.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	0.73

Figure 9. The most similar sentence groups (chunks) matched between the machine-translated Ficino query text and the comparison corpus of English-language texts.

Here again, it would be worrisome – and a sign of the deficiencies of our approach – if most of these results did not come from the 1657 *Divine Pymander*. In fact, except for the top matching result above (more on that in a moment), the top 10 matches all come from the 1657 edition – just as it did when Everard was our query text (see **Figure 9** for the top 4 results). And what about that top result? It is the exception that proves the rule for, if one looks closely, one can see that Mr. La Peyrère is paraphrasing directly from the *Divine Pymander* ('Wherefore the same Poemander would have eternity to be in God, and the world in eternity…') [33].

These results, we believe, are enough to show proof of concept for a multilingual approach using automated machine translation. They appear to be 'good enough' to obtain meaningful and reliable results. And this is before we refine it. For instance, we can include some transformation rules for each translation such that the resulting translation is closer in syntax to its early modern variant (in whatever language), instead of relying on 21st-century vocabulary and syntax. Or we could include a 'Translation Matrix' that maps translated terms between source and target languages. Both steps should bump up the similarity scores. We can also cross-reference our Text Matching results to some of the more corpus-based measures we are using in VERITRACE (not discussed in this paper), including Latent Semantic Analysis and Latent Dirichlet Allocation, which have been used on multilingual corpora before [34, 35].

Automated machine translation is not our only option, however. There are exciting possibilities with the latest multilingual transformer models. In fact, we can obtain even more accurate results – without the extra step of automated translation – by using some of these on our corpus. This is not without trade-offs – as we mention below – but there are great opportunities here. To illustrate this, let's return to our simple 3-text corpus and use some multilingual transformer models to compare the Latin search query to the corpus:



Newton's Latin text as Search Query (using Transformer Models)

QUERY (TEXT 1)

I. Corpus omne perseverare in statu suo quiescendi vel movendi uniformiter in directum, nisi quatenus illud a viribus impressis cogitur statum suum mutare. II. Mutationem motus proportionalem esse vi motrici impressæ, et fieri secundum lineam rectam qua vis illa imprimitur. III. Actioni contrariam semper et æqualem esse reactionem: sive corporum duorum actiones in se mutuo semper esse æquales et in partes contrarias dirigi.

Detected/Specified query language: la

No additional machine translation required…

RESULTS

TEXT 1: I. Corpus omne perseverare in statu suo quiescendi…
Language: la
paraphrase-multilingual-mpnet-base-v2 Cosine Similarity: 1.000
distiluse-base-multilingual-cased-v2 Cosine Similarity: 1.000
LaBSE Cosine Similarity: 1.000

—

TEXT 2: The quick brown fox jumps over the lazy dog. This …
Language: en
paraphrase-multilingual-mpnet-base-v2 Cosine Similarity: 0.304
distiluse-base-multilingual-cased-v2 Cosine Similarity: 0.183
LaBSE Cosine Similarity: 0.199

—

TEXT 3 : I. Every body perseveres in its state of rest, or …
Language: en

paraphrase-multilingual-mpnet-base-v2 Cosine Similarity: 0.481
distiluse-base-multilingual-cased-v2 Cosine Similarity: 0.677
LaBSE Cosine Similarity: 0.887

Figure 10. Newton's Latin text (TEXT 1) is used as a search query across the multilingual corpus and processed using a set of multilingual transformer models – instead of TF-IDF.

Here we send Newton's Latin text (TEXT 1) as our search query across the corpus – without performing any additional machine translation first. The multilingual transformer models have been trained on many languages and have no trouble handling multilingual texts. We have provided three of these for comparison, but the LaBSE model appears to perform best in this instance, so we have highlighted its results above (see **Figure 10**). LaBSE is a language-agnostic BERT sentence embedding model that supports 109 languages out of the box.[21] LaBSE was originally developed by Google. It "is trained and optimized to produce similar representations exclusively for bilingual sentence pairs that are translations of each other. So, it can be used for mining for translations of a sentence in a larger corpus."[22] As we expect, the TEXT 1 query is identical to itself, and the dummy English TEXT 2 is not very similar to the Latin query. But the results for TEXT 3 should make us sit up and take notice: the LaBSE model provides a cosine similarity score of 0.887 – a significantly higher score than we were able to achieve using automated translation. The model appears to detect the deep semantic similarity between the Latin query and its English translation. This is very much what we had hoped for when we began this investigation. If this kind of result were to hold up across all 6 languages, then we have found a very powerful tool indeed.

The VERITRACE team will therefore explore the use of multilingual transformer models for use with our Text Matching tool. But we should also be cautious for there are definite trade-offs in using these new tools. They require significantly more computational resources, add complexity and are much less interpretable. We cannot yet understand how these models come to the conclusions they do, nor can we consistently reproduce the same results. There is an element of indeterminism in their method, which makes reproducible research much harder to achieve [36]. Still, automated machine translation is also built on the advances of transformer models, so once we introduce this into our project, we must confront this 'black box' technology.

[21] https://huggingface.co/sentence-transformers/LaBSE

[22] https://www.kaggle.com/models/google/labse/tensorFlow2/labse/1?tfhub-redirect=true

This is the direction NLP has been headed in the past few years, and it would be obstinate to ignore these developments entirely.

A final point about capabilities: we are just scratching the surface of what one can do, even with our more traditional tools. For Text Matching, we have only used one query text at a time, but we can do so for our entire Close Reading Corpus or a subset of that, e.g. all editions of the *Corpus Hermeticum*. We could then look for matches and similarities to this larger source collection. We also will use our entire multilingual VERITRACE text collection of c.430,000 texts as the comparison corpus – not the 18,633 English texts we limited ourselves to for this specific research question. Using many query and comparison texts will, of course, dramatically increase the computational demands of the Text Matching task, and there may be a practical limit here. Whether this is at 100 or 1000 or 100,000 texts, we have yet to explore. The VERITRACE team must also consider whether the increased accuracy of transformer models is worth the trade-offs and complexity they bring with them.

4. Next Steps: A Two-Pronged Approach to Text Matching

What are the lessons for VERITRACE in this small, multilingual case study? To conclude this paper, we outline our current thinking about what it means for the creation of a more robust version of multilingual text matching.

The VERITRACE Text Matching tool should be able to measure lexical similarity between highly similar words or phrases (irrespective of their 'semantic' meaning) from different texts. If a text from the comparison corpus, for example, repeats verbatim a passage from the query text, no matter what model we use this should generate a similarity score of 1. And if two sentences or passages are very similar from a lexical standpoint – they tend to use many of the same words, even with different semantic meanings – that too will show up as high lexical similarity. This is, again, a sort of plagiarism detector, at least of the simplistic kind, where one author simply re-uses exact or very similar words and phrases from another work or set of works. And for this matching task, using TF-IDF and cosine similarity for identifying similar texts is a great choice because TF-IDF (and its variations, like Okapi BM25) is an effective, surface-level tool, focused on a text's most important keywords and vocabulary.

There are drawbacks, however, to using TF-IDF-based cosine similarity alone. If we limit ourselves to that, we are likely to get some false positives, passages that appear to overlap based on vocabulary and lexical similarity but have different meanings. Consider the pair of sentences: "He couldn't desert his post at the plant" and "Desert plants can survive droughts" [37]. They do not mean remotely the same thing – they have vastly different semantic meanings – but they share some of the same key vocabulary. They would likely receive a high similarity score, if matched on lexical similarity alone.

There would undoubtedly be false negatives as well, if we consider passages that have been extensively paraphrased. One can imagine a more sophisticated plagiariser – continuing our metaphor – who changes most of the words from a borrowed passage but keeps the sense and meaning of it. A simple example drawn from [38]: Consider "Peter is a handsome boy" and "Peter is a good-looking lad." They are arguably quite close in semantic meaning, but their keywords (beyond the proper name) do not overlap and would therefore not be identified as similar using TF-IDF. Now, we are not trying to identify early modern plagiarism per se, but traces of influence between ancient wisdom texts and natural philosophical discourse. But the need is the same: to be able to identify both lexical and deep semantic similarity between query and comparison texts. Therefore, if we want to avoid too many false positives and false negatives, we cannot limit ourselves to TF-IDF alone. To capture paraphrasing, we need a nuanced semantic similarity tool, like what transformer models seemingly supply.

And this is just in terms of monolingual text matching. If we want to capture any sort of cross-lingual similarity at all, we also need an effective semantic similarity tool for that as well. Indeed, at a minimum, this is what our case study above has demonstrated.

That means our Text Matching Tool ought to have a two-pronged approach – it needs to be able to capture and identify, on one end of the spectrum, surface-level, lexical similarity. And at the other end, deep, contextual, semantic similarity. TF-IDF-based cosine similarity is therefore likely to be our tool of choice for the lexical similarity metric. For the other prong, we can use a multilingual transformer model, like LaBSE (and cosine similarity), to capture deep semantic similarity. Both can be used together, as complements, for monolingual text matching, but multilingual text matching will lean more heavily on the latter by using a multilingual transformer model.[23]

This two-pronged approach is becoming standard, in fact, with the latest iterations of vector search. Vector database software vendors, for instance, already advertise sparse (e.g. TF-IDF) vs. dense search (e.g. using transformers), as well as hybrid search, which combines the two result sets.[24] Multilingual hybrid search is, unfortunately, harder to find.[25] In any case, whether we rely on an implementation using open-source vector database software, or customise some variation of our own, pursuing this general hybrid approach for Text Matching is a logical next step.

In short, the capabilities of VERITRACE will be expanded significantly, as we proceed, and we look forward to sharing our results with the academic community.

[23] An interesting question is: does it make sense to consider lexical similarity in a multilingual context? What would that mean? This is beyond the scope of this paper, but a promising line of investigation is found in work done on cross-lingual plagiarism detection. For languages that are not too dissimilar lexically, character n-gram vectors have been tried [39]. More recently, attempts have been made using multilingual word clusters or sets of multilingual thesauri, combined with automated translation [40, 41].

[24] https://weaviate.io/blog/hybrid-search-explained

[25] Weaviate, for instance, offers multilingual semantic search but not lexical or keyword search in anything other than English: https://weaviate.io/blog/weaviate-non-english-languages

Acknowledgements

VERITRACE would like to acknowledge funding from the European Research Council as part of ERC-StG Project VERITRACE (101076836). We would also like to thank Dr. Klaus Ceynowa, the Director General of the Bayerische Staatsbibliothek, for his support, as well as the entire staff of the Munich Digitisation Centre. VERITRACE also thanks: Dr. Arthur der Weduwen, the Co-Director and Project Manager of the Universal Short Title Catalogue (USTC: https://www.ustc.ac.uk) for allowing us access to raw USTC database files; Doug Knox, Martin Mueller, Craig Berry, Joseph Loewenstein, and Anupam Basu of the EarlyPrint Project (https://earlyprint.org) for helpful advice right at the start of our project; Paul F. Schaffner, Manager of the Text Creation Partnership (EEBO-TCP) and Senior Associate Librarian, Digital Content & Collections, at the University of Michigan Library, for clear explanations of the EEBO dataset; Samuel Pizelo and Arthur Koehl, from Project Quintessence (http://quintessence.ds.lib.ucdavis.edu) at the UC Davis Data Lab, who shared their approach to managing data; and, finally, audience feedback on our project at panel presentations at the Scientiae Conference in Brussels (June 2024), the European Society for the History of Science in Barcelona (Sept. 2024), and the 4th Humanities-Centred A.I. Workshop in Würzburg (Sept. 2024).

References

[1] VERITRACE project website. URL: https://veritrace.eu.

[2] C. J. Schilt, Traces de la Verité: The reappropriation of ancient wisdom in early modern natural philosophy, VERITRACE (ERC-2022-STG-101076836), 2022. URL: https://veritrace.eu/wp-content/uploads/2023/04/Project-Traces-de-la-Verite-Condensed.pdf.

[3] J. C. Wolf, From Data Acquisition to Latent Semantic Analysis: Developing VERITRACE's Computational Approach to Tracing the Influence of Ancient Wisdom in Early Modern Philosophy, Society and Politics 18(1:35) (April 2024, *forthcoming* 2025).

[4] M. Cohen, Narratology in the Archive of Literature, Representations 108 (2009).

[5] D. Reid, Distant Reading, 'the Great Unread', and 19th-Century British Conceptualizations of the Civilizing Mission: A Case Study, Journal of Interdisciplinary History of Ideas 15 (2019).

[6] M. J. Hill, S. Hengchen, Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study, Digital Scholarship in the Humanities 34, no. 4 (2019).

[7] P. Kurhekar, S. Nigam, S. Pillai, Automated Text and Tabular Data Extraction From Scanned Document Images, Data Management, Analytics and Innovation, in: Proceedings of ICDMAI 2021, 1(2021), 169-182.

[8] K. Imai, Quantitative Social Science: An Introduction, Princeton University Press, Princeton and Oxford, 2018.

[9] F. Karsdorp, M. Kestemont, A. Riddell, Humanities Data Analysis: Case Studies with Python, Princeton University Press, Princeton and Oxford, 2021.

[10] H. Trismegistus, The divine Pymander of Hermes Mercurius Trismegistus, in XVII. books. Translated formerly out of the Arabick into Greek, and thence into Latine, and Dutch, and now out of the original into English; by that learned divine Doctor Everard. Printed by

Robert White, London, [1650]. A digital transcription of this text can be found online. URL: https://sacred-texts.com/eso/pym/index.htm.

[11] M. P. Oakes, Author Profiling and Related Applications, in: R. Mitkov (ed.) Oxford Handbook of Computational Linguistics, 2nd ed., Oxford University Press, Oxford, UK, 2022, pp. 1165-1197.

[12] H. Trismegistus, Hermes Mercurius Trismegistus his Divine pymander in seventeen books: together with his second book called Asclepius, containing fifteen chapters with a commentary / translated formerly out of the Arabick into Greek, and thence into Latine, and Dutch, and now out of the original into English by Dr. Everard. Printed by J.S. for Thomas Brewster, London, 1657.

[13] D. Jurafsky, J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released August 20, 2024. URL: https://web.stanford.edu/~jurafsky/slp3.

[14] C. D. Manning, R. Prabhakar, H. Schütze, An Introduction to Information Retrieval. Cambridge Online Edition, Cambridge University Press, Cambridge, UK, 2009.

[15] R. Mitkov (ed.), Oxford Handbook of Computational Linguistics, 2nd ed., Oxford University Press, Oxford, UK, 2022.

[16] O. Levy, Word Representation, in: R. Mitkov (ed.) Oxford Handbook of Computational Linguistics, 2nd ed., Oxford University Press, Oxford, UK, 2022, pp. 334-358.

[17] R. Mihalcea, S. Hassan, Similarity, in: R. Mitkov (ed.) Oxford Handbook of Computational Linguistics, 2nd ed., Oxford University Press, Oxford, UK, 2022, pp. 415-434.

[18] N. Indurkhya, F. J. Damerau (Eds.), Handbook of Natural Language Processing, 2nd ed., Chapman and Hall/CRC, Boca Raton, FL, 2010.

[19] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, O'Reilly Media, Inc., Sebastopol, CA, 2009.

[20] A. Wangperawong, Multilingual Search with Subword TF-IDF, arXiv, 29 Sept 2022.

[21] G. Salton, The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice Hall, Upper Saddle River, NJ, 1971.

[22] C. E. Osgood, G. J. Suci, P.H. Tannenbaum, The Measurement of Meaning, University of Illinois Press, Urbana, IL, 1957.

[23] T. Traherne, Christian ethicks, or, Divine morality opening the way to blessedness, by the rules of vertue and reason, London: Printed for Jonathan Edwin, 1675. URL: https://name.umdl.umich.edu/A63047.0001.001

[24] C. Marks, Thomas Traherne and Hermes Trismegistus, Renaissance News, 19/2 (1966): pp. 118-131

[25] R. Cudworth, The true intellectual system of the universe. The first part wherein all the reason and philosophy of atheism is confuted and its impossibility demonstrated, London: Printed for Richard Royston, 1678. URL: https://name.umdl.umich.edu/A35345.0001.001

[26] D. P. Walker, The Ancient Theology: Studies in Christian Platonism from the Fifteenth to the Eighteenth Century, Cornell University Press, Ithaca, NY, 1972.

[27] I. Newton, Philosophiae naturalis principia mathematica. Editio tertia aucta & emendate, London: Apud Guil. & Joh. Innys, 1726.

[28] A. Motte (ed.), The mathematical principles of natural philosophy. By Sir Isaac Newton. Translated into English by Andrew Motte. To which are added, The laws of the moon's

motion, according to gravity. By John Machin ... In two volumes [volume 1], London: Printed for Benjamin Motte, 1729.

[29] W. H. Gomaa, A. A. Fahmy, A Survey of Text Similarity Approaches, International Journal of Computer Applications, 68(13), pp. 13-18. https://doi.org/10.5120/11638-7118

[30] M. Ficino, De potestate et sapientia Dei, Treviso: Gerardus de Lisa, 1471.

[31] W. J. Hanegraaff, A Suggestive Inquiry into Hermetic Rebirth: Nondual Noēsis and Bodily Fluids in Victorian England, in: S. Perez, B. van Rijn, J. Schlieter (Eds,), Intentional Transformative Experiences, De Gruyter, Berlin, pp.149-178.

[32] F. Patrizi, Nova de universis philosophia, Ferraro: Apud Benedictum Mammarellum, 1591.

[33] I. La Peyrère, Men before Adam. Or a discourse upon the twelfth, thirteenth, and fourteenth verses of the fifth chapter of the Epistle of the Apostle Paul to the Romans. By which are prov'd, that the first men were created before Adam. [A theological systeme upon that presupposition, that men were before Adam. The first part.], London : Leach, F., 1656.

[34] A.A.P. Ratna, P. D. Purnamasari, B. A. Adhi, F. A. Ekadiyanto, M. Salman, M. Mardiyah, D. J. Winata. Cross-Language Plagiarism Detection System Using Latent Semantic Analysis and Learning Vector Quantization, Algorithms 10(69), 2017. https://doi.org/10.3390/a10020069

[35] T. K. Landauer, D.S. McNamara, S. Dennis, W. Kintsch (Eds.), Handbook of Latent Semantic Analysis, Psychology Press, New York, NY, 2007.

[36] J.E. Dobson, Interpretable Outputs: Criteria for Machine Learning in the Humanities, Digital Humanities Quarterly, 15 (2), 2021.

[37] A. Ng, N. Namjoshi, Understanding and applying text embeddings [MOOC], DeepLearning.AI, URL: https://www.deeplearning.ai/short-courses/google-cloud-vertex-ai/

[38] R. Ferreira, R.D. Lins, S. J. Simske, F. Freitas, M. Riss, Assessing sentence similarity through lexical, syntactic and semantic analysis, Computer Speech and Language 39 (2016) 1–28.

[39] P. McNamee, J. Mayfield, Character n-gram tokenization for European language text retrieval, Information Retrieval 7 (2004) 73–97.

[40] M. Potthast, A. Eiselt, L. A. Barrón-Cedeño, B. Stein, P. Rosso, Overview of the 3rd international competition on plagiarism detection, in: CEUR workshop proceedings, Vol. 1177, CEUR Workshop Proceedings, 2011.

[41] K. Avetisyan, A. Malajyan, T. Ghukasyan, A. Avetisyan, A Simple and Effective Method of Cross-Lingual Plagiarism Detection, arXiv, 5 April 2023.