DIGITAL CLASSICS ONLINE

Pain and the Body in *Corpus Hippocraticum*: A Distributional Semantic Analysis

Vojtěch Linka, Vojtěch Kaše

Abstract: The authors of the medical treatises collected in *Corpus Hippocraticum* often mention pain, its qualities and origin. At the same time, however, they do not provide any explicit definition or theory of pain, of its nature and of relation to other important aspects of Hippocratic medicine. Moreover, they employ at least four word-families which are commonly suggested to denote pain in ancient Greek. This encourages modern researchers to ask how do these four pain-words semantically differ and to what extent are they based on a shared notion of pain. In this article, we attempt to answer these questions by analysing the corpus employing several computational text analysis methods, especially by employing a distributional semantic modelling approach. Our results reveal a close association between some of these pain-words, bodily parts and pathological states. The results are further compared with findings obtained through the traditional close reading of the sources.

Introduction[1]

Alleviating pain and taking away the cause of suffering is one of the maxims of Hippocratic medicine.[2] The authors of Hippocratic writings often mention pain; we read about its quality, location, causes and relation to illness. What, though, is pain? There is no explicit definition of pain in *Corpus Hippocraticum* (= *CH*); there is no treatise on its nature. While some scholars assume there might be a unified conception of pain in *CH*,[3] other researchers boldly claim there is no such thing.[4] In this article, we evaluate these propositions by combining insights obtained by the application of some computational text analyses methods on the corpus as a whole (i.e. *distant reading*), with insights based on the detailed interpretation of selected passages from the corpus (i.e. *close reading*).[5]

In particular, we focus on the semantics of a selection of words which are commonly suggested to denote pain in *CH*, namely four word-families: πόνο*, ὀδύν*, ἄλγ*, λύπ*. We are especially interested in the typical contexts in which these pain-words occur, how these words are related to each other and what other words are most closely semantically associated with them.

We begin with the Materials and Methods section, in which we offer an overview of the data we use for the computational text analysis parts of the article. We describe the procedures which were conducted to make the texts suitable for these analyses and subsequently introduce all methodological steps, especi-

1 This work was financially supported by Charles University Grant Agency, project no. 78120, entitled "Aristotle and Hippocrates on Pain", implemented at the Faculty of Arts of Charles University.

2 Hippocr. *Med. Vet*. 3.35–40; *Vict*. 15.5–6; *De arte* 3.5.

3 King (1999), 269–286.

4 Horden (1999), 295–315.

5 The phrases distant reading and close reading were coined by Franco Moretti and have since come to be used widely in digital humanities literature. See Moretti (2013); Underwood (2017), 1–12; Jänicke et al. (2015), 1–21.

ally concerning the Distributional Semantic analysis. We continue with the Results section which builds on these methods. Finally, in the Discussion, we evaluate the results obtained by the computational text analysis methods against observations based on the close reading of individual texts and passages from the corpus.

Materials and Methods

Textual corpus and its preprocessing

The computational text analyses included in this article are based on a corpus of Hippocratic texts retrieved from the Lemmatised Ancient Greek Texts dataset (LAGT).[6] LAGT combines two open-source corpora of ancient Greek texts: the Canonical Greek Literature dataset from the Perseus Digital Library[7] and the First Thousand Years of Greek dataset of the Open Greek & Latin project.[8] Both datasets are publicly available on Github and Zenodo under Creative Commons Attribution licences, which makes them suitable for further reuse within any large-scale computational text analysis project[9] or within LAGT and, subsequently, in our article.[10] Further, the works in LAGT employ a canonical reference system based on the CITE architecture,[11] which makes identification of any work in the dataset very straightforward.[12]

Within LAGT, the textual data from the Perseus Digital Library and the First Thousand Years of Greek project are subjected to several standard text preprocessing procedures, namely tokenization, POS-tagging and lemmatisation. To provide a better understanding of the subsequent analyses, it will be useful to briefly describe how they are implemented within LAGT. Tokenization is the procedure of splitting textual data into their constitutive elements, called tokens. Thus, within LAGT, each work is first divided into sentences and subsequently each sentence into words. In the next step, each token (i.e. word) is coupled with a POS tag[13] and a lemma.[14] Assignment of a POS tag works probabilistically as an output of a neural network model which has been previously trained on manually annotated ancient Greek sentences.[15] Subsequently, the lemmatisation works deterministically, trying to find a suitable word-

6 Kaše (2021).

7 Cerrato et al. (2020).

8 Crane et al. (2020).

9 E.g. Koentges (2020).

10 Because of relying on LAGT, some of our calculations might be slightly different from the ones we could obtain by employing other digital editions, namely from Thesaurus Linguae Graecae (TLG). For instance, the LAGT dataset contains a substantially shorter version of the Hippocratic treatise *Epidemiae*. This is due to the fact that the Perseus version of this work relies on Loeb's edition from 1923, which treats only books I and III as substantial representatives of epidemic medicine and does not contain books II and IV–VII.

11 Blackwell / Smith (2019).

12 Thus, the Hippocratic texts might be easily extracted using the CTS URN for author "tlg0627".

13 POS stands for part of speech. In the case of LAGT, the POS-tagging has the form of a coarse-grained analysis, which means that it assigns to a word only the part of speech itself (e.g. noun, verb, adjective, conjunction etc.) and not other morphological features such as gender, number, or tense (i.e. fine-grained analysis). For the POS tag categories, see https://universaldependencies.org/u/pos/ (Last access 10.07.2021).

14 Lemma is the dictionary form of a word. Thus, in the case of a verb, it is 1st sing. pres. ind. act. (e.g. δοκέω).

15 I.e., the model tries to predict the POS tag of a word by drawing on the structure of the current sentence and comparing it with sentences that the model encountered during its training. See the LAGT repository for more details.

form-lemma pair within the Greek part of the Morpheus Dictionary.[16] Having the words coupled with their POS tags makes it possible to filter texts according to them and focus only on lemmatised words coupled with certain POS tag categories. Since LAGT is primarily designed for semantic analysis, it returns lemmatised versions of the texts containing only words tagged as nouns, proper names, adjectives and verbs.[17] Being aware that the POS tagging and lemmatisation are semi-automatic processes, we should not be surprised that they are also prone to errors. Therefore, our data may still contain a negligible amount of incorrectly POS-tagged, improperly lemmatised or completely un-lemmatised words.[18] Despite this fact, it seems that this limitation does not bias the overall results of our analyses.

All computational text analyses introduced below have been implemented using the Python 3 programming language.[19] Since we aim to make our analyses fully reproducible and our code reusable by other scholars, all the data and the whole code used in this article are accessible via a Zenodo repository,[20] to which we occasionally refer below for details and supplementary data and figures.

Document distances

To obtain a general overview of the corpus and the relationship between individual documents, we firstly generated a document-term matrix, with rows representing individual works in the corpus and columns representing a subselection of words used in the corpus. The cell values within the matrix represent frequencies of these words across the works within the corpus. In particular, the selected words are words appearing in at least 10% of works in the corpus. This forms a set of 2,033 unique words. The rows of this matrix have been subsequently treated as vectors, expressing positions of points in a multidimensional space, with the number of dimensions equal to the number of words. Thus, we obtained a set of 52 vectors within a space with 2,033 dimensions. Having the data in this form, we can calculate distances between the vectors by measuring and inverting their cosine similarity. This way we obtain a matrix expressing distance between any two works within the corpus, with works sharing a larger proportion of words being less remote to each other than works employing less overlapping vocabulary. This distance matrix could be finally projected into a 2-dimensional space by using t-distributed Stochastic Neighbor Embedding (tSNE)[21] and plotted as a scatter plot (see Fig. 1 below).[22]

Pain-words in context

As we have already mentioned, when it comes to the concept of pain in *CH*, we have to deal with at least four word-families: πόνο*, ὀδύν*, ἄλγ*, λύπ*. Each of these word families combines several words, which we usually have in the lemmatised form. Thus, for instance, the most common lemmata from the ἄλγ* family appear to be the noun ἄλγημα and the verb ἀλγέω, with 141 and 84 instances respectively. However, in our corpus, there is also a significant number of instances of un-lemmatised words which

16 Crane (1991).

17 For a similar approach see Svärd et al. (2020), 470–502. This approach differs from computational stylometry, which commonly focuses on the usage of conjunctions, prepositions etc., which usually capture any difference in a style very distinctly. Cf. Koentges (2020), 211–41.

18 The accuracy (proportion of correctly annotated words from a text) of the POS-tagger and the lemmatiser is between 87 and 97 %, depending on the genre of the text.

19 Rossum / Drake (2009).

20 Kaše / Linka (2021).

21 van der Maaten / Hinton (2008), 2579–2605.

22 For details, see Kaše / Linka (2021), scripts/3_OVERVIEW+WORK-DISTANCES.ipynb (Last access 30.08.2021).

are not covered by the database we used for lemmatisation, like ἀλγεῦντα. We used regular expressions[23] to capture all these word forms and replaced them with a unified word pattern consisting of the word root and an asterisk: πόνο*, ὀδύν*, ἄλγ*, λύπ*.[24] In what follows, whenever we point out Greek pain-words, we refer to these word patterns.

Having the pain-words captured, we can focus on the context in which they appear. While the analysis of work distances treats the corpus as a list of individual works and each work as a list of words coupled with their frequencies, this analysis approaches the corpus as a list of sentences. For each pain-word, we firstly extract all sentences containing it. Subsequently, we compute term frequency (TF) for all words within these sentences. This measure gives us a general overview of the terms most commonly co-occurring with each of the pain-words. However, this measure does not distinguish between frequently appearing words in the sentences containing the pain-words, as they are semantically associated with them, and words frequently appearing here because of their distribution over the corpus as a whole. To overcome this limitation, we weight the TF measurement using a TFIDF algorithm. TFIDF stands for term frequency-inverse document frequency. Inverse document frequency (IDF) is obtained by dividing the total number of documents by the number of documents containing the term. Typically, the IDF value is logarithmically normalized. TFIDF is then a multiplication of the two measures:

$$TFIDF = TF \times log_2(IDF)$$

In effect, the weighting by IDF proportionally reduces the TF values of context general words while increasing the values of context-specific words. Using this measure, for each of the pain-words, we are able to identify the words which are most typical for their context.

Distributional Semantics and PPMI[3]

A step further is to adopt some methods from the field of distributional (or vector) semantics. The term distributional semantics designates a broad palette of methods from the areas of natural language processing and computational linguistics inspired by the distributional hypothesis of meaning,[25] henceforth it is also called Distributional Semantic Modeling (DSM).[26] Since these methods usually transform words into vectors, some scholars use the designation vector semantics.[27] According to the distributional hypothesis, words that occur in similar contexts tend to have similar meanings. Thus, to capture the meaning of a word requires an analysis of words most frequently surrounding it. But, as we will see, this is only a starting point. DSM goes further and constructs matrices and vector representations for whole corpora, which are subsequently transformed and analysed using complex algorithms from linear algebra and statistics. However, to work properly, most of the DSM algorithms require very large textual data (typically at least 1 million words) to be trained on. In this respect, our corpus consisting of 171,332 words is rather small and therefore allows us to employ only certain distributional semantic models.[28]

23 López / Romero (2014).

24 We use the asterisk character (*) to mark that the word is a product of a regular expression match.

25 For distribution hypothesis, see Harris (1954), 146–62.

26 Lenci (2018), 151–71.

27 For a basic overview of the most common algorithms, see Jurafsky / Martin (2020), 270–85.

28 For instance, it has been demonstrated that the well-known word2vec model outperforms other methods when trained on 1 billion words of data. However, when trained on a smaller dataset, consisting of 1, 10 or 100 million words, it is outcompeted by much simpler models. For word2vec, see Mikolov et al. (2013), 3111–3119. For comparison of word2vec with other models, see Sahlgren / Lenci (2016); Altszyler et al. (2016).

In what follows, we employ a DSM approach combining Pointwise Mutual Information (PMI) and Singular Value Decomposition (SVD). In its basic version, PMI has the following form:

$$PMI(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)},$$[28]

where x and y represent two words, *P(x,y)* their probability of appearing together within a predefined context within a corpus, and *P(x)* and *P(y)* their probabilities of appearing independently, i.e. their respective term frequencies within the corpus.[30] The ratio is subsequently normalized by a logarithm with base 2. However, a well-known problem associated with this measure is that it gives very high scores to word pairs involving infrequent words, as the denominator is rather small in such cases. Therefore, several of modifications of PMI have been proposed to overcome this limitation.[31] Here we employ the so-called PMI^3, which modifies the measure by cubing the *P(x,y)* value and which has been documented to produce reasonable results:

$$PMI^3(x,y) = log_2 \frac{P(x,y)^3}{P(x)P(y)}$$[31]

Finally, since the fraction $\frac{P(x,y)^3}{P(x)P(y)}$ usually returns values lower than 1 and since log_2 for numbers smaller than 1 is a negative number (which might be confusing for a visual inspection), we finally create a PPMI^3 measure transposing the PMI^3 values to a scale from 0 to 1, with PPMI^3=0 for all word pairs with joint probability P(x,y) equal to 0 (i.e. for words which do not appear together at all) and the rest on a scale from 0.5 to 1, with PPMI^3=0.5 assigned to a word pair with the minimal PMI^3 value in total (but different from 0) and PPMI^3=1 assigned to a word pair with the maximal PMI^3 value in total.

Drawing on this, we can generate a PPMI^3 matrix by calculating the PPMI^3 value for each possible word pair of all words appearing in at least 5 works within the corpus. Such a matrix gives us straightforward access to weighted *first-order co-occurrence* (also called *syntagmatic association*) between any two words forming the matrix. Thus, for instance, in the case of English, the word "blue" tends to co-occur with the word "colour". In this respect, the PPMI^3 attempts to capture the same type of semantic relatedness as the TFIDF metric we described in the previous section.

However, the PPMI^3 matrix allows us to access *second-order co-occurrence* (also called *paradigmatic association*) as well.[33] This means that, after a subsequent transformation and analysis of the matrix, we are able to capture the semantic association between words that perhaps do not occur so often together but do tend to co-occur with similar third-words. Thus, there might be a strong paradigmatic association between the words "blue" and "green", since they both co-occur with a third word "colour" and a number of other colour-related third words. In principle, we can measure this sort of semantic relatedness between any two words by comparing the row vectors corresponding to them within the PPMI^3 matrix. In fact, this sort of vector comparison lies at the core of vector semantics as such, and gives it its name.

However, to make the vector comparison more robust, we further employ Singular Value Decomposition (SVD) to reduce their dimensionality, i.e. we transform them from sparse high-dimensional vectors

29 Church / Hanks (1990), 22–29.

30 For all the subsequent analyses, see Kaše / Linka (2021), scripts/5_VECTORS.ipynb (Last access 30.08.2021).

31 Levy et al. (2015), 211–25.

32 Role / Nadif (2011), 218–23.

33 For the difference between first-order and second-order co-occurrence, see Jurafsky / Martin (2020), 274–75 and Schütze / Pedersen (1993), 104–13.

with 2,033 dimensions to lower-dimensional (denser) vectors with 250 dimensions.[34] The outcome is a PPMI^{3}SVD matrix, in which each row corresponds to a 250-dimensional vector representation of a word. Subsequently, we employ cosine similarity to construct a similarity matrix comparing any two-row vectors against one another.[35] Using this similarity matrix, for any word we choose we can easily extract a certain number of the most similar words to it, i.e. its nearest neighbours. As we have already mentioned, these similarities between words attempt to capture the so-called paradigmatic association between them. It has been repeatedly demonstrated that, when trained on large and representative language corpora, this sort of method is able to automatically detect synonymity and some other types of semantic relatedness – a capability that might be evaluated against benchmark tests based on manually coded data.[36]

Results

Corpus overview and document distances

The *Corpus Hippocraticum* (*CH*) extracted from LAGT consists of 52 works.[37] These works are formed by 24,456 sentences and 171,332 lemmatised words tagged either as nouns, proper names, adjectives or verbs.[38] To obtain a basic overview of the corpus, we have produced Figure 1, which plots distances between individual works in *CH* based on similarities and dissimilarities in their vocabulary. The term vocabulary here refers to this subselection of lemmatised words. Works depicted closer to each other tend to share more words than works that are farther from each other.

Upon analysis of Figure 1, we see that it produces some local clusters of works. For instance, quite unsurprisingly, on the left side of the figure we see very close to each other two works which have been classified by Jouanna as "Dietetics".[39] This suggests that the method performs well in capturing this thematic relatedness. Furthermore, on the right side, we see a cluster of works formed by five texts which have been classified as "Surgical". Again, drawing on the vocabulary, our measurement properly captures that these five texts are indeed related. At the top of the figure, there is another relatively homogeneous cluster (*Lex*, *De decente habitu*, *De arte*, *Praeceptiones*, *De medico*, *Epistulae*), which, however, does not fall under any single category proposed by Jouanna. Yet, thematically, these writings appear to be related, since all of them somehow concern the profession and social role of the physician.

34 In the context of vector semantics, SVD was originally popularized by Latent Semantic Analysis, where it serves to reduce the dimensionality of a word-document matrix (see Deerwester et al. [1990], 391–407). Here we employ it to reduce the dimensionality of the PPMI^3 matrix, which might be considered a weighted variant of the word-word co-occurrence matrix. For the same approach and its rationale, see Levy et al. (2015), 211–25.

35 This analysis shares several features with the analysis of distances we have introduced above. The main difference is that there the vectors represented works, whereas here they represent words.

36 See e.g. Levy et al. (2015), 211–25; Sahlgren / Lenci (2016); Baroni et al. (2014), 238–47.

37 19 works are included in the Perseus Digital Library; the rest originate from the First Thousand Years of Greek project.

38 In total, the corpus consists of 333,443 raw words. For more details, see Kaše / Linka (2021), scripts/1_EXTRACTING-CORPORA.ipynb (Last access 30.08.2021).

39 In Figure 1, we adopt a classification proposed by Jouanna (1999), 66–71. He acknowledges that particular writings of the corpus vary in both date and authorship, and that despite it being problematic to categorize the corpus, he attempts to do so and classifies particular writings into groups indicated on the right in Figure 1. This classification is based on the criteria of content and the date of composition. Other authors propose different classifications, emphasising different groups of writing; thus, the discussion about the classification of *CH* is ongoing (see e.g. Craik [2015]), xiv–xxxv). For comparison, we have also generated a figure using Craik's categories; see Kaše / Linka (2021), figures/c_hip_distances_by_cat_craik.png (Last access 30.08.2021).

It is encouraging to see that our method is able to capture this aspect as well. At the bottom right, we can see another strongly distinguishable group, formed mainly by works classified as "Later" or "Other". There are also two other works from the category of "Female medicine". This cluster of works, which represents different categories, can be explained by several factors: It can either mean that at least some of the works classified by Jouanna as "Later" and "Other" are indeed also related to the topic of "Female medicine", or that the two works from the category of "Female medicine" reveal substantial similarities with some later works. We are not able to decide which is the case here, since it would either require a detailed close reading of the texts or an employment of another CTA (computational text analysis) method, e.g. a stylometric analysis. However, both would divert us from the main topic of this article, which is the understanding of pain.

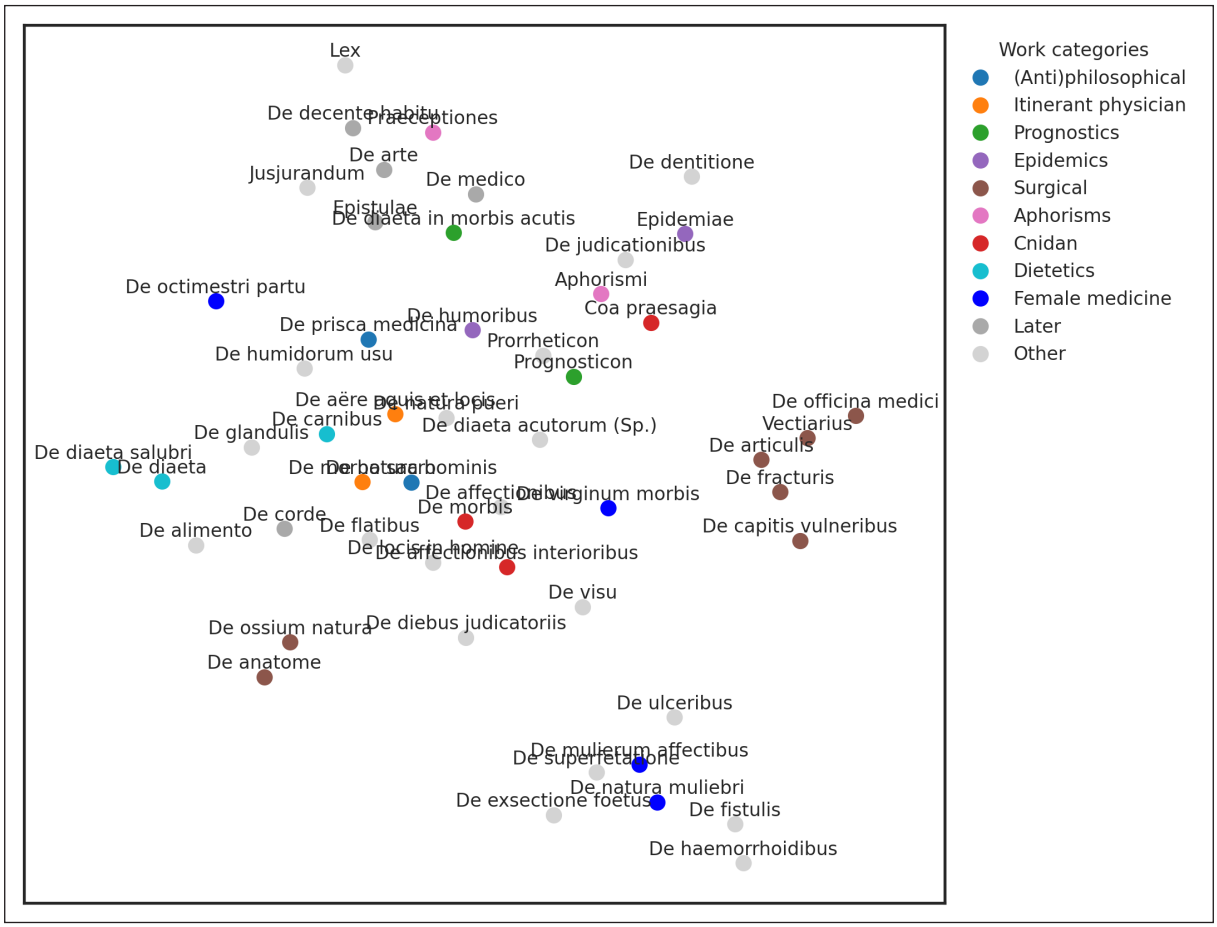

Fig. 1: Work distances based on shared vocabulary.

Taken together, Figure 1 helps us demonstrate several different things. First, it allows us here to validate our overall approach, since the work clustering obtained by this method might easily be evaluated against any work classification offered by experts conducting close reading of the sources. In that respect, we should realise that the classification of works within *CH* is an important prerequisite for almost any inquiry concerned by the history of ancient medicine, given the fact that *CH* is a heterogeneous corpus of works written by different authors over the span of more than a century.[40] At the same time, it can also direct future research, identifying subtle similarities of works that are otherwise treated separately, as in the case of the cluster at the bottom-right of the figure discussed above. However, we should also not ignore the limitations of this particular method. Firstly, it completely ignores word order, employing what is known as a bag-of-words approach.[41] This substantially constrains the possibility of inferring

40 Craik (2015), xxiv–xxviii.

41 Jurafsky / Martin (2017), 76.

anything substantial concerning the semantics, since the meaning of words is determined by their context of usage on the level of sentences etc., as captured by the DSM. Secondly, here we focus exclusively on a subselection of lemmatised words, namely nouns, proper names, adjectives and verbs. This naturally flattens any differences in style, which are typically mirrored in the usage of function words like καί, δέ, μέν or τε and which are therefore commonly used in stylometric analyses for authorship attribution.[42]

Pain words across work categories and sentences

After the analysis of work distances, we proceed to the problem of pain in *CH*. In this case, we have to focus on usage of the four pain-words. Our dataset contains 657 instances of πόνο*, 645 instances of ὀδύν*, 315 instances of ἄλγ*, and 58 instances of λύπ*.[43] Thus πόνο* and ὀδύν* appear to be the most frequent ones, while λύπ* tends to be used only rarely. Remarkably, the proportion of usage of these word families is completely different than the one we observe in other ancient Greek texts from a similar period, which represent a different genre. For instance, from the four pain-words, Aristotle most often uses λύπ* (406 instances), followed by πόνο* (103 instances); there are only 34 instances of ἄλγ* and 3 instances of ὀδύν*.[44] Furthermore, as shown in Figure 2, the proportional distribution of the pain-words also broadly varies across individual work categories within the corpus.[45]

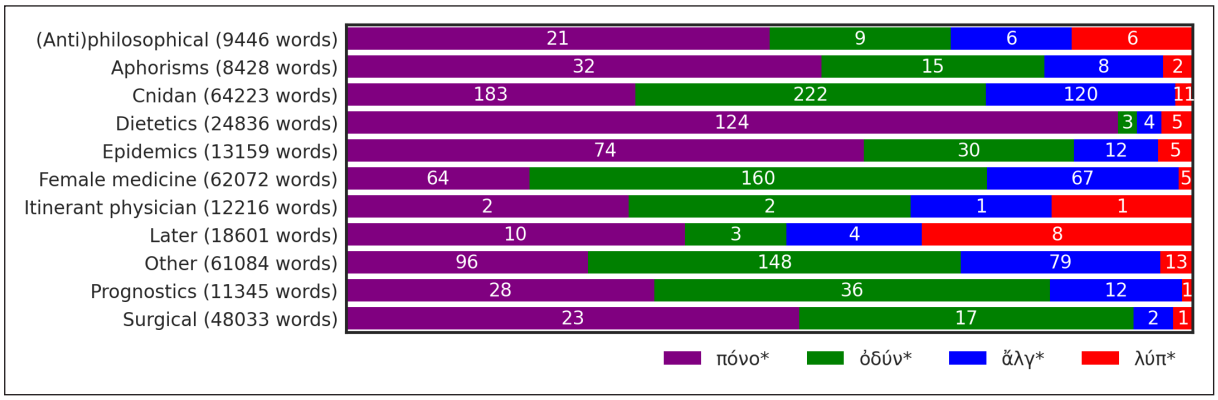

Fig. 2: Ratios of pain-words across work categories by Jouanna.

Figure 3 plots 20 words with the highest TFIDF scores within all sentences containing the individual pain-words. We have manually classified the terms into several categories (pain-word, pathological state, body and its parts, dietetic term, quality, and general term) and differentiated these terms by colours (see legend on Figure 3).[46]

42 Koentges (2020), 211–41.

43 In TLG, we can find 891 instances of ὀδύν*, 709 instances of πόνο*, 379 instances of ἄλγ*, and 60 instances of λύπ*. This numerical difference has at least three reasons: Firstly, the TLG search engine includes composite words like κεφαλαλγία, while we focus only on words beginning with the root. Secondly, TLG employs different lemmatisation. Finally, TLG includes editions of some works that are different from ones available via open resources.

44 For the extraction of pain-words and comparison with Corpus Aristotelicum, see Kaše / Linka (2021), scripts/2_EXPLORATIONS+REPLACEMENTS.ipynb (Last access 31.08.2021).

45 For the proportion of pain-words across the work categories proposed by Craik, see Kaše / Linka (2021), figures/c_hip_ratios_by_cat_craik.png (Last access 31.08.2021).

46 For details, see Kaše / Linka (2021), scripts/4_PAIN-SENTENCES.ipynb (Last access 31.08.2021). For the full list of terms classified by categories and accompanied by automatic translations, see Kaše / Linka (2021), data/terms_by_category.csv (Last access 31.08.2021).

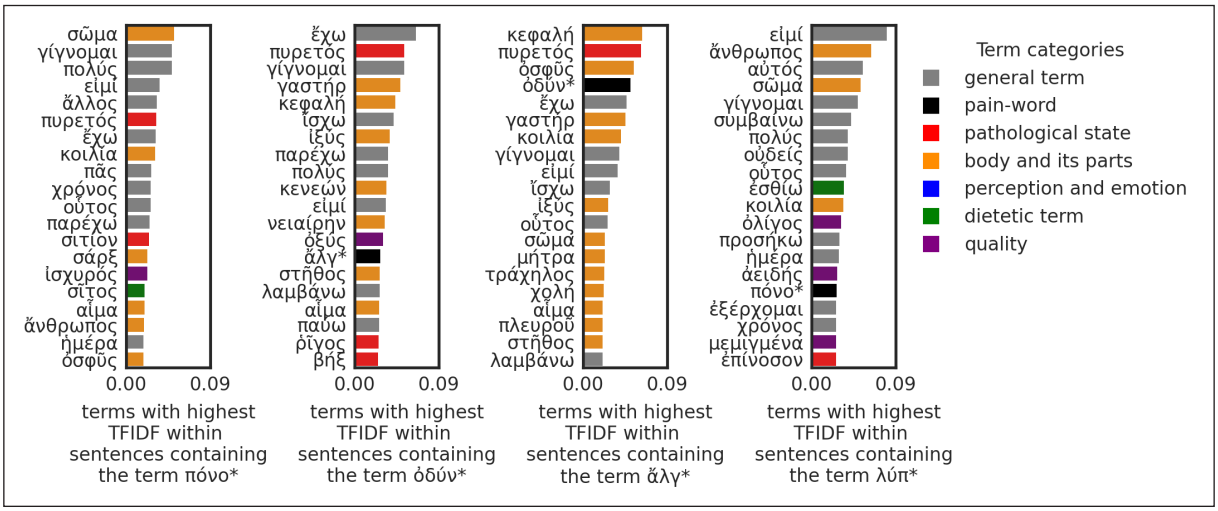

Fig. 3: 20 terms with highest TFIDF within sentences containing the pain-words.

We can see that the terms listed in Figure 3 substantially overlap between the four subplots. However, there are also some remarkable differences. From the four word-families, ἄλγ* seems to co-occur very frequently with individual bodily parts and constituents (12 from 20 terms with the highest TFIDF value), followed by ὀδύν* (7 terms) and πόνο* (6 times).[47] The usage of πόνο* tends to be more general, and is associated with terms like σῶμα ("body"), σάρξ ("flesh") or αἷμα ("blood"). Looking at this data, λύπ* appears to be a term from a slightly different semantic domain, only marginally connected with the somatic and medical domain. This is unsurprising given the fact that λύπ* in classical Greek literature usually denotes sorrow or some other negatively evaluated emotional state.[48] Thus, this analysis of sentences using the TFIDF algorithm gives us some preliminary insights concerning the contexts in which pain-words appear. The advantage of this method is that it is computationally rather straightforward and easy to interpret. However, it does not allow us to go as deep concerning the semantics of the terms under scrutiny. This requires the adoption of more advanced methods, which will be the subject of the following section.

Distributional Semantics and Word Embeddings

By calculating the PPMI^3 value for each possible word pair of all words appearing in at least 5 works within the corpus, we obtained a square matrix of 2,033 rows and 2,033 columns. Subsequently, for each of the pain-words, we used this matrix to extract 20 words having the highest PPMI^3 association with them (see Figure 4). It should not surprise us that the results are highly comparable to the results we obtained using the TFIDF algorithm (see Figure 3). Since both measures attempt to capture the same type of semantic relatedness, we can consider this observation as a sort of validation of this second, more complex measure. It is important, since the PPMI^3 matrix serves us here as a middle step in the construction of a PPMI^{3}SVD matrix, which we can use to calculate word vector similarities in an attempt to capture the paradigmatic association between any two words of our interest.

47 Remarkably, the method captures this feature even while it does not include words like κεφαλαλγία.

48 The sense of this word becomes broader in the context of the Greek tragedy, and its authors – Aeschylus, Sophocles, Euripides – also use it in the sense of mental pain. However, it is only in Plato and Aristotle where λύπ* is used for denoting physical pain as well, and it works as a general term for pain in opposition to ἡδονή (pleasure). See Cheng (2019), 47–71. Also, our close reading analysis introduced below shows that in *CH*, λύπ* usually keeps its non-physical-pain sense, even though there are some rare exceptions.

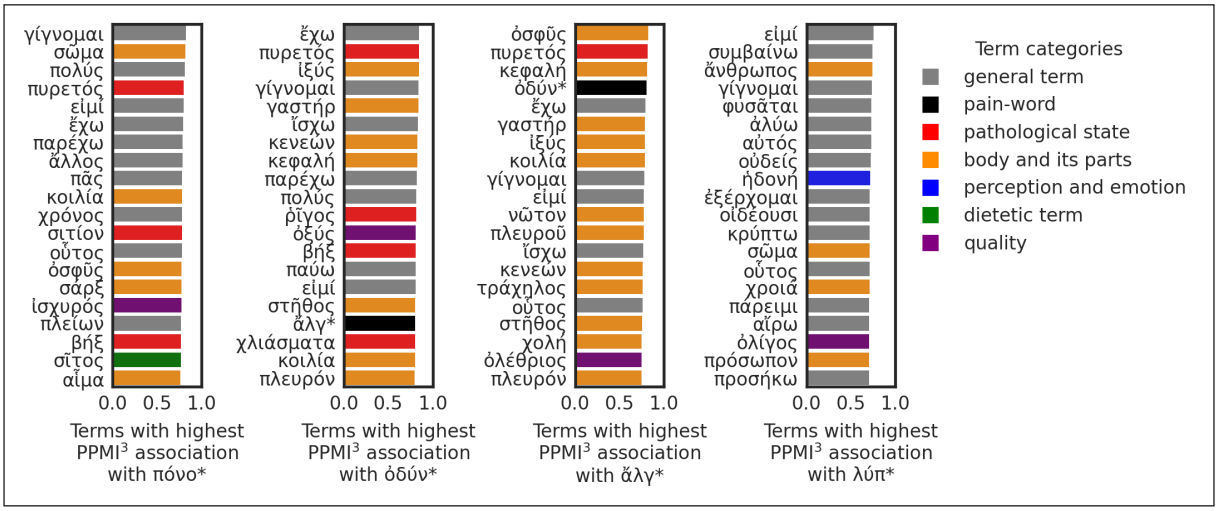

Fig. 4: Pain-words coupled with 20 terms with highest PPMI³ association with them.

Figure 5 is based on the cosine similarity of words within the PPMI³SVD matrix, in which each row corresponds to a 250-dimensional vector representation of a word. In particular, it contains the 20 nearest neighbours for each of the pain-words together with horizontal bars expressing a cosine similarity score on a scale from 0 to 1. Firstly, when looking at the third column, we see that ἄλγ* is no longer as strongly associated with the body and its parts as was the case of TFIDF and the original PPMI³ matrix scores (cf. Figure 3 and 4). This should not surprise us, since we are now capturing the second-order (paradigmatic) association and not the first-order (syntagmatic) co-occurrence. Following this, we see in the third subplot that the nearest neighbour of ἄλγ* is ὀδύν*. At the same time, we observe in the second subplot that in the case of ὀδύν*, ἄλγ* occupies the 9th position. The score associating ὀδύν* and ἄλγ* is the same in both cases, but in the case of ὀδύν* there are other terms with higher scores.[49]

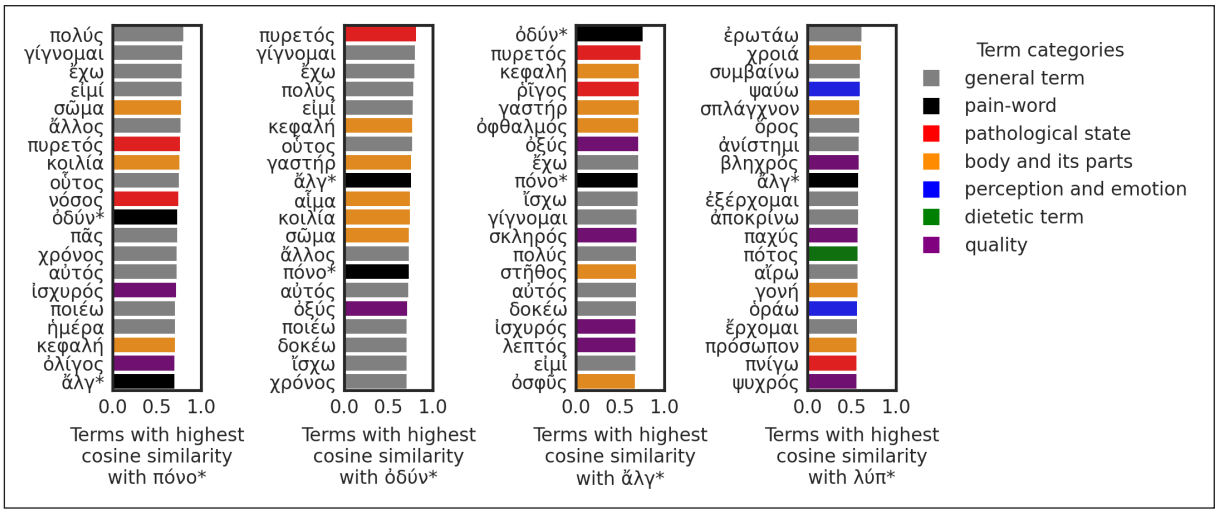

Fig. 5: Pain-words coupled with 20 terms with the highest cosine similarity of vectors based on the PPMI³SVD.

Taken together, there seems to be much overlap between ὀδύν* and ἄλγ*. In both cases, we see a very strong association with πυρετός ("fever"). Both terms reveal a highly medicine-specific context without any clear semantic difference. πόνο* also reveals some association with πυρετός, but the predominance of general terms suggests that the semantic context is slightly different. When we look at the λύπ* column, it appears that we are deviating even farther from the medicine context than in the case of πόνο*.

49 Again, we observe here a significant number of general terms in the figure. It is a consequence of the PPMI³ metric, which we used to construct the PPMI³SVD matrix.

Again, as we have already discussed above, this is unsurprising because λύπ* was originally used in the sense of sorrow. The results of Figure 5 will be elaborated upon further within the Discussion.

Relying on the same data that we used for the creation of Figure 5, we can proceed further with another visualisation, which will be to a certain extent similar to the one we used for plotting distances between individual works by drawing on their shared vocabulary. This time we will plot distances between words by inverting similarity scores from the PPMI^{3}SVD similarity matrix. As in the case of work distances, we firstly apply tSNE to project the data from the distance matrix into a 2-dimensional space. Subsequently, we plot these data using a scatter plot, a standard way to visualise word-embeddings. However, since there are 2,033 data points (i.e. words), it is not possible to plot all of them in a meaningful way together within one plot. Therefore, in Figure 6, we instead introduce a series of four subplots.

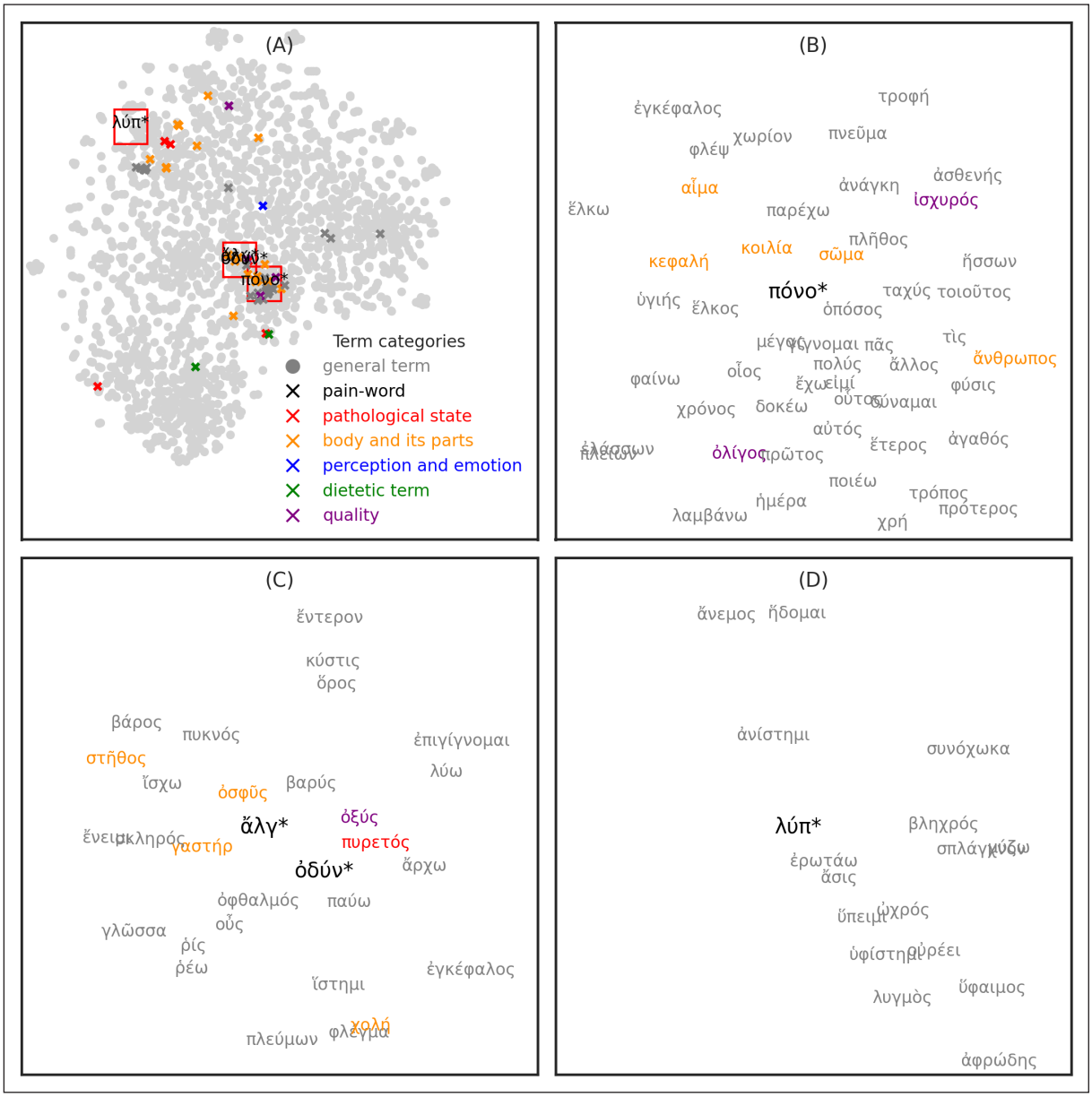

Fig. 6: Word-embeddings based on the PPMI^{3}SVD matrix. Subplots (B), (C) and (D) represent cutouts containing the four pain-words.

Subplot (A) gives us a general overview of the spatial distribution of the 2,033 words within the model. This distribution is based on distances between these words within the PPMI^{3}SVD distance matrix (an inverted version of the PPMI^{3}SVD similarity matrix). Points depicted using a cross sign depict words already contained in Figures 4 and 5. As we move further from the middle of the figure, we can identify some more densely clustered groups of points corresponding to semantically closely related groups of words. We also see that three of the pain-words appear rather close to the centre. This suggests that these words are strongly connected with the rest of the corpus, appearing in more than one specific context. In that respect, λύπ* appears to be much less anchored within it, which also reflects its substantially lower frequency. The subplot (A) further depicts three squares surrounding the pain-words, which are used for a subselection of data for subplots (B), (C), and (D).

In subplot (C), we see that ἄλγ* and ὀδύν* appear very close to each other (this is also seen in the Figure 5). This also allows us to capture their neighborhood using one subplot. Furthermore, we also observe here the close association with πυρετός, which appears to be much stronger than between πόνο* and πυρετός. Finally, we can identify here a number of terms from the category of the body and its parts. The neighbourhood of the term πόνο* as depicted by subplot (B) seems to be preoccupied by semantic general terms which are commonly the most frequent terms in the corpus as a whole. The subplot with λύπ* (D) is extracted from a much less densely populated part of the embedding. This also helps us to understand why its similarity values with its closest neighbours as depicted in the fourth subplot of Figure 5 are comparatively much lower than the values we observe in other subplots. Taken together, it seems that the usage of λύπ* in *CH* does not reveal any specific semantic context, what is also caused by its limited extent of usage.

Discussion

In the previous section, we have captured a significant semantic association between ἄλγ* and ὀδύν*. The similarity of the two pain words to each other is clearly manifested in Figure 6 as well as in Figures 3–5. Both terms are closely associated with bodily organs or pathological states (see especially Figures 3–5, where the connection to bodily organs is substantially stronger than in the case of the other pain words). These insights can be further validated and elaborated by close reading of the texts. When we go through various thematically dissimilar texts, for example *Coa praesagia*, *De fracturis*, *De natura muliebri*, *Prognosticon*, or *Epidemiae*, we find ἄλγ* and ὀδύν* used usually as examples of pain occurring in some particular body part as a result of an illness or other pathological state.[50] It is worth mentioning that ὀδύν* maintains the sense of a specific physical pain even in the treatises which are more theoretical and general, for example *De natura hominis* or *De prisca medicina*,[51] whereas ἄλγ* can be used in these types of treatises as a general term denoting pain, by which the author proposes his theory of the nature of pain.[52]

Whereas in the case of ἄλγ* and ὀδύν* the DSM analysis reveals some clear connections to bodily organs or pathological states, in the case of πόνο* the results are less decisive. We have seen above that this word is closer to general terms rather than to some special medical vocabulary (see especially Figure 6 [B]). Of course, this word is related to other medical terms, too, but remarkably, Figure 3–5 depict it as being closely associated with rather general terms such as σῶμα, rather than to some particular bodily organ. Yet, when we conducted a close reading of some representative selection of Hippocratic texts, we

50 Hippocr. *Coac*. 18.1, 195.1, 265.5, 274.7; *Fract*. 7.2, 9.21, 17.6; *Nat. Mul*. 2.7, 5.2–4, 5.2, 6.2–3, 7.3, 18.1; *Progn*. 5, 7, 19, 24; *Epid*. 1.2.6.1–14, 1.3.13.17, 1.1.3.26.

51 Hippocr. *Nat. Hom*. 4.10–14, 11.13, 11.36, 12.2, 15.2, 12.2, 15.2, *Med. Vet*. 19.5, 22.51.

52 Hippocr. *Nat. Hom*. 2.8–12; 4.3–5.

found πόνο* used in a way very similar to ἄλγ* and ὀδύν*, i.e. in connection with a bodily organ and a pathological state.[53] Thus, we expected that this association will be apparent in the DSM analysis as well. However, we must take notice of the fact that 107 of 657 instances of πόνο* in the whole corpus appear in *De diaeta*, where it has a meaning different than pain. In this dietetic work, πόνο* usually designates exercise.

Thus, to explore the possibility that the overall meaning of this term is substantially influenced by this one work, we re-ran the whole DSM analysis, this time without *De diaeta*.[54] In this version, we found the term πόνο* to be more closely related to the other pain-words, especially ὀδύν*.[55] Thus, it appears that in the case of πόνο*, the overall results are substantially influenced by this particular writing and the specific meaning of πόνο* in it. For instance, we can also see in Figures 3–5 that πόνο* is connected to some temporal attributes such as ἡμέρα or χρόνος, which should not surprise us, because time and duration play an important role in the dietetics. Nevertheless, taken together, it seems that πόνο* has a slightly broader meaning than ἄλγ* and ὀδύν* in *CH*, a feature which is captured within the DSM analysis by its close association with more general terms. This feature is also recognised by at least some translators, who choose to render it as "suffering" or "souffrance".[56]

The DSM analysis of ἄλγ*, ὀδύν* and – to some degree – πόνο* seems to support an interpretation of the problem of pain in *CH* made by some scholars over the past thirty years.[57] They all agree that pain in *CH* figures as a symptom of illness and that it is usually connected with a concrete bodily organ or area. We believe that the close association between pain-words on the one hand and bodily organs and pathological states on the other captured by the DSM analysis supports this claim. It is of interest that all pain-words can relate to various words specifying the quality of pain (sharp, intensive etc.), which, possibly, says something about how the patients classified their pain (this is most noticeable in Figure 5). Yet, with the methods we use in this paper, it is difficult to elaborate on the problem of how the patients felt their pain. To enhance this question, we would need to focus more on semantic and psychological analyses of *CH*.

As we have already mentioned several times, the word λύπ* occupies a specific position within the corpus, which is most clearly seen in Figure 6 (A). Figures 3–5 reveal that λύπ* is not associated as much with bodily organs or pathological states, and is more connected with words of other types, e.g. ἡδονή (see Figure 4). The word λύπ* is also the only pain-word connected to sense-perception (see Figure 5), a trend which is also documented in the philosophical literature of the time.[58] Furthermore, it is noticeable that in the case of Figure 3, both λύπ* and πόνο* maintain a strong connection to general terms like σῶμα ("body"), ἄνθρωπος ("human being"), ἡμέρα ("day") or χρόνος ("time"). Thus, it seems that λύπ* is usually not meant in the sense of a concrete physical pain, an observation which is evidenced by close reading as well. In the majority of writings, this word is either completely absent or used only exceptionally. Even in the works where it is used more often, it usually means an emotion of sorrow[59] or pain or

53 *Fract*. 2.32, 3.8, 5.30, 6.8. *Coac*. 31.1–3, 76.3, 138.2, 139.2–3; *Nat. Mul*. 5.4, 12.14, 18.2, 23.1. *Epid.* 1.2.6.5–11, 1.2.3.4.123, 1.3.13(2).25, 1.3.13(4).5–7; *Progn*. 5, 11, 19.

54 See Kaše / Linka (2021), scripts/6_VECTORS_without-de-diaeta.ipynb (Last access 31.08.2021).

55 The 20 terms with the highest PPMI^3 score were: γίγνομαι, πολύς, σῶμα, πυρετός, εἰμί, ἔχω, ἄλλος, κοιλία, πᾶς, παρέχω, ὀσφῦς, βήξ, χρόνος, ἰσχυρός, οὗτος, αἷμα, ὀδύν*, ὀξύς, κεφαλή, τράχηλος.

56 See especially theoretical writings like *Vict*., *Nat. Hom*., *Med. Vet*. translated by W. H. S. Jones and E. Littré.

57 King (1998); Horden (1999), 295–315; Rey (1995).

58 For Aristotle, for example, sense-perception is a necessary condition for feeling pain and pleasure, and the relation between pain, pleasure and perception is an important topic in Aristotelian scholarship. See *DA* II, 3, 414a32–b16. Cf. Corcilius (2008), 79–82.

59 Hippocr. *Epid*. 3.3.17(11)3, 3.3.17(15)3.

suffering in general without any explicit connection to a bodily organ or pathological state.[60] However, some moderation is required in ascribing λύπ* to any specific context in *CH*, because of its scarcity and relative distance to other terms (see the interpretation of Figure 6 [B] above).

We should not overlook that our methods are not able to capture some important specifics and exceptions which occur in some particularly important writings of *CH*. Especially in writings such as *Nat. Hom*., *Med. Vet*. and *Vict*., we find intriguing passages about the nature of pain, its generation and further scientific and philosophical implications. However, in *CH* as a whole, the treatises containing an explicated theory of pain make up a minority.[61] In this respect, the value of the DSM analysis lies in its capability to look at the corpus as a whole, without being biased by a few writings that represent an exception. Nevertheless, if we are interested in the problem of pain not in the perspective of the whole corpus, but, for instance, in the treatises particularly influential for the reception of Hippocrates in Western thought, it is important not to overlook some intriguing claims connected to the theory of pain presented in them.[62]

In the future, we envision the possibility of employing other computational text analysis approaches when studying pain in *CH*. In particular, stylometric analysis is a promising research pathway to evaluate some hypotheses discussed within the scholarship. Rey, for example, claims that there is a difference between ὀδύν* and πόνο* based on the prepositions with which these words occur: ὀδύν* usually occurs together with more concrete prepositions, so it is a more precisely localised type of pain, whereas πόνο* denotes a more general type of pain because it is connected with prepositions which are not particularly specific.[63] Without a doubt, it is possible to quantitatively evaluate such claims without employing any advanced computational techniques, e.g. by exploring available word indices and concordances. However, computational stylometry allows us to do this in a more controlled fashion, comparing a large number of features at once. Furthermore, a more complex distributional semantic analysis could also help us evaluate King's claim that πόνο* usually means natural pain (for example birth pain), whereas ὀδύν* unnatural pain (being a result of some damage to the organism).[64] Finally, the methodological framework we employed here can easily be transferred and applied to other comparable or even much larger textual corpora of ancient Greek texts. Thus, for instance, we could analyse the understanding of pain in *Corpus Aristotelicum* or *Corpus Galenicum*, both of which are covered by the LAGT dataset that we have used here. Furthermore, the algorithms in the core of our scripts might also be modified and reused by other scholars to study different topics.

60 Hippocr. *Med. Vet*. 14.23–28; *Vict*. 15.5–6. Thus, it seems that λύπ* in *CH* has a different meaning than in classical philosophical literature. Only in works like *Med. Vet*. and *Vict*. is the meaning similar. See for example Aristotle, *EN* 1152b1–8; 1153b1–4; 1154a22–31; Plato, *Gorg*. 492a–499a; *Phlb*. 31a–34a, 44a–45a; *Resp*. 583b–584b; *Phaed*. 65b–c, 68e–69b, 83d–84e. For broader discussion about pleasure and pain in Plato and Aristotle, see Cheng (2015); Frede (2016), 255–76; Wolfsdorf (2013). For the semantics of pain in Aristotle and his contemporaries (including *CH*), see Cheng (2019). Plato and Aristotle, however, use λύπ* also in cases where Hippocratic authors would use different pain-words, i.e. in the case of a specific bodily pain.

61 This is emphasised by Horden (1999), 295–315, who underlines that in respect to pain, *CH* differs from the philosophical corpora of classical antiquity in its absence of any theoretical conception of pain.

62 For the theory and origins of pain as well as its relation to the nature of the human body, see Hippocr. *Nat. Hom*. 2.8–12; *Med. Vet*. 14.23–28; *Vict*. 66.42–46. For the connection between pain, sense-perception and mind, see *Aph*. II.6, II.46.

63 Rey (1995), 18–19.

64 King (1998), 267–286.

Conclusion

In this article, we approached the problem of pain in *Corpus Hippocraticum* by combining a distributional semantic analysis of the corpus with the close reading of selected works. We have especially focused on the semantic similarity between pain-words and other relevant terms. Our interpretation indicates that in the case of ἄλγ*, ὀδύν*, there seems to be a shared close association between pain, bodily organs and pathological states. Thus, as far as we deal with these word families, our findings are in accord with the interpretation advocated by some other scholars, who view pain in *CH* as a symptom of a pathological state located within some part of the body. From the same perspective, the meaning of πόνο* tends to be similar, but slightly more general, revealing substantially weaker association with the medical domain; the word λύπ* stands completely aside. We find it remarkable that, even though the Hippocratic authors offer neither an explicit conception of pain nor its definition, we are able to uncover some general features of its understanding typical for the corpus and to capture some semantic differences between the relevant terms.

Text editions

Littré (1839–1861): É. Littré (ed./transl.), Oeuvres complètes d'Hippocrate, Paris 1839–1861, repr. Amsterdam 1961–1962 & 1973–1991.

References

Altszyler et al. (2016): E. Altszyler / M. Sigman / S. Ribeiro / D. F. Slezak, Comparative Study of LSA vs Word2vec Embeddings in Small Corpora: A Case Study in Dreams Database, arXiv [cs.CL] 5 (2016) http://arxiv.org/abs/1610.01520 (Last access 31.08.2021).

Baroni et al. (2014): M. Baroni / G. Dinu / G. Kruszewski, Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2014, 238–47.

Blackwell / Smith (2019): C. W. Blackwell / N. Smith, The CITE Architecture: a conceptual and practical overview, in: M. Berti (ed.), Digital Classical Philology, Ancient Greek and Latin in the Digital Revolution, Berlin 2019, 73–93.

Cerrato et al. (2020): L. Cerrato et al., PerseusDL/canonical-greekLit 0.0.2711 (Version 0.0.2711), Zenodo, http://doi.org/10.5281/zenodo.4067170 (Last access 10.07.2021).

Cheng (2015): W. Cheng, Pleasure and Pain in Context: Aristotle's Dialogue with his Predecessors and Contemporaries, PhD diss. Humboldt Universität Berlin 2015.

Cheng (2019): W. Cheng, Aristotle's vocabulary of pain, Philologus 163(1) (2019), 47–71.

Church / Hanks (1990): K. W. Church / P. Hanks, Word Association Norms, Mutual Information, and Lexicography, Computational Linguistics 16 (1990) 22–29.

Corcilius (2008): K. Corcilius, Streben und Bewegen, Aristoteles' Theorie der animalischen Ortsbewegung, Berlin / New York 2008.

Craik (2015): E. M. Craik, The 'Hippocatic' corpus, London / New York 2015.

Crane (1991): G. R. Crane, Generating and Parsing Classical Greek, Literary and Linguistic Computing 6 (1991), 243–245.

Crane et al. (2020): G. R. Crane / L. Mueller / B. Robertson / A. Babeu / L. Cerrato / T. Koentges / R. Lesage / L. Stylianopoulos / J. Tauber, First1kGreek (Version 1.1.5070), Zenodo, http://doi.org/10.5281/zenodo.4091475 (Last access 10.07.2021).

Deerwester et al. (1990): S. Deerwester / S. T. Dumais / G. W. Furnas / T. K. Landauer / R. Harshman, Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science. American Society for Information Science 41 (1990), 391–407.

Frede (2016): D. Frede, Pleasure and Pain in Aristotle's Ethics, in: R. Kraut (ed.), The Blackwell Guide to Aristotle's Nicomachean Ethics, Blackwell publishing 2016, 255–76.

Harris (1954): Z. S. Harris, Distributional Structure, Word & World 10 (1954), 146–62.

Horden (1999): P. Horden, Pain in Hippocratic Medinine, in: J. R. Hinnells et al. (eds.), Religion, Health and Suffering, London 1999, 295–315.

Jänicke et al. (2015): S. Jänicke / G. Franzini / M. F. Cheema / G. Scheuermann, in: Eurographics Conference on Visualization (EuroVis) 2015, 1–21.

Jouanna (1999): J. Jouanna, Hippocrates, Baltimore / London 1999.

Jurafsky / Martin (2020): D. Jurafsky / J. H. Martin, Speech and Language Processing 2020, https://web.stanford.edu/~jurafsky/slp3/ (Last access 31.08.2021).

Kaše (2021): V. Kaše, sdam-au/LAGT v1.0.0 (Version v1.0.0), Zenodo, http://doi.org/10.5281/zenodo.4552601 (Last access 10.07.2021).

Kaše / Linka (2021): V. Kaše / V. Linka, PIA – article supplementary (Version v1.0.2), Zenodo, http://doi.org/10.5281/zenodo.5089410 (Last access 10.07.2021).

King (1998): H. King, Hippocrates' Woman, Reading the Female Body in Ancient Greece, London 1998.

King (1999): H. King, Chronic pain and the creation of narrative, in: J. Porter (ed.), Constructions of the Classical Body, Michigan 1999, 269–286.

Koentges (2020): T. Koentges, Measuring Philosophy in the First Thousand Years of Greek Literature. Digital Classics Online 6,2 (2020), 1–23, https://doi.org/10.11588/dco.2020.2.73197 (Last access 31.08.2021).

Lenci (2018): A. Lenci, Distributional Models of Word Meaning, Annu. Rev. Linguist 4 (2018), 151–71.

Levy et al. (2015): O. Levy / Y. Goldberg / I. Dagan, Improving Distributional Similarity with Lessons Learned from Word Embeddings, Transactions of the Association for Computational Linguistics 3 (2015), 211–25.

López / Ramero (2014): F. López / V. Romero, Mastering Python Regular Expressions, Birmingham 2014.

Mikolov et al. (2013): T. Mikolov / I. Sutskever / K. Chen / G. S. Corrado / J. Dean, Distributed Representations of Words and Phrases and their Compositionality, in: C. J. C. Burges et al. (eds.), Advances in Neural Information Processing Systems 26 (2013), 3111–3119.

Moretti (2013): F. Moretti, Distant Reading, London / New York 2013.

Rey (1995): R. Rey, The History of Pain, England 1995.

Role / Nadif (2011): F. Role / M. Nadif, Handling the impact of low frequency events on co-occurrence based measures of word similarity, in: J. Filipe et al. (eds.), Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR-2011), Scitepress 2011, 218–23.

Sahlgren / Lenci (2016): M. Sahlgren / A. Lenci, The Effects of Data Size and Frequency Range on Distributional Semantic Models. arXiv [cs.CL] 27 (2016), http://arxiv.org/abs/1609.08293 (Last access 31.08.2021).

Svärd et al. (2020): S. Svärd / T. Alstola / H. Jauhiainen / A. Sahala / K. Lindén, Fear in Akkadian Texts: New Digital Perspectives on Digital Semantics, in: S.–W. Hsu / J. L. Raduà (eds.), The Expression of Emotions in Ancient Egypt and Mesopotamia, Leiden / Boston 2020, 470–502.

Schütze / Pedersen (1993): H. Schütze / J. Pedersen, A vector model for syntagmatic and paradigmatic relatedness, Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research, Oxford 104–113.

Underwood (2017): T. Underwood, A Genealogy of Distant Reading, Digital Humanities Quarterly 11 (2017), 1–12.

van der Maaten / Hinton (2008): L. van der Maaten / G. Hinton, Visualizing Data Using T-SNE, Journal of Machine Learning Research: JMLR 9 (2008), 2579–2605.

van Rossum / Drake (2009): G. van Rossum / F. L. Drake, Python 3 Reference Manual, Scotts Valley 2009.

Wolfsdorf (2013): D. Wolfsdorf, Pleasure in ancient Greek philosophy, Cambridge 2013.

Author contact information[65]

Vojtěch Linka

Department of Philosophy and Religious Studies
Faculty of Arts
Charles University
nám. Jana Palacha 2
11638, Praha 1
Czech Republic
E-Mail: vojtech.p.linka@gmail.com

Vojtěch Kaše

School of Culture and Society – History
Jens Chr. Skous Vej 5
building 1463, 528
8000, Aarhus C
Denmark

E-Mail: kase@cas.au.dk

&

Department of Philosophy
Faculty of Arts
University of West Bohemia
Sedláčkova 19
30614, Plzeň
Czech Republic
E-Mail: kase@kfi.zcu.cz

65