Umanistica Digitale – ISSN: 2532-8816 – n. 23, 2026: Special Issue
D. Longrée, L. Vanni – *Hyperbase Web*
DOI: http://doi.org/10.60923/issn.2532-8816/24319

Hyperbase Web

A web based tool box for linguistic analysis: the example of the LASLA files

Dominique Longrée

Univ. Liège, L.A.S.L.A.
dominique.longree@uliege.be

Laurent Vanni

Univ. Côte d'Azur
laurent.vannni@univ-cotedazur

Abstract

This article presents the main features of *Hyperbase Web* based on the Latin databases of L.A.S.L.A. This software stands out above all for its tools that enrich the reading by statistical measurements on word distribution throughout texts, as well as its deep learning algorithm applied to text analysis. After describing the L.A.S.L.A. corpus implemented in the various Latin databases, we will present the documentary and static functions accessible from the *Hyperbase Web* interface. Each main feature will be illustrated with machine output and enriched with linguistic interpretation suggesting complementary reading paths.

Keywords: Latin databases; search engine; intertextual distances; text analysis; deep learning

Questo articolo presenta le caratteristiche principali di *Hyperbase Web* basato sui *database* latini di L.A.S.L.A. Questo *software* si distingue soprattutto per i suoi strumenti che arricchiscono la lettura con misurazioni statistiche sulla distribuzione delle parole nei testi, nonché per il suo algoritmo di *deep learning* applicato all'analisi del testo. Dopo aver descritto il *corpus* L.A.S.L.A. implementato nei vari *database* latini, presenteremo le funzioni documentarie e statiche accessibili dall'interfaccia *Hyperbase Web*. Ogni caratteristica principale sarà illustrata con *output* automatico e arricchita con interpretazioni linguistiche che suggeriscono percorsi di lettura complementari.

Parole chiave: banche testuale latine; motore di ricerca; distanze intertestuali; analisi testuale; deep learning

1. Introduction

Statistical analysis of textual data is one of the major challenges in corpus linguistics. There are many software programs available, but they do not all share the same methods and resources. *Hyperbase Web* offers a corpus-driven approach [12] based on exploratory statistical tools. The software allows users both to query private corpora (by creating personal databases) and to consult a large collection of ready-to-use databases regrouping texts in different languages according to various themes such as literature, political discourse, and media discourse.

While the interface is designed around classic search engine ergonomic principles, *Hyperbase Web* stands out above all for its tools that enrich the reading by statistical measurements on word distribution throughout texts, as well as its deep learning algorithm applied to text analysis. The software uses both local methods (co-occurrence calculation) and global methods (factorial correspondence analysis) and offers various representations of texts and corpuses.

This article presents the main features of *Hyperbase Web* based on the Latin databases of L.A.S.L.A. (Laboratoire d'Analyse statistique des Langues anciennes – UR mondes Anciens). Using all the morphosyntactic annotations included in the reference corpus, these databases combine the technical expertise implemented by the software with the philological expertise provided by the University of Liège. The Latin databases of *Hyperbase Web* illustrate the many possibilities of the software and offer the community a new perspective on the analysis of classical Latin.

After describing the L.A.S.L.A. corpus implemented in the various Latin databases, we will present the documentary and static functions accessible from the *Hyperbase Web* interface. Each main feature will be illustrated with machine output and enriched with linguistic interpretation suggesting complementary reading paths.

2. Corpus

Since 1961, the L.A.S.L.A. has created a large database of classical Latin texts (constantly expanding: over 2,000,000 words for the time being). The database is currently the only one that includes, for each of the text's units, the lemma and the complete morphological analysis systematically checked by a philologist. *Hyperbase Web* contains several regularly updated databases that group LASLA texts according to various criteria: the "LASLA database", which contains all available texts; the "Latin database", which contains a selection of the most significant texts, grouped or divided to form sets of sizes suitable for statistical analysis; and various databases containing texts by the same author (Plautus, Cicero, Tacitus) or of the same literary genre, also prepared in the same way.

The traditional structure of the LASLA filing system is as follows:

1. The lemma, such as it is found in the dictionary chosen as a reference work (*Lexicon totius latinitatis de Forcellini*, ed. de Corradini, Padoue, 1864);
2. An index that allows us to distinguish different homographic lemmas, or to mark the proper names and adjectives that are derived from them;

3. The form such as it appears in the text;
4. The reference, which conforms to the rules of the *ars citandi*;
5. A complete morphological analysis in an alphanumeric format; in other words, for example, for a noun, the declension, case and number, or for a verb, conjugation, voice, mood, tense, person and number;
6. For verbs, syntactical indications; main clauses are distinguished from subordinate ones, and the latter are classified by type of subordination.

In certain cases, the number of indications attached to a word reaches as high as ten. So, for a form such as *regnante*, are indicated: the reference data, the lemma, the grammatical category, the conjugation, the voice, the case, the number, the mood, the tense and the gender are all indicated. Finally, we will indicate, if necessary, when such a verb is the predicate of an ablative absolute. The software *Hyperbase Web* allows searches on every type of information: lemma, index, every morphologic and syntactical feature.

3. Empirical approaches to texts

The recommended entry point for *Hyperbase* is the "Edition" function. It displays the corpus profile and provides some initial measurements performed automatically by the software. The corpus editing window contains the minimum statistical information you need to know before undertaking a study with *Hyperbase*, namely the size of the corpus in terms of number of words, vocabulary size, and number of hapaxes according to the contrast (metadata) chosen. This interface also allows you to edit certain information relating to the description of the database, namely its name, title, and author, as well as to export data, share the database, or delete it permanently.

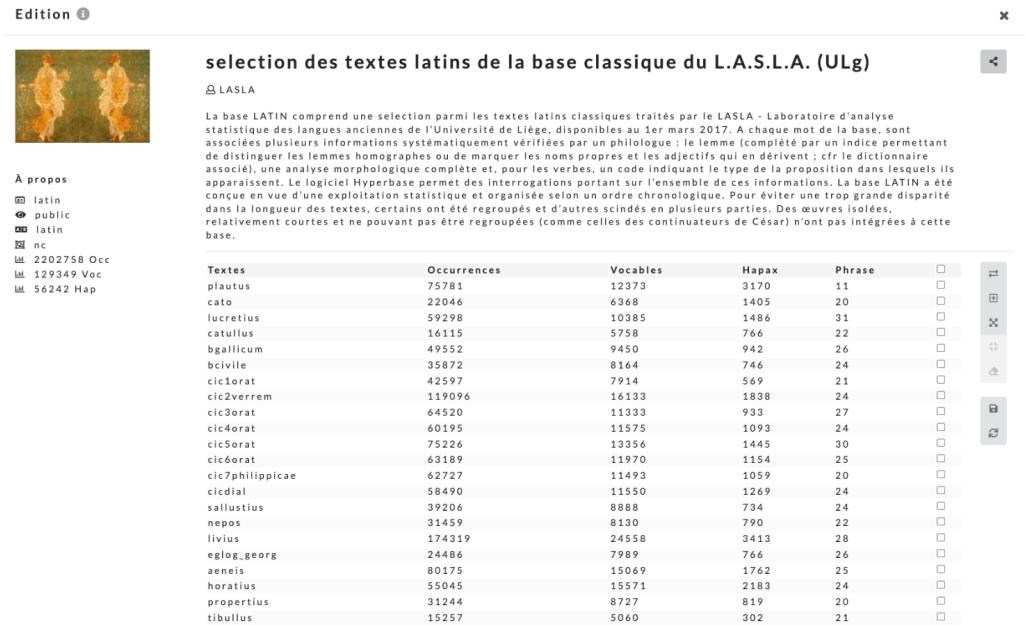


Figure 1 - Menu Edition – Base Latin

Figure 1 shows the profile of the Latin database, whose structure, as mentioned above, allows users to explore the statistical differences between each of the text sets constituting the database. The metadata collected for each text allows users to modify this view from the interface and switch, for example, to an analysis by work, book, or author. The editing tools offered by *Hyperbase* allow the available metadata to be cross-referenced at any time to create new representations of the corpus and test new hypotheses. The central table shows the status of the corpus, which must be kept in mind when a statistical calculation is performed in the interface. The overall size of the corpus and the size of each part are essential pieces of information when using statistical methods. When changes are applied, the entire indexing of the corpus is modified, and all results are recalculated to take these changes into account.

As a prelude to any analysis, the "Edition" menu allows you to check the status of the corpus (number of word forms, lemmas, "hapax" defined as lemmas appearing only in one of the text of the corpus) and presents some statistical analyses on data available from this menu. The "Richness" column, for example, reports on lexical richness in hapaxes; more precisely, the difference between the actual number of hapaxes in each part and the theoretical number (see the technical note on calculating specificities in section 3.2).

3.1 Intertextual distance

Another exploratory function that can be queried as soon as the database is loaded is the "Distance" function. This function corresponds to the intertextual distance calculation proposed by [2]. This calculation, combined with a tree analysis, reveals a structure in the corpus organized according to the selected metadata. Each leaf of the tree represents a part of the corpus, and each intermediate node of the tree constitutes a step in the distance separating two parts. The longer

this path (the greater the number of nodes and the longer the branches) between two leaves, the greater the distance between two parts of the corpus. The interpretation of this distance is based on the choice of calculation, in this case a variant of the Jaccard distance, which causes two leaves of the tree to move closer to the same node if their proportion of common vocabulary increases.

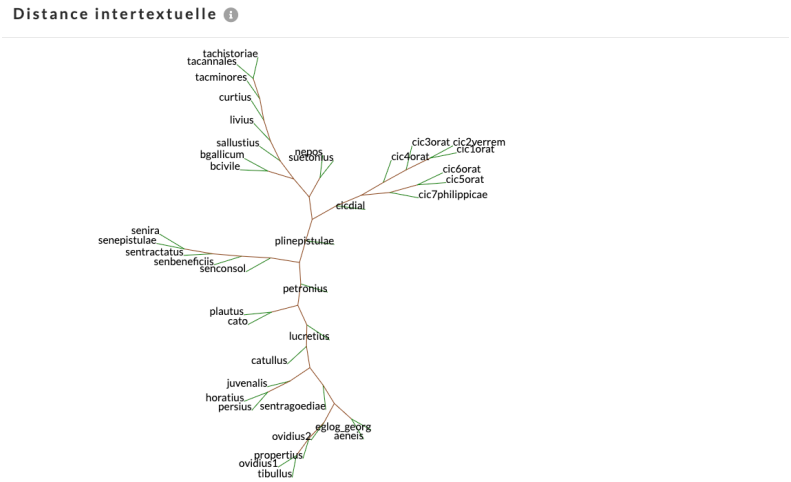


Figure 2 - Intertextual Distance – Base latin

Figure 2 shows this distance based on the Latin corpus. This figure illustrates above all a contrast between poetry in the lower part of the tree and prose in the upper part. In the upper part, one branch groups together all the historical texts and highlights the proximity of biographies (Nepos – Curtius). The other two branches each correspond to a different author: Cicero and Seneca.

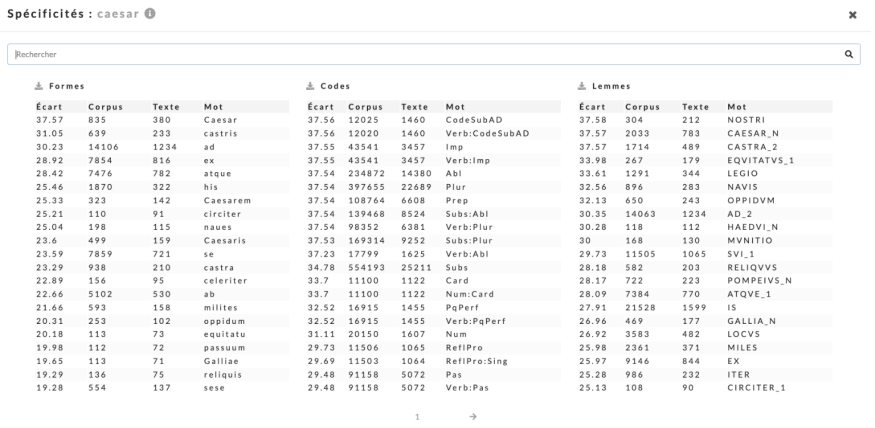
Regardless of the metadata selected, intertextual distance is a good way to assess the balance of the corpus. The tree generated allows us to identify a robust structure or, conversely, unexpected alterations; it is a simple and effective way to begin a study of a corpus based on its lexical topology [9]. While this function provides a quick overview of the organization of data in the corpus, it is generally combined with other complementary analyses that go beyond a simple comparison of the presence or absence of words in the texts.

3.2 Lexical specificities

Lexical specificities are at the origin of the lexicometric method. They are commonly based on a calculation of the difference between theoretical and actual numbers [11].

Lexical specificities are accessible once a text has been selected in the interface (from the drop-down list). The suite of statistical tools adapts to the data requested. The intertextual distance (across the entire corpus), previously displayed in the main menu, gives way to specificities when the researcher focuses on a part of the corpus. The table that appears then gives a list of specific words, i.e., words that are overused in the text compared to an average distribution. This vocabulary list represents the first real contrasts that can be seen in the corpus. These linguistic markers are the observable phenomena that the working hypothesis, i.e., the partitioned corpus,

allows us to study. The list is not filtered but is sorted by decreasing deviation. All tokens in the corpus are subject to calculation, which generally leads to a mixture of nouns, verbs, speech markers, function words, symbols of all kinds, etc., which share the characteristic of being overrepresented in the text compared to the corpus as a whole.



Spécificités : caesar

Rechercher

Formes

Écart	Corpus	Texte	Mot
37.57	835	380	Caesar
31.05	639	233	castris
30.23	14106	1234	ad
28.92	7854	816	ex
28.42	7476	782	atque
25.46	1870	322	his
25.33	323	142	Caesarem
25.21	110	91	circiter
25.04	198	115	naues
23.6	499	159	Caesaris
23.59	7859	721	se
23.29	938	210	castra
22.89	156	95	celeriter
22.66	5102	530	ab
21.66	593	158	milites
20.31	253	102	oppidum
20.18	113	73	equitatu
19.98	112	72	passuum
19.65	113	71	Galliae
19.29	136	75	reliquis
19.28	554	137	sese

Codes

Écart	Corpus	Texte	Mot
37.56	12025	1460	CodeSubAD
37.56	12020	1460	Verb:CodeSubAD
37.55	43541	3457	Imp
37.55	43541	3457	Verb:Imp
37.54	234872	14380	Abl
37.54	397655	22689	Plur
37.54	108764	6608	Prep
37.54	139468	8524	Subs:Abl
37.54	98352	6381	Verb:Plur
37.53	169314	9252	Subs:Plur
37.23	17799	1625	Verb:Abl
34.78	554193	25211	Subs
33.7	11100	1122	Card
33.7	11100	1122	Num:Card
32.52	16915	1455	PqPerf
32.52	16915	1455	Verb:PqPerf
31.11	20150	1607	Num
29.73	11506	1065	ReflPro
29.69	11503	1064	ReflPro:Sing
29.48	91158	5072	Pas
29.48	91158	5072	Verb:Pas

Lemmes

Écart	Corpus	Texte	Mot
37.58	304	212	NOSTRI
37.57	2033	783	CAESAR_N
37.57	1714	489	CASTRA_2
33.98	267	179	EQVITATVS_1
33.61	1291	344	LEGIO
32.56	896	283	NAVIS
32.13	650	243	OPPIDVM
30.35	14063	1234	AD_2
30.28	118	112	HAEDVI_N
30	168	130	MVNITIO
29.73	11505	1065	SVI_1
28.18	582	203	RELIQVVS
28.17	722	223	POMPEIVS_N
28.09	7384	770	ATQVE_1
27.91	21528	1599	IS
26.96	469	177	GALLIA_N
26.92	3583	482	LOCVS
25.98	2361	371	MILES
25.97	9146	844	EX
25.28	986	232	ITER
25.13	108	90	CIRCITER_1

1

Figure 3 - Intertextual Distance – Base latin

Figure 3 illustrates the results of this method for the partition "caesar". From the point of view of the word forms, the list of the specificities highlights word forms appearing in typical Caesarian expressions like *Caesar ex castris…* or *circiter passuum*, and most of the specific lemmas are related to the military terminology (*castra, equitatus, legio…* as the possessive *nostri*, with the meaning "our troops). While the method used here is primarily lexical, it can easily be extended to a more formal linguistics by adding grammatical categories and lemmas, which lead to an exploration of other contrasts : in terms of grammatical codes, the methods allows some specific morphologic an syntactical features (absolute ablatives, imperfect tense, ablative, nouns, verbs, and numerals) to be identify as highly distinctive of the caesarian style.

This robust approach for detecting contrasts in the corpus is not limited to simply counting occurrences. It can easily be extended to detect more complex linguistic phenomena such as specific co-occurrences. There are many approaches to assessing co-occurrence phenomena within corpora [5], i.e., the particular associations between words that distinguish their usage. The first exploratory function of co-occurrences offered by the software is based on generalized co-occurrences and is called "Corrélats – Correlate".

3.3 Correlates

"Corrélats – Correlates" is one of the software's historical functions [3]. The idea promotes a semantic visualization of words based on their co(n)textual distribution [18]. Co-occurrence thus offers analyses that reveal semantic, thematic, or ideological choices throughout the corpus. Correlates is one of the tools that allows for a global representation of the organization of words, whether at the level of the entire corpus or of each metadata taken separately.

Technically, the tool first sorts words by frequency to retain only the n most frequent words (n being a parameter that can be adjusted in the *Hyperbase* interface). This selection allows the

analysis to focus on shared words and avoid empty matrices that would give rise to distortions linked to marginal phenomena that are difficult to interpret. It should be noted that this automatic cleaning is fully configurable and the morphosyntactic labeling of words allows filtering by grammatical categories (for example, the top 100 verbs).

Secondly, a word×word matrix (square table) shows the frequency of occurrence of each pair in the corpus, considering a sliding contextual window. The nature and size of this window is part of *Hyperbase*'s general settings and uses 10 words by default. This choice can be changed to another value (20, 30, 50 words, etc.) or by using sentences (strong punctuation marks) or paragraphs (line breaks) as the basic context.

Finally, the data generated by this type of information extraction is analyzed using Correspondence Analysis (CA) to compress the information and reproduce it on two axes [1]. This is a slightly different use of PCA than traditional applications, as there are no words/texts (rectangular data) to visualize, but only words/words (square data). In addition to PCA, hierarchical classification is applied to the table (a method similar to that used for tree analysis). The classes selected (set by a number chosen in the interface options) are represented by different colors on the graph. The final representation obtained produces a lexical map of the texts and provides a quick visualization of the general co-occurrence and specific co-occurrence profiles within the selected texts (Figure 4).

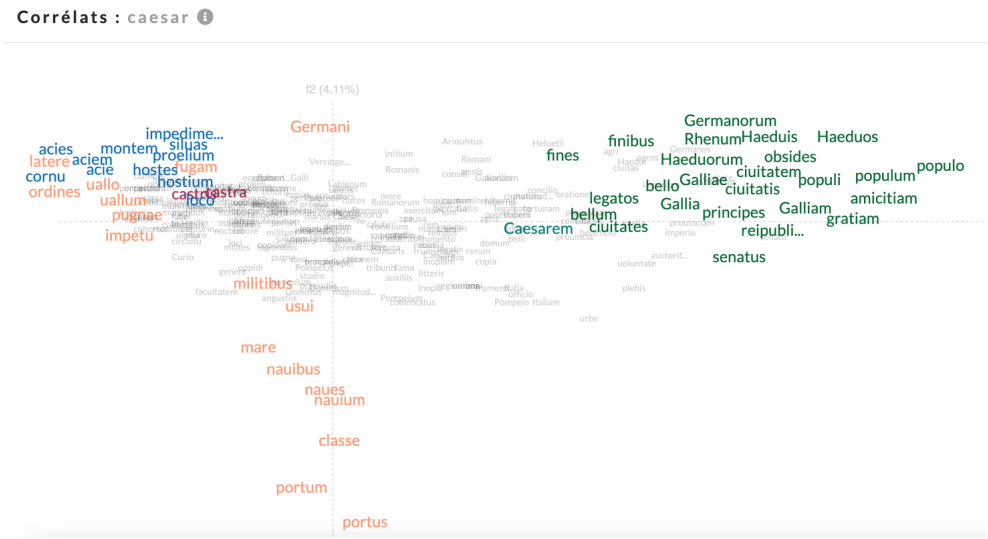


Figure 4 - Correlates in Caesar – Base latin

Figure 4 illustrates the application of the method to Caesar's work. Here, the analysis focuses on the 300 most frequent nouns. This overview reveals two poles on the horizontal axis: on the right, terms related to political relations; on the left, terms related to military maneuvers. On the vertical axis, there is a clear opposition between terms related to the fleet and all other terms.

The "Corrélats – Correlates" function therefore provides food for thought on the lexical content and its thematic-semantic organization as a whole. Another function, the function "Associations", allows to take into account the specific arrangements of words in texts.

3.4 Associations

Unlike "Correlates," which is based on established co-occurrence profiles (the word×word matrix), the "Associations" function is based solely on the statistical assumption of the co-presence of two words in context. Similar to the statistical calculation of specificities, this approach uses a theoretical co-occurrence calculation between two words that uses two probabilities: the probability of the absence of the first word in a text segment and the probability of the absence of the second word. The probability is obtained from the same law used to calculate specificities, the hypergeometric law. Based on the frequency of the word in the corpus f, the size of the corpus T, and the size of the context t chosen to observe co-occurrence (10 words by default), the probability is given by the formula:

$$p = \frac{f!\,(T-f)!\,t!\,(T-t)!}{f!\,t!\,(T-f-t)!\,T!}$$

This is a simplification of the calculation presented in section 3.2 (see also [11]) (with k=0). By symmetry, the inverse probability q=1-p gives the chances of finding each word in the context regardless of their frequencies. Finally, the product of the two results gives the chances of finding both words together in the same context:

$$q = q_1 \times q_2$$

The theoretical frequency of a pair is then obtained by multiplying this result by the number of contexts T/t. This frequency is finally compared to the actual frequency within a text, and a difference is then obtained that reflects the strength of attraction between these two words in the text. This process, unique to *Hyperbase*, provides a list of statistically significant associations related to metadata and allows for another form of visualization that illustrates strong co-occurrence links within texts.

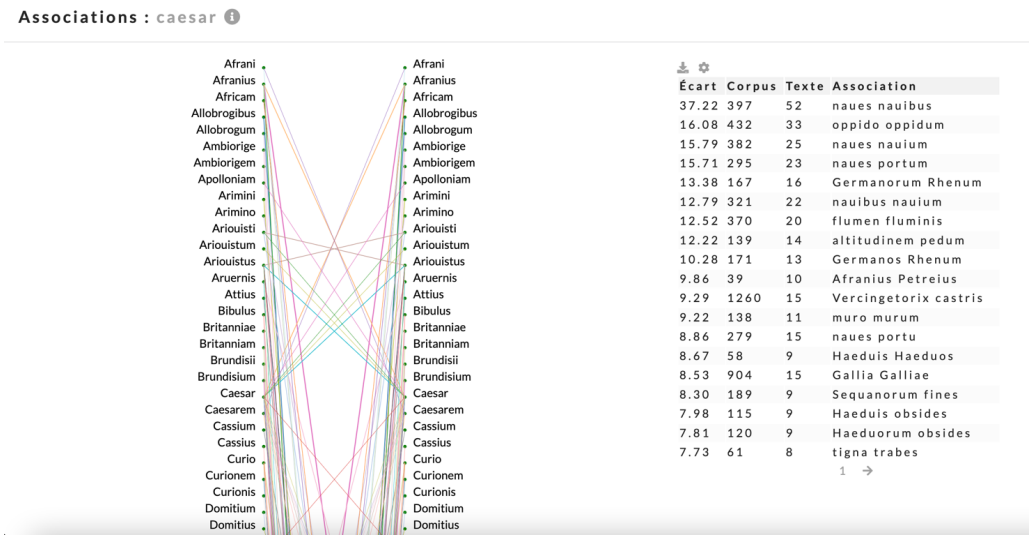


Figure 5 - Associations in Caesar – Base latin

The preferred associations noted in Caesar's text (Figure 5) are, in some cases, associations within expressions such as *murum in altitudinem pedum sedecim perducit* (*Gall.* 1, 8, 1), but they also illustrate a tendency in Caesar to repeat the same words at short distance (*naues nauibus*)

We can see that moving from occurrence to co-occurrence provides a more detailed level of analysis than a decontextualized lexicon. In reality, it is the combination of all the methods offered by ADT that provides researchers with rich interpretative pathways and precise analyses. Each tool is based on a particular representation of texts, which has both advantages and limitations. And the automatic cross-referencing of all these representations remains difficult for traditional statistics. Therefore, in order to broaden the methodological field and enable this cross-referencing, *Hyperbase* proposes the use of deep neural networks dedicated to data description.

3.5 Deep learning

In the humanities and social sciences, deep learning or deep neural network training corresponds to black box methods offering tools with often intriguing performance [7]. In linguistics, many automatic tasks such as translation, classification, and generation have been significantly improved. With *Hyperbase*, and when applied to texts, deep learning questions the intelligibility of AI. Indeed, the primary purpose of ADT, namely the description of the corpus, is not the same as that of deep learning, which focuses more on pure performance. However, considering automatic classification as the predictive counterpart of contrastive corpus analysis, *Hyperbase* aims to extract textual information encoded in the hidden layers of neural networks in order to 1) explain model performance and 2) discover new linguistic markers that question the contrasts that determine each part of the corpus.

To be interpretable, the architecture used is based on layers of standard neurons associated with information extraction methods. Several types of layers are used to combine different types of

text abstraction. More specifically, three approaches are used. An embedding layer responsible for vectorizing words (similar to an AFC); a Convolution layer that modifies this representation based on the co(n)text of the words; and an Attention layer [17] that specializes in detecting distant links between words. *Hyperbase* offers the unique feature of combining each of these approaches in a single architecture, called *Multichannel Convolutional Transformer* (MCT) ([14],[16]), where *Multichannel* means that words are considered as both a graphic form, a grammatical category, and a lemma. This system guarantees the extraction of complex linguistic markers used by the model to perform an automatic classification task (e.g., predicting an author). As with all traditional ADT methods, the applicability of deep neural networks depends on the organization of the corpus. The metadata chosen represents the classes to be distinguished for the AI. Learning is based on the corpus studied, and the accuracy of the model depends on the organization of the corpus. The metadata chosen represents the classes to be distinguished for the AI. Learning is based on the corpus studied, and the accuracy of the model depends on the volume and homogeneity of the data.

The availability and accuracy of the AI on the loaded corpus can be seen from the *Hyperbase* main menu. If the number of words is sufficient and the accuracy acceptable, a Hyperdeep entry is added to the main menu.

One of the main functions of deep learning applied to text is prediction. Using a new text that is not part of the training corpus, it is possible to ask the AI to predict the class, i.e., to provide the most likely metadata associated with the corresponding text segment. In addition to the final functionality, which can be useful in areas such as text authorship research, the description provided by the interface reveals a different use for ADT. *Hyperbase* offers researchers the opportunity to explore the language markers that prevail in the AI's class determination. Whether lexical, semantic, selective, or associative, these markers are generally distinguished by their particular combinations and add an additional level of analysis for ADT. With the MCT architecture, *Hyperbase* embraces text in all its possible representations. The structure of language is covered both on the syntagmatic axis and the paradigmatic axis by "Convolution" and "Attention", which capture particular associations in the present (based on the organization of words in the text to be predicted) and in memory (based on relationships learned from the training corpus).

One of the uses of text prediction is the detection of intertextuality [6]. By projecting a target author into a corpus containing a set of other source authors (who may have inspired the former), it is possible to assess the passages attributed to each as traces of intertextuality. By using the description of markers provided by *Hyperbase*, these traces become clearer and the work of interpretation becomes more specialized. Each word that contributed to the decision-making process is highlighted in the text, and the links between the selected words are represented by colored lines.

Figure 6 shows an example of key passage extraction from Caesar in the Latin database.

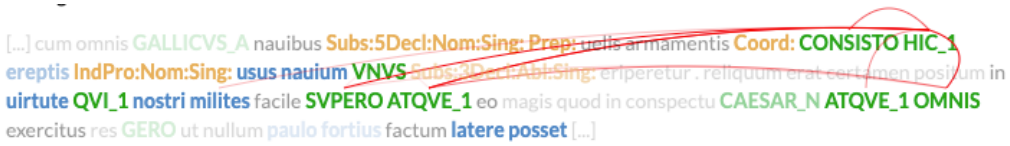


Figure 6 - Key passage from the Latin text attributed to Caesar

This excerpt presented illustrates how deep learning highlights the elements that structure the cesarean text (qui, atque, atque) and to detect "deep motifs" [15].

Deep learning necessarily involves returning to the text in order to analyze it correctly. Whether manually or using sophisticated statistical calculations, verifying certain observations is useful for confirming certain phenomena. To achieve this, *Hyperbase* is linked to a search engine, a means of targeting specific linguistic markers in order to assess their uses.

4. Search engine

The main database query engine consists of a central search field with three tabs corresponding to three different approaches to corpus search (Figure 9). The first tab, which is the default, corresponds to documentary "Search," i.e., a return to the raw text presented in the form of a concordance. The second tab uses co-occurrence analysis to display the "Theme" of the selected word or expression. The last tab displays the statistical "Distribution" of the requested search across the metadata that make up the corpus. All these functions are based on a query system that ranges from simple keywords to complex (regular) expressions. This engine easily combines graphic forms of words, grammatical categories, or lemmas, whether contiguous or not. Several expressions can be grouped together in the same search by using quotation marks to delimit each one. This ad hoc search system is designed to be broad enough to cover most linguistic questions asked in ADT without resorting to computer language that is discouraging for the humanities. Details on how the query engine works are available from the interface in the "Advanced Search" entry in the main menu. Finally, general "Settings" extend the search engine's capabilities. In particular, it is possible to change the representation of words by grouping them by lemmas or morphosyntactic tags. The corpus can also be filtered by grammatical categories to ensure that the statistical calculations take linguistic considerations into account. The context window can also be manipulated to calculate co-occurrences or use wildcards associated with queries.

4.1 Concordance creation

The concordance tool, a classic tool used by philologists, is a means of weighing certain hypotheses through qualitative machine outputs that can be directly interpreted by humans [10]. The tool presents the search results in the form of a table containing three columns representing the left context, the pivot (the search object), and the right context. This visualization can be sorted alphabetically from right to left or left to right and allows for a (manual) analysis of the redundancies displayed. To this traditional view, *Hyperbase* adds a fourth column to display metadata related to the outputs, and a return to the full text with a click on the line in question.

This contextual view of a word or expression is generally a recurring step in Text Data Analysis. Statistical interpretations are thus regularly reinforced by concrete examples in the text and vice versa. For example, the expression "*his rebus*".

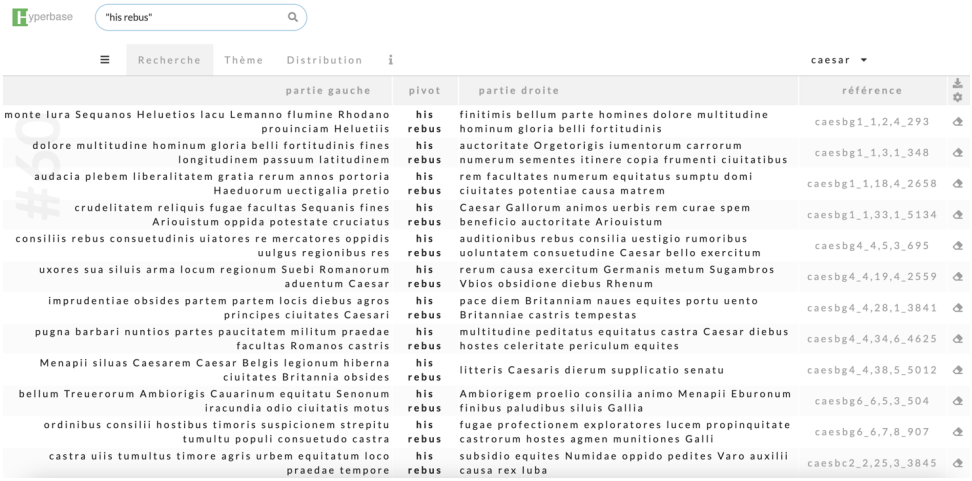


Figure 7 - Concordance of the string *his rebus* in Caesar

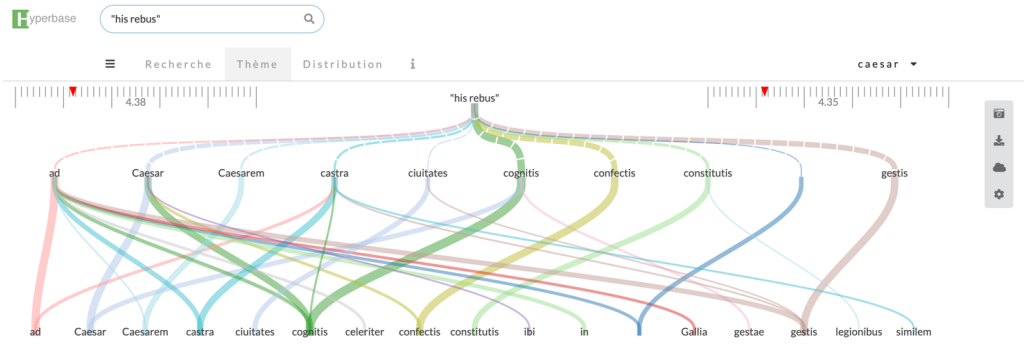
This particular reading of the text is a valuable aid in identifying all occurrences of textual motifs such as *his rebus gestis, his rebus cognitis*, etc. [4].

Navigation by tabs allows you to switch from a plain text "Search" (corresponding to the query) to frequency representations for the same query by clicking on either 'Theme' or "Distribution".

4.2 Thematic search

Along with the "Correlates" and 'Associations' functions, the "Theme" tab is the third type of co-occurrence analysis offered by *Hyperbase*. The Theme corresponds to the calculation of specific co-occurrences, allowing you to list words that are statistically overrepresented around the chosen word or expression. The result takes the form of a graph associated with an exhaustive list of terms whose specificity score exceeds 2 (minimum specificity threshold). Two visualizations are possible. The first takes a hierarchical form on two levels, representing poly-co-occurrences (words which co-occur with fixed co-occurrence pairs; see [13]). The second, more traditional visualization displays a word cloud representing all direct co-occurrences. In both cases, the size of the words or links is proportional to the specificity score. These visualizations provide a global overview of co-occurrence that is no longer static. The interface is dynamic and adapts to the researcher's journey. Some links are highlighted and others fade when the user points to a particular word. To interpret the results, the text can be accessed at any time from the interface, which is then divided between statistical outputs and plain text.

This function has additional uses in text analysis. It gives semantic weight to the words searched for by associating them with specific lexical fields. It reveals stylistic or ideological vocabulary choices in the case of political or media discourse analysis. The use of filters or changes in word representation also allows for more complex linguistic queries. In particular, it is possible to study the variation of certain grammatical categories or even to query syntax by playing with the "Parameters" to study the use of certain parts of speech around a particular syntactic sequence (before, after, or bidirectional). This advanced co-occurrence analysis function addresses questions related to associative criteria in language of all kinds.



Indice de spécificité	Probabilité	Fréquence corpus	Fréquence partie	Taille corpus	Taille partie	Mot
9.27	2.26e-19	69	9	2202758	1200	cognitis
7.18	1.24e-12	22	5	2202758	1200	confectis
6.84	1.17e-11	86	6	2202758	1200	gestis
6.06	1.42e-9	835	9	2202758	1200	Caesar
5.19	1.82e-7	20	3	2202758	1200	constitutis
4.84	0.00000111	938	7	2202758	1200	castra

Figure 8 - Thème of the string *his rebus* in Caesar

Fig. 8 shows how a statistical calculation can confirm the existence of textual motifs constructed on the association of the expression *his rebus* and a passive perfect participle *cognitis, confectis, gestis, constitutis*...

4.3 Statistical distributions

The last tab of the *Hyperbase* search engine displays the statistical "Distribution" of words. This function cross-references the words or expressions searched with the texts in a contingency table where the columns represent the metadata and the rows represent the units observed. This type of multivariate analysis can be approached in different ways depending on the density of the table A simple search for a single term is generally assessed using a classic histogram, which allows the measurement to be visualized. In the case of *Hyperbase*, this involves measuring the absolute or relative frequencies, as well as calculating specificities (see section 3.2.), and, in this last option, the graph shows the positive or negative deviations for each metadata item.

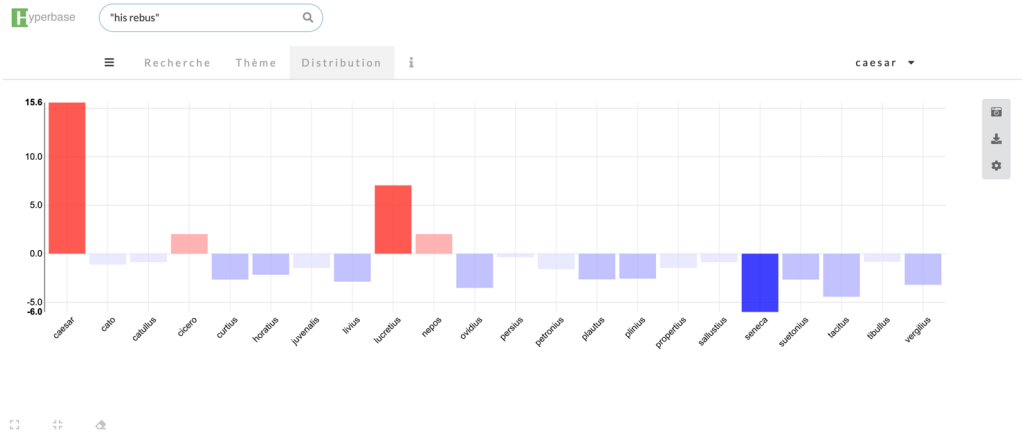


Figure 9 - Distribution of the string *his rebus* in the base Latin

Figure 9 illustrates how much the string *his rebus* is characteristic of the *Caesar language*.

Hyperbase's search engine allows you to analyze one or more words at a time, either by listing them in the search field or using a grammatical code and requesting details via the data table located below the graph. Several icons appear above the table header, allowing you to manipulate the selected rows (checkbox at the beginning of the row). Among the actions available are "Display most frequent," "Group words," and "Delete words from list." All of these actions modify the measurements taken and change the visualization according to your needs. The contingency table representation automatically switches from a histogram to an AFC when the number of columns or rows exceeds 50 entries. The tools available on the graph (vertical menu to the right of the graph) allow you to switch between views at any time and use dedicated functions such as displaying additional elements on an AFC. Note that it is possible to call up certain additional analyses, such as tree diagrams, using these same tools if the number of texts allows it (> 4 for a tree diagram or an AFC).

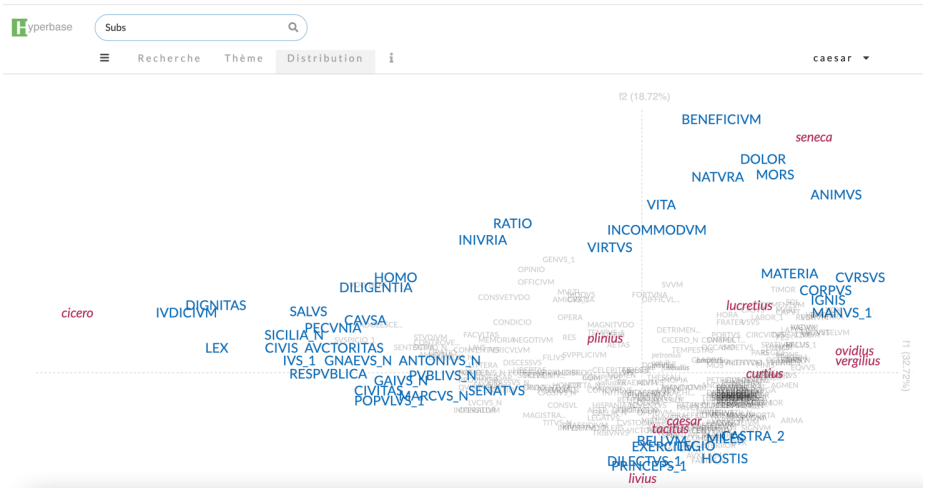


Figure 10 - AFC of the most frequent substantives in the base Latin

The AFC of the most frequent subtantives in the base Latin (Figure 10) pinpoints the lexical proximity of the historians in the lower part of the figure or of the poets Ovid and Vergil on the right, and the isolation of authors like Cicero and Seneca, with a quite specific vocabulary (as for instance*dignitas, iudicium, lex* for Cicero and *beneficium, dolor, mors* for Seneca)

Far from being exhaustive, these examples provide an overview of the range of tools offered to researchers by *Hyperbase*. The search engine, which includes text return, co-occurrence analysis, and statistical distribution, offers virtually unlimited possibilities for analyzing text, metadata, and the corpus in general.

5. Conclusion

Hyperbase Web offers researchers a suite of Textual Data Analysis tools, ranging from the most basic to the most sophisticated. Advances in computing enable increasingly sophisticated analyses and are driving constant paradigm shifts. However, the software is not intended to objectify language, but simply to describe texts. Human language remains a complex phenomenon that needs to be materialized in order to be studied. The software chooses words as its raw material and the corpus as its object of study. The interface offers a comprehensive interpretative journey, maintaining a permanent link between quantitative measurement and qualitative text feedback. Built around an advanced search engine, the linguistic observables extracted from the texts are put into perspective by graphical means that illustrate the organization of words in the corpus. These textual artifacts primarily express statistical realities that do not seek to naturalize humans and culture (nor to humanize the machine) but rather to move toward scientific analyses of language in action, through descriptions of specific idiolectal or sociolinguistic phenomena.

References

[1] Benzecri, Jean-Paul. 1973. *L'Analyse des données. Tome 2: L'Analyse des correspondances*, Dunod.

[2] Brunet, Étienne. 2003. "Peut-on mesurer la distance entre deux textes?". *Corpus* 2: 47-70.

[3] Brunet, Étienne. 2010. *HYPERBASE Manuel de référence*.

[4] Longrée, Dominique, and Sylvie Mellet. 2013 "Le motif: une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours". *Langages* 189: 65-79.

[5] Mayaffre, Damon. 2012. *Le discours présidentiel sous la Ve République. Chirac, Mitterrand, Giscard, Pompidou, de Gaulle*. Honoré Champion.

[6] Mayaffre, Damon, and Laurent Vanni. 2020. "Objectiver l'intertexte ? Emmanuel Macron, deep learning et statistique textuelle". *JADT 2020 - 15èmes Journées Internationales d'Analyse statistique des Données Textuelles, Jun 2020, Toulouse*: ffhal-02894990f.

[7] Mayaffre, Damon, and Laurent Vanni. 2021 (eds). *L'intelligence artificielle des textes. Des algorithmes à l'interprétation*, (Lettres Numériques 15), Honoré Champion.

[8] Mayaffre, Damon, and Jean-Marie Viprey. 2012. "La cooccurrence. Du fait statistique au fait textuel". *Corpus* 11: 7-19.

[9] Mellet, Sylvie, and Jean-Pierre Barthélemy. 2007, "La topologie textuelle : légitimation d'une notion émergente". *Lexicometrica* 7. http://lexicometrica.univ-paris3.fr/numspeciaux/special9/mellet.pdf.

[10] Pincemin, Bénédicte, Fabrice Issac, Marc Chanove, and Michel Mathieu-Colas. 2006. "Concordanciers : thème et variations". *JADT 2006,* edited by Jean-Marie Viprey: 773-784.

[11] Pincemin, Bénédicte. 2024. "Specificities and other applications of the Fisher exact test to textual data: What's the matter with lexical frequencies?". In *JADT 2024 - 17th International Conference on Statistical Analysis of* Textual Data, UCLouvain, Site Saint-Louis, June 2024, Bruxelles: 703-712.

[12] Tognini-Bonelli, Elena. 2001. "Corpus linguistics at work". *Computational Linguistics*, 28: 583–583.

[13] Vanni, Laurent, and Adiel Mittmann. 2016. "Cooccurrences spécifiques et représentations graphiques, le nouveau" Thème " d'Hyperbase". In *JADT 2016 - Statistical Analysis of Textual Data, Jun 2016, Nice, France.* 295-305.

[14] Vanni, Laurent, Melanie Ducoffre, Damon Mayaffre, Frederic Precioso, Dominique Longrée, Veeresh Elango, Nazly Santos Buitrago, Juan Gonzalez Huesca, Luis Galdo,

and Carlos Aguilar. 2018a. “Textual deconvolution saliency TDS: a deep tool box for linguistic analysis”. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers.* 548–557.

[15] Vanni, Laurent, Damon Mayaffre, and Dominique Longrée. 2018b. “ADT et deep learning, re-gards croisés. Phrases-clefs, motifs et nouveaux observables”. *JADT 2018 – Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*. 459-466.

[16] Vanni, Laurent, Hadi Mahmoudi, Dominique Longrée. and Damon Mayaffre. 2024. “Multichannel Convolutional Transformer and Intertextuality: A Latin Case Study”. In *New Frontiers in Textual Data Analysis*, edited by Giuseppe Giordano and Michelangelo Misuraca. Springer Nature. 197-207.

[17] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasx Kaiser, and Illia Polosukhin. 2017. “Attention is all you need”. In *Proceedings of the 31st International Conference on Neural Information Processing Systems NIPS'17*.

[18] Viprey, Jean-Marie. 2006, “Structure non-séquentielle des textes”. *Langages* 163: 71-85.