Category theory, Document Analysis, and Philological Operations. A formal approach: limitations, and challenges.

Riccardo Del Gratta[1,✉] , Simone Zenzaro[1], Angelo Mario Del Grosso[1] and Federico Boschetti[1,2]
[1]Institute for Computational Linguistics "A. Zampolli", Italian National Research Council, Pisa, Italy
Email: riccardo.delgratta@ilc.cnr.it, simone.zenzaro@ilc.cnr.it, angelo.delgrosso@ilc.cnr.it, federico.boschetti@ilc.cnr.it
[2]Venice Centre for Digital and Public Humanities (VePDH)
Email: federico.boschetti@unive.it

***Abstract*—Starting with a formal definition of the process of scholarly editing, we further formalize it by exploiting Category Theory. We then apply this formal model to automated Natural Language Processing tools, highlighting the parallelism between composition and associativity and linguistic tool pipelines. We then discuss the notion of interoperability between tools. Finally, we propose some future improvements to the proposed formal model.**

***Index Terms*—Interoperability, NLP, Linguistic and philological operations, Category Theory, Formal Models**

I. Introduction

In our earlier papers [1]–[3], we defined a formal model to describe the complexity of scholarly editing processes. The model commences with the formalization of electronic *documents*, associated *operations*, and collections of documents at specific temporal instances named *base-spaces*. Within this framework, we emphasized that *operations* on documents impact both linguistic and philological aspects.

The model outlines the *evolution* of documents as a group of *operations* that boost their transitions from one *base-space* to the next. Besides, the model presents the *actors* as the entities (*performers*) accountable for executing the *operations*. An *actor* is broadly defined and can be associated with either a human agent or an automated tool, illustrating the diverse range of contributors involved in the evolutive process of documents.

In this paper, we review and refine our model by integrating the conceptual framework of Category Theory [4]. This integration represents significant progress in advancing the formalization of our model. The adopted strategy to connect the proposed model to Category Theory relies on identifying documents and operations as objects and morphisms within a category. (Readers may refer to [5] for an earlier investigation of this approach). In Sec. VII-B we will show a much deeper link between our model and Category Theory, namely the formal equivalence between the compatibility and interoperability among *operation* with the composition and associativity principles of Category Theory Furthermore, the introduction of Category Theory also reveals two weaknesses of the proposed model: (i) the relevance of the evolution path for the scholars; (ii) the need for tailored treatment of the role of the time.

Therefore, we introduce two categories and provide an exhaustive investigation of their advantages and limitations. In Sec. VI, we illustrate that combining composition and associativity with the explicit time labeling of the *base-spaces* generates a range of theoretical evolution paths for identical objects. It's essential to clarify that this aspect might be misinterpreted as an indication of the unavailability of missing or presumed sources instead of being correctly interpreted as an outcome derived from a theoretical framework.

In Sec. VII, we limit the model to Natural Language Processing (NLP) tools as *operations*. This decision is appropriate because using NLP tools, such as deriving part-of-speech tagging for a given text, has become an indispensable part of the modern philological process.

Likewise, it is plausible to conclude that the duration required to execute an operation utilizing NLP tools is negligible when compared with the decades or even centuries necessary for philological tasks. This element allows us to neglect the time labeling of the *base-spaces* and results in fundamental importance in establishing the suggested category.

Finally, we discuss some implications and future developments of our formalization.

II. A brief recap on document formalization

We will review some of the definitions and axioms outlined in the paper [2].

An electronic text document indicated as D in (1), is a conceptual entity that includes, at the very least, the following components: (i) a content (c), (ii) a format (f), and (iii) a set of para-textual layers $\{p_0 \dots p_k\}$:

$$D = D(c, f, \{p_0, \dots, p_k\}), \tag{1}$$

where, in (1), the content c expresses the quantity of information carried by D; f is the format used to *formalize* such information, and the set $\{p_0 \dots p_k\}$ represents additional para-textual layers that add more details to enhance the understanding of document D.

A base-space $\mathscr{B}_\tau$ is the term used to describe a collection of documents $\{D_1, D_2, \ldots, D_n\}$ that are accessed synchronously at a specific time, τ.

$$\mathscr{B}_\tau := \mathscr{B}(\{D_1, \ldots, D_n\}; \tau) \tag{2}$$

An operation op takes a subset of documents $S_t = D_{i_1}, \ldots, D_{i_n} \subseteq \mathscr{B}_t(D)$ as input and produces a new **single** document $D_j \in \mathscr{B}_\tau(D)$ as output, where $\tau > t$:

$$op(\underbrace{\{D_1, \ldots, D_n\}}_{S_t \subseteq \mathscr{B}_t(\{D\})}) = \underbrace{D_j}_{\{D_j\} \equiv S_\tau \subset \mathscr{B}_\tau(\{D\})}, \tag{3}$$

where D_j belongs to S_τ, a subset of $\mathscr{B}_\tau(\{D\})$. Actors carry out operations in (3). An actor denoted as *ac* can be a human or an automated tool.

Fig. 1 shows a conceivable evolution of the document D_i:

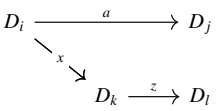

Figure 1: An *Evolution* graph from D_i to D_j and from D_i to D_l (through D_k).

We conceptualize an evolution $\mathscr{H}$ as a graph whose nodes correspond to the documents D, while the edges correspond to the agents that perform the operations, However, the above figure is only meant to illustrate the evolution graph. We know from (3) that operations act on subsets of documents and produce a new one. For example, the path $D_i \xrightarrow{a} D_j$ should be written as $\{D_i\} \xrightarrow{a} D_j$.

$\mathscr{H}$ is a collection of evolution steps ev_{st} (from source s to target t). We formalize the elements ev_{st} in Fig. 1 as follows:

$$\begin{aligned} ev_{ik} &= (\{D_i\}, D_k)_x \\ ev_{kl} &= (\{D_k\}, D_l)_z \\ ev_{ij} &= (\{D_i\}, D_j)_a \end{aligned} \tag{4}$$

The last concept we take from [2] is the total space, designed to relate documents to their evolution steps[1]. By definition, the total space $\mathscr{E}$ is a *bundle* obtained by relating each document $D_i \in \mathscr{B}_\tau(\{D\})$ to its evolution $ev(\{D_j, \ldots, D_{j_k}\}, D_i)$, with $\{D_j, \ldots, D_{j_k}\} \in \mathscr{B}_t$ and $D_i \in \mathscr{B}_\tau$:

$$\mathscr{E} = \mathscr{B} \times \mathscr{H}. \tag{5}$$

It turns out to be useful to write the bundles explicitly for the evolution shown in Fig. 1. The elements of the bundles are the following pairs b_t[2]:

$$\begin{aligned} b_k &= (D_k, ev_{ik}) = (D_i, (\{D_i\}, D_k)_x) \\ b_l &= (D_l, ev_{kl}) = (D_l, (\{D_k\}, D_l)_z) \\ b_j &= (D_j, ev_{ij}) = (D_i, (\{D_i\}, D_j)_a) \end{aligned} \tag{6}$$

[1] A concatenation of steps, as it results from applying bundles in total spaces defines an evolution path.

[2] Again, t stands for target.

We note that we can reach D_l through the intermediate document D_K, $D_i \xrightarrow{x} D_k \xrightarrow{z} D_l$. We thus add the last bundle b'_l to formalize this two-step evolution:

$$\begin{aligned} b'_l &= (D_l, ev_{kl}) = (D_l, (\{D_k\}, D_l)_z) \\ &= \left(\left(D_l, \underbrace{(\overbrace{(D_i, (\{D_i\}, D_k)_x)}^{b_k}, D_l)}_{b_l})_z \right) \right) \end{aligned} \tag{7}$$

III. Issues and consideration on the model, part I

We propose two categories in this paper. The first category "uses" *base-spaces* and philological operations, see Sec. VI. The second category "uses" documents and NLP tools as morphisms. cf. Sec. VII.

To set the stage for our upcoming discussion on the success and failures of these categories, we present some issues and additional considerations related to the model described in [2].

A. Evolution of documents is the evolution of base-spaces

We defined the *base-spaces* as a collection of documents $\mathscr{B}_\tau$ that are synchronously available at a specific time, τ. We also defined the operations as actions on a subset S_τ of $\mathscr{B}_\tau$. The first point to note is that S_τ constitutes a proper base-space, since it is a collection of documents accessed synchronously. The second aspect to consider is that equation (3) can be understood as follows: operations transform one base-space into another.

We rewrite the equation (3) here by making this aspect explicit.

$$\begin{aligned} op(\underbrace{\{D_1, \ldots, D_n\}}_{S_t \subseteq \mathscr{B}_t(\{D\})}) &= \underbrace{\{D_j\}}_{\{D_j\} \equiv S_\tau \subset \mathscr{B}_\tau(\{D\})} \quad \text{or} \\ op(S_t \subseteq \mathscr{B}_t(\{D\})) &= \{D_j\} \text{ where} \\ \{D_j\} &\equiv S_\tau \subset \mathscr{B}_\tau(\{D\} \end{aligned} \tag{8}$$

The interpretation of (8) is as follows: the set $\mathscr{B}_t$ contains m documents, and the operation op takes a subset S_t containing $n \leq m$ documents to produce a new single document D_j. D_j belongs to another set $\mathscr{B}_\tau$ containing at least D_j which was not on $\mathscr{B}_t$. We view D_j as a subset S_τ of $\mathscr{B}_\tau$ with one document:

$$D_j \in S_\tau \equiv \{D_j\}$$

In the evolutionary perspective presented in [2], the base-space $\mathscr{B}_\tau$ originates from the base-space $\mathscr{B}_t$ and may evolve into a base-space $\mathscr{B}_{t'}$ with $t < \tau < t'$ and:

$$\mathscr{B}_t \neq \mathscr{B}_\tau \neq \mathscr{B}_{t'} \tag{9}$$

Let us introduce a notation that will be useful later in this article: a pair p_t to identify the base-space $\mathscr{B}$ along with its time label t:

$$\mathscr{B}_t \to \{t, \mathscr{B}\} := p_t. \tag{10}$$

Figure 2 shows the same evolution graph of Fig. 1 with the pairs p as nodes:

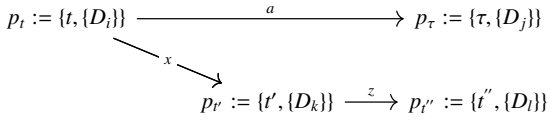

Figure 2: An *Evolution* graph from p_t to p_τ and from p_t to $p_{t''}$ (through $p_{t'}$).

Equation (9) in terms of (10) (and using the time labels of Fig. (2)) reads:

$$\begin{aligned} &p_t \neq p_\tau \\ &p_t \neq p_{t'} \neq p_{t''} \\ &p_\tau \neq p_{t''} \end{aligned} \tag{11}$$

B. Bundles and available operations

This section takes advantage of the fact that we can apply the evolutionary model to pairs $p_t = \{t, \mathscr{B}\}$ and illustrates the correlation that exists between total-spaces (bundles) and available operations within a base-space.

We examine the process depicted in Fig. 3. Our primary goal is to outline the requirements that oversee the feasibility of condensing a two-step process (x, z) into a one-step process, (α).

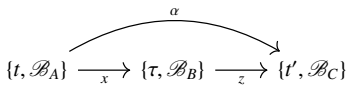

Figure 3: Evolution from $\mathscr{B}_A$ to $\mathscr{B}_C$

To see if it is possible to condense a two-step process into a one-step process, we set the following initial conditions:

$$\mathscr{B}_A = \{D_A\} \tag{12a}$$

$$p_t = \{t, \mathscr{B}_A\} \tag{12b}$$

$$x(\mathscr{B}_A) = x(\{D_A\}) = D_B \tag{12c}$$

$$\alpha(\mathscr{B}_A) = \alpha(\{D_A\}) = D_C^\alpha \tag{12d}$$

Then we apply (12c) to (12b) to obtain the base-space $\mathscr{B}_B$, its bundle b_B (using definition (5), equations (6), and (7))) and the corresponding pair, p_τ:

$$b_B = \{D_B, (\{D_A\}, D_B)_x\} \tag{13a}$$

$$\mathscr{B}_B = \{D_B, D_A\} \tag{13b}$$

$$p_\tau = \{\tau, \mathscr{B}_B\} \tag{13c}$$

Similarly, we apply (12d) to (12b):

$$b_C^\alpha = \{D_C^\alpha, (\{D_A\}, D_C^\alpha)_\alpha\}. \tag{14a}$$

$$\mathscr{B}_C^\alpha = \{D_C^\alpha, D_A\} \tag{14b}$$

$$p_{t^\alpha} = \{t^\alpha, \mathscr{B}_C^\alpha\}, \tag{14c}$$

The base-spaces $\mathscr{B}_B$ and $\mathscr{B}_C^\alpha$ contain $\{D_B, D_A\}$ and $\{D_C^\alpha, D_A\}$ respectively, since the evolution $x : D_A \to B_B$ and $\alpha : D_A \to B_C^\alpha$ are certain.
Figure 4 is a graphical representation of (13) and (14):

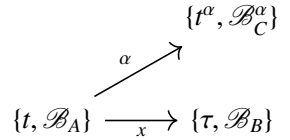

Figure 4: A graphical representation of (13) and (14).

In the figure above τ and t^α may differ, while $\mathscr{B}_B, \mathscr{B}_C^\alpha$ certainly do.

Finally, we have to figure out when going from $\mathscr{B}_B$ to $\mathscr{B}_C^\alpha$ is possible, even at a different time $t' > \tau$, cf. a) in Fig. 5 or when $\mathscr{B}_B$ evolves to a different $\mathscr{B}_C$, as in b).

Figure 5: Evolution graphs of p_τ. a) The operation "?" is α shifted in time, identified by α'; b) "?" is a placeholder for different operations whose number depends on the number of documents in the base space, so that $\mathscr{B}_C \neq \mathscr{B}_C^\alpha$. In (15b) and (15c) we use y and k.

The base-space $\mathscr{B}$ contains two documents, $\{D_B, D_A\}$, thus we have three distinct subsets where we can apply operations:[3]

$$S = \{D_B, D_A\}, S_A = \{D_A\}, \text{ and } S_B = \{D_B\}$$

The corresponding operations are the following:

$$\alpha'(S_A) = D_C^{\alpha'} \tag{15a}$$

$$y(S_B) = D_C^y \tag{15b}$$

$$k(S) = D_C^k \tag{15c}$$

Operations k and y in (15c) and (15b) act on S and D_B, respectively. They produce the documents D_C^k, D_C^y which are different from each other and distinguishable from D_C^α.
On the contrary, both operations (12d) and (15a) take $\{D_A\}$ as input. We may assume that α' is α only applied at a different time, τ instead of t, so that $D_C^\alpha = D_C^{\alpha'}$.
We move on to define the base-spaces that follow from (15):

$$\mathscr{B}_C^{\alpha'} = \{D_C^{\alpha'}, D_A\} = \mathscr{B}_C^\alpha = \{D_C^\alpha, D_A\} \tag{16a}$$

[3]This excludes the empty set $S_\emptyset = \emptyset$.

$$\mathscr{B}_C^k = \{D_C^k, D_B, D_A\} \neq \mathscr{B}_C^\alpha = \{D_C^\alpha, D_A\} \tag{16b}$$

$$\mathscr{B}_C^y = \{D_C^y, D_B\} \neq \mathscr{B}_C^\alpha = \{D_C^\alpha, D_A\} \tag{16c}$$

As a result, if we apply y we follow the evolution graph reported in Fig. 5 b), while if we use α' we follow the graph in a). The application of k is subtle: Depending on how k handles input documents, k can produce either D_C^α or another document, D_C^k. However, the base-space created by k differs $from \mathscr{B}_C^\alpha$, again the path b).

C. Time labeling of base-spaces

The previous section focused on the evolution of base-spaces, the second element of the pair $p_t = \{t, \mathscr{B}\}$, while this section concentrates on the time, the first element.
Whether philological or NLP, operations on documents take a certain amount of time (years for the former, minutes for the latter) to achieve the expected result. If we call the duration of operation op Δ_{op}, the time labeling of the spaces is a s shown in Fig. 6:

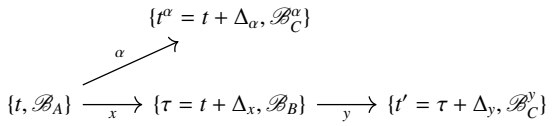

Figure 6: Time labeling of base-spaces evolving under op : $t_f = t_i + \Delta_{op}$

IV. Category Theory

Category Theory, [4], [6]–[9], is a branch of pure mathematics that provides a framework towards an abstraction of mathematical concepts and structures in various mathematical disciplines. The essential notion of *category* includes a collection of *objects* and a collection of *arrows*, which we call *morphism*, to model the *transitions* among these objects.

A. Definition of a Category

A Category C consists of the following entities:

***Objects*:** A collection of *objects*, $Ob(C)$;
***Morphism*:** A collection of morphisms;
Source and Target objects For every morphism f, two objects $X = s(f)$, and $Y = t(f)$ which we refer to as source (or domain) and target or (codomain);
***Composition*:** For every pair of morphisms f, g such that the target of f is the source of g, $t(f) = s(g)$, a binary operation, called *composition* which we identity as $g \circ f$, see (17). In other words, let f and g be morphisms: $A \xrightarrow{f} B$, $B \xrightarrow{g} C$, with $A = s(f), B = t(f) = s(g)$, and $C = t(g)$, then there must exist $h = g \circ f$ from A to C which is the composite of f and g such that $A = s(h), C = t(h)$. Notably, there can be a morphism $k : A \xrightarrow{k} C, k \neq h$.

$$h = f \circ g \tag{17}$$

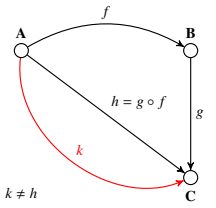

Figure 7: Composition in Category Theory.

The composition of morphisms comes with two additional axioms:

***Associativity*:** Composition is associative. For all f, g, h in $Hom_C(X, Y)$ results, see Fig. 8:

$$h \circ (g \circ f) = (h \circ g) \circ f = h \circ g \circ f, \tag{18}$$

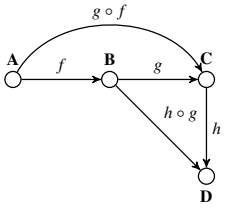

Figure 8: Composition is associative.

***Identity*:** There must exist exactly one[4] *identity arrow* which starts and ends on the same object. The *identity* morphism, $X \xrightarrow{id_X} X$, can be composed, see (19) and Fig. 9:

$$f \circ id_A = f = id_B \circ f, \tag{19}$$

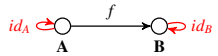

Figure 9: Identity and Composition.

V. Issues and consideration on the model, part II

After we have presented Category Theory, let us proceed to interpret the concept of associativity and identity morphism within the proposed model.

A. Associativity of base-spaces and categories with "families of morphisms"

In Category Theory, the composition of morphisms is associative by definition. In Fig. 10, we schematize a three-step process useful to model associativity.

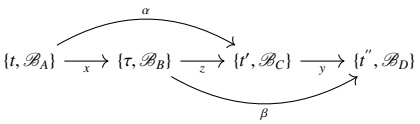

Figure 10: Associativity from $\mathscr{B}_A$ to $\mathscr{B}_D$

[4]If Id_X can not be defined, we speak of semi-category.

Unfortunately, we know from Sec. III (more precisely from (16)) that, apart from time-shifting, $\alpha \neq z \circ x$ and $\beta \neq y \circ z$. However, even if we assume true the following equations:

$$\alpha = z \circ x \tag{20a}$$

$$\beta = y \circ z, \tag{20b}$$

we have to check if the following statement

$$y \circ \alpha = \beta \circ x \tag{21}$$

is verified.
From (20a), z is α shifted in time:

$$\begin{aligned} &x : \{D_A\} \to D_B \\ &\alpha : \{D_A\} \to D_C \text{ if} \\ &z : \{D_A\} \to D_C \text{ and, as a result:} \\ &\mathscr{B}_C = \{D_C, D_A\} \end{aligned} \tag{22}$$

Equation (20b) adds more restrictions on the available operations:

$$\begin{aligned} z : \{\mathscr{B}_B\} &\to D_C \text{ but, using (22)} \\ z : \{D_A\} &\to D_C. \text{ Using the } y \text{ defined in (20b)} \\ y : \{D_C, D_A\} &\to D_D \text{ we create the base-space} \\ \mathscr{B}_D &= \{D_D, D_C, D_A\}. \text{ In addition} \\ \beta : \{\mathscr{B}_B\} &\to D_D \text{ instantiates (at most)} \\ \mathscr{B}'_D &= \{D_D, D_B, D_A\} \end{aligned} \tag{23}$$

If (20b) has to hold, the base-space $\mathscr{B}'_D$ must be the same as $\mathscr{B}_D$. This equality requires both y and β take only D_A in input. Since y acts on D_A, the left-hand side (LHS) of (21), combined with (22) defines:

$$\mathscr{B}_D = \{D_D, D_A\} \tag{24}$$

Similarly, Since y acts on D_A, the right-hand side (RHS) of (21), combined with (23) instantiates:

$$\mathscr{B}_D = \{D_D, D_A\} \tag{25}$$

Associativity requires the following set of operations, $\{\mathbb{O}\}$:

$$\{\mathbb{O}\} = \begin{cases} x : \{D_A\} \to D_B \text{ from } \mathscr{B}_A, \\ z : \{D_A\} \to D_C \text{ from } \mathscr{B}_B, z = \alpha \\ \alpha : \{D_A\} \to D_C \text{ from } \mathscr{B}_A, \\ y : \{D_A\} \to D_D \text{ from } \mathscr{B}_C, y = \beta \\ \beta : \{D_A\} \to D_D \text{ from } \mathscr{B}_B. \end{cases} \tag{26}$$

As a final remark, we add that we may define categories with "families of morphisms": We substitute the definition of *Morphisms* in Sec. IV-A with the following:

Morphisms For every pair of objects $X, Y \in Ob(C)$, a collection (even empty) of *morphisms* (*arrows*), between *objects*, Hom_C.

Utilizing this definition, it is no longer necessary for z and y to be the same as α and β respectively. In theory, it might be possible to find $z' \neq \alpha$ and $y' \neq \beta$ which verify

$$\alpha = z' \circ x$$

$$\beta = y' \circ z,$$

so that both composition and associativity make sense.

B. Identity and semi-category

In this section, we focus on identity morphism Id_x and how its definition according to Category Theory disagrees with the proposed model.

We rewrite (3) using the identity morphism Id_x as the operation:

$$Id(\underbrace{\{D_i\}}_{S_t \subseteq \mathscr{B}_t(\{D\})}) = \underbrace{D_i}_{\{D_i\} \equiv S_\tau \subset \mathscr{B}_\tau(\{D\})}, \tag{27}$$

or according to (8):

$$Id\{D_i \subseteq \mathscr{B}_t(\{D\})\} = D_i. \tag{28}$$

In (28), D_i (on the LHS) belongs to the group of documents at a given time t, which we labeled $\mathscr{B}_t(\{D\})$. The identity Id, by definition, does nothing on D_i, and after its application, the *same* D_i (on the RHS) belongs to the collection of documents at the time $t = \tau > t$, which we tag $\mathscr{B}_\tau(\{D\})$.

Following the formalism of pairs $\{t, \mathscr{B}\}$, introduced in Sec. III:

$$Id : \underbrace{\{t, \mathscr{B}\}}_{p_t} \to \underbrace{\{\tau, \mathscr{B}\}}_{p_\tau}, \text{ and } p_t \neq p_\tau. \tag{29}$$

p_t differs from p_τ both for two reasons:

Time labeling: If we assume Id lasts for a time Δt_{Id} then $\tau = t + \Delta t_{Id}$;

Collection of documents: $\mathscr{B}_\tau = \{D_i\}$ only, while D_i is an element of $\mathscr{B}_t$ which may, in principle, contain many other documents.

The result is that Id in (29) is *not* a valid identity morphism. However, there is a different notion of category, where identity morphism is not required. More precisely, a *semicategory* is a category where the identity morphism is not defined. Consequently, (29) does not hold. In *semicategory*, the identity morphism is not required but may exist. In this case, the identity morphism in a *semicategory* is an additional property, not an extra structure.

VI. First proposed category

In the evolutionary perspective presented in [2], the base-space $\mathscr{B}_\tau$ originates from the base-space $\mathscr{B}_t$ at time $t < \tau$ and evolves into a base-space $\mathscr{B}_{t'}$ at time $t' > \tau > t$. According to the same paper, subsets of documents $S_\tau \subseteq \mathscr{B}\tau$ transform into a single document $D_j \in \mathscr{B}_{t'}$ through appropriate operations. This new $\mathscr{B}_{t'}$ includes at least one document D_j that does not belong to $\mathscr{B}_\tau$. Consequently, in the evolution path, we have $\mathscr{B}_t \neq \mathscr{B}_\tau \neq \mathscr{B}_{t'}$.

Even if in Fig. 1, the focus is on the evolution of D_i, we know that the real evolving objects are the base-spaces (better the pairs p), as explained in Sec. III-A. The structure of the base-space, which is a collection of documents, remains constant while the number of contained documents changes.

To conclude, the most intuitive approach is to associate the entities within the category with the base-spaces and the operations between base-spaces with the morphisms. The proposed category has "families of morphisms" instead of a single morphism, which results more pertinent in the domain of Philology.

According to Sec. IV, we define a category C with the following characteristics:

***Objects*:** The collection of *objects* is represented by $Ob(C)$ and includes all base-spaces $\mathscr{B}$ that evolve, such as $\mathscr{B}_i$, $\mathscr{B}_j$, $\mathscr{B}_k$, $\mathscr{B}_l$, and so on, as the outcome of applied of operations.

***Morphisms*:** The set of all operations that make $\mathscr{B}_i$ evolve into $\mathscr{B}_j$, as shown in Fig. 11, is denoted as $Hom_C(\mathscr{B}_i, \mathscr{B}_j)$.

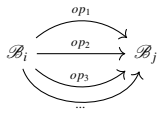

Figure 11: Morphisms from $\mathscr{B}_i$ to $\mathscr{B}_j$.

Following the discussion of Sec. III, and according to (10) we replace Fig. 11 with Fig.12:

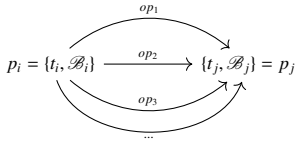

Figure 12: Morphisms from p_i to p_j.

Performing the operation described in (3) (or the analogous (8)) will increase the time labeling t of the pairs (see Sec. III-C) and send $\mathscr{B}_i$ to $\mathscr{B}_j$ (p_i in p_j), where $\mathscr{B}_i$ and $\mathscr{B}_j$ are objects in C.

We rewrite here the definition of the identity morphism (29)

$$Id_p : \{t, \mathscr{B}\} \to \{t + \Delta t_{Id}, \mathscr{B}\} = \{t_1, \mathscr{B}\} = p' \neq p \tag{30}$$

Since the identity morphism is not defined, the proposed category is a *semicategory*, which is not a problem.

For base-spaces containing only one document, the identity morphism can be defined as an extra property:

$$Id_p : \{t, \mathscr{B}\} \to \{t, \mathscr{B}\} = p' = p, \tag{31}$$

where t is left unchanged.

For such types of base-spaces, even (19) holds, as explained below.

If $\mathscr{B} = \{D\}$ and $f : D \to D'$ is a morphism, then the application of f on $\mathscr{B}$ results in:

$$f : \underbrace{\{t, \mathscr{B}\}}_{p} \to \underbrace{\{t + \Delta t_f, \mathscr{B}'\}}_{p'},$$

where $\mathscr{B}' = \{D'\}$. Composing the identity morphism with f on p' is equivalent of composing f with p:

$$Id_{p'} \circ f : \{t, \mathscr{B}\} \xrightarrow{f} \{t + \Delta t_f, \mathscr{B}'\} \xrightarrow{Id_{p'}} \{t + \Delta t_f, \mathscr{B}'\}$$

is equal to

$$f \circ Id_p : \{t, \mathscr{B}\} \xrightarrow{Id_p} \{t, \mathscr{B}\} \xrightarrow{f} \{t + \Delta t_f, \mathscr{B}'\}$$

The identity morphism is defined and can be combined as stated in Sec. IV-A.

In addition, when we come to model composition, see (17), to associativity, as shown in Figures 8 and 10(with base-spaces A, B, and C), we find:

$$\begin{aligned} &f : \{t, A\} \to \{t + \Delta t_f, B\} \\ &g : \{t + \Delta t_f, B\} \to \{t + \Delta t_f + \Delta t_g, C\}, \text{ but} \\ &h : \{t, A\} \to \{t + \Delta t_h, C\}, \end{aligned} \tag{32}$$

since $h = g \circ f$ is a single operation, it increases the time t by Δt_h.

Rules of composition are an integral part of the category definition and we prescribe:

$$\Delta t_h = \Delta t_g + \Delta t_f$$

With this prescription, the two objects C are identical: same time label and same contained documents.

While a formally sound category has been established, the documents go through different evolutionary paths, as described in Sec. III-B. This difference can have an impact on the philological analysis of the text.

Below we use the formalism defined in Sections III and V to analyze the following example: D_A is a Greek poem, D_B its Latin translation, and D_C is the paraphrase of D_B. Formally:

$$f : \underbrace{\{D_A\}}_{D_A \in \mathscr{B}_A} \to D_B, \text{ with } \mathscr{B}_B = \{D_B, D_A\} \tag{33a}$$

$$g : \underbrace{\{D_A\}}_{D_A \in \mathscr{B}_B} \to D_C, \text{ with } \mathscr{B}_C = \{D_C, D_A\} \tag{33b}$$

$$h = g \circ f : \underbrace{\{D_A\}}_{D_A \in \mathscr{B}_A} \to D_C, \text{ with } \mathscr{B}_C = \{D_C, D_A\} \tag{33c}$$

Analogously for pairs:

$$p_A = \{t_A, D_A \in \mathscr{B}_A\} \xrightarrow{f} \{t_A, \mathscr{B}_B = \{D_B, D_A\}\} = p_B \tag{34a}$$

$$p_B = \{t_B, D_A \in \mathscr{B}_B\} \xrightarrow{g} \{t_C, \mathscr{B}_C = \{D_C, D_A\}\} = p_C \tag{34b}$$

$$p_A = \{t_B, D_A \in \mathscr{B}_A\} \xrightarrow{g \circ f} \{t_C, \mathscr{B}_C = \{D_C, D_A\}\} = p_C \tag{34c}$$

Finally, we apply equations (5), (4), and (6) to find out that the albeit formally correct paths previously (f, g, and $g \circ f$) have implications from the point of view of philological analysis.

$$\begin{aligned} E_{t_B} &= \{D_B\} \times (\{D_B\}, D_A)_f \\ E_{t_C} &= \{D_C\} \times (\{D_C\}, D_B)_g \\ &= \{D_C\} \times (\{D_C\}, \underbrace{\{D_B\} \times (\{D_B\}, D_A)_f)_g}_{E_{t_B}} \end{aligned} \tag{35a}$$

$$E'_{t_C} = \{D_C\} \times (\{D_C\}, D_A)_{g \circ f} \tag{35b}$$

In the equation (35), the contrasts between E and E' lie in the internal relationships among the documents. At time t_C, scholars access different sources depending on the evolution paths from $A \to C$. In E', only D_A is connected to D_C, while in E, D_C is connected to D_B, and D_B to D_A. It is very different for a scholar at t_C to create the paraphrase D_C, having access

to the original Greek poem (D_A) and its Latin translation (D_B) than to have access to only the Greek poem D_A.

After analyzing this simple example, we concluded that although the proposed *semicategory* is formally correct, it does not consider all the features associated with philological research, including one of the most significant, namely access to sources.

VII. A different approach

The considerations of sections III and V made on the model and the first proposed category suggest two clear types of modifications to the model.

First, we focus exclusively on the documents, disregarding the time labeling of the base-spaces. In this case, we have a single base-space that contains all documents. This approach provides the single base-space $\mathscr{B}$ an algebraic structure beyond just being a collection of documents. The set $\mathscr{B}$ is an algebraic structure with a trivial signature, meaning no relations or operations are formally defined. However, it is necessary to determine how documents are assembled and accessed by the operations that determine their evolution. In other words, we must specify both the *arity* of the operations and, if possible, the relationships between the elements of the set.

Secondly, we add some restrictions to the model, as discussed in Sec. VII-A

A. Restrictions to the model

The main restrictions to the model are the following:

(a) NLP tools as automatic agents responsible for document evolution. NLP tools operate on a time scale shorter than that required by philological operations. This aspect allows us to argue that input and output documents belong to the same base-space;

(b) Unary operations:

$$D_i(c_i, f_i, \{p\}_i) \xrightarrow{op} D_j(c_j, f_j, \{p\}_j). \tag{36}$$

Therefore, the subset $S \subseteq \mathscr{B}$ to which op applies, see (3) consists of a single document;

(c) We also assume that the format f in (36) remains unmodified during the process:

$$D_i(c_i, \mathbf{f}, \{p\}_i) \xrightarrow{op} D_j(c_j, \mathbf{f}, \{p\}_j)$$

This last assumption pertains to the notions of interoperability and compatibility, see Sec.VII-B.

B. Tools Compatibility and Interoperability

Interoperability is commonly used in various fields, including healthcare, [10], emergency management, citizen services, and computer science, [11]. When discussing interoperability in computer systems two terms frequently used are: *syntactic* and *semantic* interoperability. *Syntactic* interoperability refers to the agreement on data formats and communication protocols, essentially the structural basis of interoperability.
Examples of agreed-upon data structures include XML, SQL dumps, and JSON. Semantic interoperability deals with agreeing on the meaning of exchanged data using available standards. This is of great importance for Language Resource and Technologies (LRT) and several efforts have been made to map different LRTs, [12]–[14].

In this paper, the term compatibility, specifically tool compatibility, will be used. Compatibility refers to the degree to which tools share input/output features, such as file requirements and formats, MIME types, and linguistic information. The higher the number of shared characteristics, the greater the level of compatibility. Compatibility and interoperability are closely related. Indeed, when the output features of a tool fit the input characteristics of the next one, we assume these two tools are interoperable and can be *pipelined*.
We focus on *unary* operations as in (36) that leave the format f untouched. This approach ensures that both the read and produced documents have the same structure (f), ensuring syntactic interoperability. We take semantic interoperability for granted.

Composition and associativity are the foundation of the concept of interoperability. Consider we have a system with 3 operations, op_0, op_1, op_2, which act on an initial document, D_0, to produce the final D_3. Operation op_i takes input document D_i and outputs D_{i+1}. The process from D_0 to D_3 is formalized as $D_0 \xrightarrow{op_0} D_1$, $D'_1 \xrightarrow{op_1} D_2$, and $D'_2 \xrightarrow{op_2} D_3$. A prime sign ($'$) suggests the unlikelihood of one operation's output being the exact input of another. However, if op_0 and op_1 are compatible, meaning that D_1 and D'_1 have similar features such as format, exchanged data, and their meaning, they can be pipelined.
This involves creating a new operation, op_3, by composing op_1 with op_0, resulting in $D_0 \xrightarrow{op_3} D_2$. Similarly, op_2 can be composed with op_1 to create op_4, resulting in $D_1 \xrightarrow{op_4} D_3$. There are five interoperable operations, namely $op_0, op_1, op_2, op_3, op_4$, that can be associated in different ways such as $D_0 \xrightarrow{op_2 \text{ after } op_3} D_3$ or $D_0 \xrightarrow{op_4 \text{ after } op_0} D_3$.

VIII. Second proposed category

In this section, we show that the notions of composition and associativity in Category Theory map onto those of interoperability. Section VII-B explains how two interoperable tools (op_1, op_2) can be combined to create a more complex, *composite*, tool ($op_2 \circ op_1$) that produces identical results, as depicted in Fig. 13.

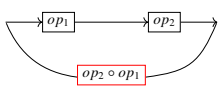

Figure 13: Composition of linguistic tools. Cf. (17)

When we deal with multiple tools, the processing pipeline(s) may take different paths. The final result can be achieved by using either individual or composite tools[5] as in Fig. 14 and 15.

[5]By atomic tools, we mean tools that go directly from A to B: $A \to B$; by composite, tools that need an intermediate C to go from A to B: $A \to C \to B$

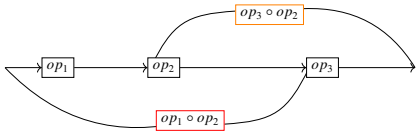

Figure 14: Associativity.

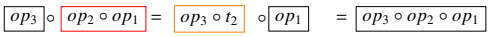

Figure 15: Associativity as an equation. Cf. (18).

Following Sec. IV, we define the category C as follows:

***Objects*:** The collection of *objects*, $Ob(C)$, comprises all documents (D_i,D_j,D_k,D_l ...) that can be processed by a linguistic application;

***Morphisms*:** The set of any linguistic application which consumes D_i and produces D_j as shown in Fig. 16 is denoted as $Hom_C(D_i, D_j)$.

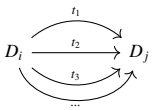

Figure 16: Tools from D_i to D_j.

The identity morphism is defined as in (37):

$$Id_X : X \to X \tag{37}$$

Id_X is a dummy tool that consumes and returns the same document. There must be an identity morphism for every document D_i, see Fig. 17.

Figure 17: Identities.

Equation (19) also holds. If f is a morphism,

$$f : D \to D'$$

then

$$Id_{D'} \circ f : D \xrightarrow{f} D' \xrightarrow{Id_{D'}} D'$$

is equal to

$$f \circ Id_D : D \xrightarrow{Id_D} D \xrightarrow{f} D'$$

The identity morphism is properly defined and can be combined, see Sec. IV-A. If we apply (18) to Fig. 7, with A, B, C as documents, we have

$$\begin{aligned} &f : A \to B \\ &g : B \to C, \text{ and} \\ &h : A \to C. \end{aligned} \tag{38}$$

Since $h = g \circ f$ is obtained by combining two interoperable tools, the final object C is the same. From a linguistic perspective, it is rational to model operations around documents and to disregard the temporal aspect. NLP tools operate on a much shorter timescale than the philological operations: minutes opposed to years. As a result, we can ignore the time aspect in the base-spaces and the various paths connecting the documents. By doing so, we can create a unique base-space and the corresponding evolution spaces:

$$\mathscr{B} = \{D_A, D_B, D_C\} \tag{39a}$$

$$\mathscr{H} = \{(\{D_C\}, D_B)_g, (\{D_B\}, D_A)_f, (\{D_C\}, D_A)_{g \circ f}\}. \tag{39b}$$

The total space $\mathcal{E}$ includes the documents in $\mathscr{B}$ along with their evolution paths. At $t = t_C$, scholars have access to all sources.

Once interoperability is established for the tools able to operate on a set of documents $\{D_i, D_j, D_k, D_m\} \in \mathscr{B}$, D_i is processed to obtain D_m in several ways: using t_5 (the dashed line); either composing t_4 with t_3 (which is $t_2 \circ t_1$) or t_6 and t_1 and doing the same with t_4, t_2, and t_1. Fig. 18 shows different paths from D_i to D_m.

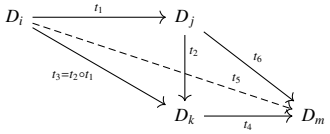

Figure 18: Composition and associativity.

We write here the total space for D_m, E_m:

$$E_m = \Big(\{D_m\}, (\{D_m\}, D_i)_{t_5}, (\{D_m\}, D_j)_{t_6}, (\{D_m\}, D_k)_{t_4}\Big)$$

IX. Final remarks and open research problems

This section analyzes the model restrictions presented in Section VII-A. The different timescale of NLP tools - prescription (a)- justifies the category proposed in Sec. VIII rather than that in Sec. VI. It is because the approach in Sec. VIII treats documents (both input and output) as part of the same base-space. This perspective is valuable from an evolutionary perspective since it makes all evolutionary paths accessible to scholars.

The morphisms described in the second approach are unary operations - prescription (b). These morphisms endow $\mathscr{B}$ with an algebraic structure having a nontrivial signature: documents which are directly connected by these operations and, together with prescription (a), documents which are connected with relations: for example, D_1 will be in relation with D_2 only if both D_1 and D_2 are nodes in an evolution graph.

It is important to note that assuming NLP tools take only one document in input is extremely restrictive. For example, the modeling of a piece of software that reads a list of places (D_1) to identify them on a text (D_2) and produces an annotated text (D_3) is a binary operation:

$$op(\{D_1, D_2\}) \to D_3$$

In addition, [2] postulated two binary operations: the union $\cup$ and the intersection $\cap$ of two documents. The operations $\cup$ and $\cap$, (along with the empty document D_e also postulated in [2]) provide robustness to the model but make considering unary operations insufficient. Consequently, the algebraic structure

with which $\mathscr{B}$ is equipped must include at least binary operations between documents:

$$* : D \times D \to D'$$

Prescription (c) is the most restrictive because it requires syntactic interoperability among different tools. In NLP this is quite unlikely, see [15]–[17].
The chain $D_0 \to D_3$ illustrated in Sec. VII-B

$$D_0 \to D_3 = D_0 \xrightarrow{op_0} D_1 \xrightarrow{op_1} D_2 \xrightarrow{op_2} D_3,$$

should be rewritten like this:

$$\begin{aligned} &D_0(f_0) \to D_3(f_3) = \\ &D_0(f_0) \xrightarrow{op_0} D_1(f_0) \xrightarrow{conv_{f_0-f_1}} D_1(f_1) \\ &D_1(f_1) \xrightarrow{op_1} D_2(f_1) \xrightarrow{conv_{f_1-f_2}} D_2(f_2) \\ &D_2(f_2) \xrightarrow{op_2} D_3(f_2), \end{aligned} \tag{40}$$

where f_i is the format of D_i and $conf_{f_i-f_j}$ converts f_i into f_j so that op_j can read $D_j(f_j)$.
Format converters operate on documents. Following Sec. VIII format converters are proper NLP tools. This fact has two consequences: (i) $Hom_C(D_i, D_j)$ must be expanded to include format converters; (ii) we have to prescribe composition and associativity on any possible combination of tools and format converters.

A more elegant solution is to treat the format change as a *functor* from a category A whose documents have the format f_A to B whose documents have the format f_B. The functor should be prescribed on the objects of the category, on morphisms (including identity), and their composition.

X. Conclusion and further reading

Interoperability among linguistic tools is a relevant issue, although whether this is a scientific or an engineering problem is still open.
Software integration platforms such as GATE[6] [18], UIMA[7] [19], [20], and Research Infrastructures, such as CLARIN[8], DARIAH[9], and H2IOSC[10] use syntactic and semantic interoperability for the (linguistic) services they provide to users.
Interoperability is also important in computational linguistics. In [**?**], the authors describe how to design Language Resources and NLP tools as web services to address the needs of digital humanists. The authors used interoperability to connect lexicons, semantic resources, and fine-grained text management. Category Theory and Linguistics are closely related. For example, [21], [22] use Category Theory and *Pregroups* to model grammar and interactions among words. [23], [24] define and update *DisCoCat*, a model that provides compositional semantics for studying the meaning of sentences in natural languages.

In this paper, we have used interoperability and Category Theory to augment the evolutionary model presented in [2] by endowing base-spaces with an algebraic structure consisting of rules about operations on documents and relations between them. It was made possible by defining a category with documents as objects and NLP tools as morphisms. Interoperability between language tools guarantees that the axioms of composition and associativity are satisfied.

A second important aspect shown in the paper concerns the short time scale of executing NLP tools compared to the philological time scale. It was possible to affirm that the base-space does not change the time labeling; the same base-space is extended by including input and output documents. Consequently, it is possible to keep track of all the evolutionary histories of documents belonging to a base-space, effectively solving the philological problem that made the *semicategory* with base-spaces as objects and evolution as morphisms unsatisfactory.

[6]https://gate.ac.uk/
[7]https://uima.apache.org/
[8]https://www.clarin.eu/
[9]https://www.dariah.eu/
[10]https://www.h2iosc.cnr.it/

References

[1] R. Del Gratta, F. Boschetti, L. Bambaci, and F. Sarnari, "Approaching document analysis with a formal model," in *6th International IEEE Colloquium on Information Science and Technology*, (Agadir, Morocco), pp. 208–214, 2020.
[2] R. Del Gratta, F. Boschetti, L. Bambaci, and F. Sarnari, "Document analysis and Textual philology: A Formal Perspective," *International Journal of Information Science and Technology*, vol. 5, no. 1, pp. 5–15, 2021.
[3] R. Del Gratta, F. Boschetti, A. Del Grosso, S. Zenzaro, and L. Bambaci, "Philology as a dynamic system," vol. 6, p. 1–20, Jan. 2022.
[4] S. Awodey, *Category Theory*. New York, NY, USA: Oxford University Press, Inc., 2nd ed., 2010.
[5] R. D. Gratta, "A Category Theory Approach to Interoperability," 2020.
[6] S. Mac Lane, *Categories for the Working Mathematician*. Graduate Texts in Mathematics, Springer, second ed., 1998.
[7] J. Baez and M. Stay, "Physics, Topology, Logic and Computation: A Rosetta Stone," *Lecture Notes in Physics*, vol. 813, pp. 95–172, 2011.
[8] T.-D. Bradley, "What is Applied Category Theory?," 2018.
[9] E. Riehl, *Category Theory in Context*. Aurora: Dover Modern Math Originals, Dover Publications, 2017.
[10] "HIMSS - Interoperability in Healthcare." https://www.himss.org/resources/interoperability-healthcare.
[11] "NIFO - National Interoperability Framework Observatory." https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/3-interoperability-layers.
[12] N. Ide and J. Pustejovsky, "What does Interoperability Mean, Anyway? Toward an Operational Definition of Interoperability," in *Proceedings of the Second International Conference on Global Interoperability for Language Resources ICGL 2010*, (Hong Kong, China), 2010.
[13] N. Ide, J. Pustejovsky, N. Calzolari, and C. Soria, "The SILT and flarenet international collaboration for interoperability," in *Proceedings of the Third Linguistic Annotation Workshop, LAW 2009, August 6-7, 2009, Singapore*, pp. 178–181, 2009.
[14] C. Cieri, K. Choukri, N. Calzolari, D. T. Langendoen, J. Leveling, M. Palmer, N. Ide, and J. Pustejovsky, "A Road Map for Interoperable Language Resource Metadata," in *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*, 2010.
[15] N. Ide and K. Suderman, "Bridging the Gaps: Interoperability for GrAF, GATE, and UIMA," in *Proceedings of the Third Linguistic Annotation Workshop, LAW 2009, August 6-7, 2009, Singapore*, pp. 27–34, The Association for Computer Linguistics, 2009.
[16] R. Del Gratta and D. Albanesi, "OpeNER and PANACEA: Web Services for the CLARIN Research Infrastructure," in *Proceedings of CLARIN Annual Conference 2019, CAC 2019* (K. Simov and M. Eskevich, eds.), (Leipzig, Germany), 2019.
[17] J. Odijk, "Discovering software resources in CLARIN," *Linköping electronic conference proceedings (Print)*, vol. 159, pp. 121–132, 2018.
[18] H. Cunningham, "Software Architecture for Language Engineering," 2000.

[19] D. Ferrucci and A. Lally, "UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment," *Natural Language Engineering*, vol. 10, pp. 327–348, sep 2004.
[20] D. Ferrucci, A. Lally, K. Verspoor, and E. Nyberg, "Unstructured Information Management Architecture (UIMA) Version 1.0." OASIS Standard, mar 2009.
[21] A. Preller and J. Lambek, "Free compact 2-categories," *Mathematical Structures in Computer Science*, vol. 17, no. 2, p. 309–340, 2007.
[22] J. Lambek, *From Word to Sentence: a computational algebraic approach to grammar*. Polimetrica sas, 2008.
[23] B. Coecke, E. Grefenstette, and M. Sadrzadeh, "Lambek vs. Lambek: Functorial vector space semantics and string diagrams for Lambek calculus," *Annals of pure and applied logic*, vol. 164, no. 11, pp. 1079–1100, 2013.
[24] B. Coecke, M. Sadrzadeh, and S. Clark, "Mathematical foundations for a compositional distributional model of meaning," *arXiv preprint arXiv:1003.4394*, 2010.