A formal ontology of texts

Pawel GARBACZ [1]

Department of Philosophy, The John Paul II Catholic University of Lublin

Abstract. The paper outlines a formal ontology of texts. The main ontological distinction I try to capture is the difference between tokens of texts (physical texts) and texts (abstract texts) themselves. The latter are understood here as ontologically dependent on the former in the sense spelled out by the axioms of the ontology. I formally characterise both types by means of their criteria of identity and existence, parthood, spatial occupancy, text precedence, and intentionality.

Keywords. ontology of texts; ontological relations; layered ontology

1. Introduction

Although applied ontologies have been being developed for almost all kinds of domains, ontologies for the humanities stand out as an under-represented minority in this crowd. One can explain this fact by pointing out that there is a number of theoretical issues that are specific to this type of research which have not been properly addressed in ontology (be it philosophical or applied). One of these issues concerns the ontological status of texts. This paper attempts to solve this problem by outlining a certain philosophical view on texts by means of a formal theory. Section 2 below reports some previous research in the vicinity of this issue. Section 3 explains the informal assumptions that support the formal theory being developed in the next two sections. The last section suggests some perspectives of further research.

The key notions of my formal ontology of texts either are "borrowed" from the BFO ontology or remain undefined. In particular, the ontology does not provide a definition that says "x is a text iff $\phi(x)$". Still, the formalism is hoped to cast a definite and relatively unambiguous conceptualisation of texts. It characterises them indirectly, i.e., in the axiomatic way, by establishing the conceptual links between the basic ontological relations relevant for texts. If the resulting conceptual nexus turns out to be rich enough, the ontology may contribute to the theoretical foundations for the future, more extensive, research in knowledge representation for the humanities.

Due to the page limit the paper specifies the axioms of the ontology and, as a rule, only mentions certain theorems that follow from these axioms. All proofs of these theorems were generated by automatic theorem prover PROVER9 via the TPTP front end (`http://www.cs.miami.edu/~tptp/cgi-bin/SystemOnTPTP`). They are are accessible on `http://metaontology.pl/deliverables/papers/ontology-of-texts/`.

[1] Email: garbacz@kul.pl

2. Related Work

There exist at least two major attempts to capture the ontological aspects of texts in formal terms: Information Artifact Ontology and FRBRoo model. More precisely speaking, these ontologies conceptualise texts as one category among other categories of information-laden entities, so they are not ontologies of *just* texts. The lack of space prevents me from discussing (or even presenting) the details of both ontologies. In what follows I will focus only on those main features that are relevant for the purposes of this paper.

2.1. Information Artifact Ontology

Referring to the BFO ontology as its base Information Artifact Ontology (aka: IAO) is to capture the kinds of information entities that are involved in scientific research.

> A foundational idea in the IAO is that information content entities are related to other things be being about them or denoting them. Information content entities are a subtypes of BFO's "generically dependent continuant". The ideas, tables, and figures being communicated right now in this book are examples of information content entities that denote other things such as the OWL and RDF languages, various ontologies, and entities of many other types.(Arp et al., 2015, p. 168).

IAO is available, as of now, as an OWL ontology.[2] The stable version of this ontology, dated on the 23rd of February 2015, contains 180 classes, 52 object properties, and 5 datatype properties – but some of those come from the BFO ontology. Among other things, it contains the class (`IAO_0000300`) that classifies textual entities of various kinds: words, sentences, paragraphs, written parts of publications. `IAO_0000300` is a subclass of the class of information content entities (`IAO_0000030`), so as aforementioned in the above quote textual entities are indeed generically dependent continuants that are "about" some (other) things. It is worth to mention that IAO has it that information content entities are concretized (via `RO_0000059` object property) as qualities (i.e., as information carriers) of material information bearers (`IAO_0000178`). A comment to the `RO_0000059` property explains this pattern by means of an example:

> A journal article is an information artifact that inheres in some number of printed journals. For each copy of the printed journal there is some quality that carries the journal article, such as a pattern of ink. The quality (a specifically dependent continuant) concretizes the journal article (a generically dependent continuant), and both depend on that copy of the printed journal (an independent continuant).

2.2. FRBRoo Model

FRBRoo is a formal ontology to capture the semantics of bibliographic information, which was developed to harmonise the FRBR model with the CIDOC CRM ontology.[3] The ontology consists of 52 classes and 74 relations (i.e., properties in the sense of

[2] It is available on `http://purl.obolibrary.org/obo/iao.owl/`

[3] For the sake of this paper I refer to the version 2.1 of the ontology, which was published in February 2015 – see Bekiari et al. (2015).

CIDOR CRM). The conceptual basis of FRBRoo is the Functional Requirements for Bibliographic Records model (aka: FRBR), which was published in 1998 by International Federation of Library Associations and Institutions – see IFLA (1998). The main refinements of this model, which were put forward by FRBRoo, concern introduction of temporal entities and re-conceptualisation of the so-called group 1 entities from FRBR: work, expression, manifestation, and item. I will focus here on the latter refinement.

FRBR defined works as abstract entities that are recognised as common contents of their expressions. FRBRoo ontology finds this description ambiguous and distinguishes four, more specific and less ambiguous, subclasses thereof: individual works, publication works, complex works, and container works.

FRBR defined expressions (of works) as realisations of works in the form of certain notations. As a result an expression (of a work) is the "specific intellectual or artistic form that a work takes each time it is realized. (IFLA, 1998, p. 16). FRBRoo specializes this notion by means of two more specific kinds: self-contained expressions and expression fragments.

FRBR defined manifestations as physical embodiments of expressions in material of the specific kind. "As an entity, manifestation represents all the physical objects that bear the same characteristics, in respect to both intellectual content and physical form." (IFLA, 1998, p. 21). In FRBRoo this notion is split into two sub-kinds. Manifestations as product types are understood as collections of physical entities and, for that reason, are abstract objects. On the other hand, manifestations as singletons are physical objects themselves.

Consequently, FRBR's notion of item, i.e., a concrete physical object that is an exemplar of manifestation, turned out to be in need of a similar adjustment. Namely, FRBRoo distinguishes between manifestations singletons and items that are (multiple) examples of manifestation product types.

3. Informal assumptions

First, let me emphasize that the ontology presented in this paper concerns traditionally understood texts as entities that are intended to be read in the sequential order as specified by their authors. In particular, the ontology does *not* cover the so-called ergodic texts (in the sense of Aarseth (1997)).

Secondly, the particular approach adopted here is focused on (two kinds of) relations between (two kinds of) texts. Namely, it is taken for granted that there are physical texts (aka: text tokens) and abstract texts (aka: texts). Both exist in time (i.e., are present at certain times), have parts (at these time at which they are present), and may precede one another, i.e., one text token (*resp.* text) may precede another text token (*resp.* text) when they are both parts of some more comprehensive text. The former, but not the latter, occupy regions of space. There exists also a kind of ontological dependency between them. Namely, text tokens may be segregated into sets or collections such that for each set all its elements are equivalent to each other. Then for every text there exists exactly one such set so that the text in question existentially depends on any element from the set.

Both notions: "physical text" and "text" are taken to be primitive here, i.e., they are not defined within the ontology. Still, they are understood in such a way that books, articles, sections, paragraphs, sentences, and words are considered as, respectively, physical texts and texts. The formal theory below is to provide more details on the content, structure, and ontological roles of these notions. In particular, the theory implies that there are other types of texts besides those aforementioned in the last sentence but one, e.g., each concatenation of adjacent texts is a text itself.

One may find my distinction between text tokens and texts similar to the IAO distinction between material information bearers and information content entities – although the latter involves information carrier qualities, which are ignored by the former. On the other hand, the FRBR quadripartite account of bibliographic entities is much more fine grained than these two distinctions.

Both physical texts and texts are intentional entities, where "intentional" is understood in accordance with the Latin etymology as synonymous to "pointing towards" or "referring to'. So both physical texts and texts refer, although it is not assumed that every text refers to something. In other words, the theory below *can* accommodate (separately) each of the following sets of claims:

1. set #1
   (a) "John is rich" is a text that refers to something (say, a situation).
   (b) "John" is a text that refers to an object.
   (c) "is" is a text that does not refer to anything.
2. set #2
   (a) "John is rich" is a text that does not refer to anything.
   (b) "John" is a text that refers to an object.
   (c) "is" is a text that does not refers to anything.
3. set #3
   (a) "John is rich" is a text that refers to something (say, a situation).
   (b) "John" is a text that refers to an object.
   (c) "is" is a text that refers to something (say, the relation on inherence between qualities and objects).

Also the notion of text equivalence, which is critical to the ontological analysis of texts, is left undefined. Intuitively, two texts are equivalent when they are of the same shape (given a certain notation) but the formal account of the latter relation is not provided here. Instead, the theory below attempts to characterise the relation of text equivalence in terms other ontological relations.

In sum, in contrast to the other two ontologies the approach taken in this paper focusses on the ontological relations that relate (physical) texts with one another: temporal existence, identity, parthood, text precedence, intentionality, and text equivalence. I will try to characterise their formal properties in section 5, but first I need to define the language of the ontology.

4. Sorted language of the ontology

In order to reduce the number of primitive terms and to simplify some of the axioms I will render the formal ontology of text in an order-sorted language.[4]

The *alphabet* of the formal language of the ontology of texts is an order-sorted language that contains:

1. five sorts from set $\mathtt{S} = \{S_{region}, S_{object}, S_{non-time}, S_{time}, S_{non-time}, S_{\top}\}$ that are ordered by the following two sort declarations:

   (a) $S_{region}, S_{object} \sqsubseteq S_{non-time}$
   (b) $S_{time}, S_{non-time} \sqsubseteq S_{\top}$

2. the usual logical connectives, including identity[5], and quantifiers
3. sorted individual variables:

   (a) $t, t', \cdots : S_{time}$
   (b) $r, r', \cdots : S_{region}$
   (c) $x, y, z, \cdots : S_{object}$
   (d) $s, s', \cdots : S_{non-time}$

4. eight primitive predicates with the following function declarations – the informal readings of these predicates are given in table 1:

   (a) $\mathrm{PhTxt} :< S_{object} >$
   (b) $\mathrm{Txt} :< S_{object} >$
   (c) $\mathrm{dep_{txt}} :< S_{object}, S_{object} >$
   (d) $\mathrm{pre} :< S_{non-time}, S_{time} >$
   (e) $\mathrm{occ} :< S_{non-time}, S_{region}, S_{time} >$
   (f) $\mathrm{ppt} :< S_{non-time}, S_{non-time}, S_{time} >$
   (g) $\mathrm{prior_{ph}} :< S_{object}, S_{object}, S_{time} >$
   (h) $\mathrm{eq_{ph}} :< S_{object}, S_{object}, S_{time} >$

Let $\mathtt{P}$ be the set of the aforementioned nine primitive predicates. Also let $\mathtt{D}$ contain the two sort declarations and the eight function declarations above. Then $\Sigma =< \mathtt{S}, \emptyset, \mathtt{P}, \mathtt{D} >$ will be the *sorted signature* for the language of our ontology.

All other elements of this order-sorted language (i.e., Σ-words, Σ-substitutions, and Σ formulae) can be defined in the usual way – see (Socher-Ambrosius and Johann, 1997, p. 170-172).

5. Axiomatisation

The formal ontology developed in this paper is constituted by the following axioms, which I arranged in three groups:

1. general axioms and definitions that specify the top-level ontological categories and relations – section 5.1 below;

[4] Speaking about order-sorted languages I refer to the idea from Socher-Ambrosius and Johann (1997) using the notation from Kaneiwa (2004).

[5] Note the sort declaration for identity $=:< S_{\top}, S_{\top} >$.

Formal Predicate	**Informal Reading**
$\mathrm{PhTxt}(x)$	x is a text token
$\mathrm{Txt}(x)$	x is an (abstract) text
$\mathrm{dep_{txt}}(x,y)$	text x depends on text token y
$\mathrm{pre}(x,t)$	x is present at time t
$\mathrm{occ}(s,r,t)$	s occupies region r at time t
$\mathrm{ppt}(x,y,t)$	x is (a proper) part of y at time t
$\mathrm{prior_{ph}}(x,y,t)$	text token x precedes text token y at time t
$\mathrm{eq_{ph}}(x,y,t)$	text token x is equivalent to text token y at time t
$\mathrm{int_{ph}}(x,y,t)$	text token x represents y at time t

Table 1. Primitive predicates of the formal ontology of texts

2. axioms and definitions that characterise physical texts – section 5.2 below;
3. axioms and definitions that characterise abstract texts – section 5.3 below.

5.1. Top-level Ontology

5.1.1. Existence and Time

The first axiom implies that the ontology of texts concerns only actual entities, i.e., those that either existed, exist, or will exist. This excludes all merely possible entities, e.g., those that exist only in counterfactual situations.

$$\exists s\ \mathrm{pre}(s,t) \tag{1}$$

The next two axioms are to express that being a physical text and being a text are rigid properties:

$$\mathrm{PhTxt}(x) \rightarrow \forall t[\mathrm{pre}(x,t) \rightarrow \mathrm{PhTxt}(x)] \tag{2}$$

$$\mathrm{Txt}(x) \rightarrow \forall t[\mathrm{pre}(x,t) \rightarrow \mathrm{Txt}(x)] \tag{3}$$

5.1.2. Parthood

As far as general mereology is concerned I assume a system that is equivalent to the system which Smith et al. (2012) propound for `continuant_part_of` relation.

Definitions

$$\mathrm{pt}(s,s',t) \triangleq \mathrm{ppt}(s,s',t) \vee s = s' \tag{4}$$

$$\mathrm{o}(s,s',t) \triangleq \exists s''\ [\mathrm{pt}(s'',s,t) \wedge \mathrm{pt}(s'',s',t)] \tag{5}$$

$$\mathrm{Sum}(s,s',s'',t) \equiv \mathrm{pt}(s',s,t) \wedge \mathrm{pt}(s'',s,t) \wedge \forall s'''[\mathrm{pt}(s''',s,t) \rightarrow \mathrm{o}(s''',s',t) \vee \mathrm{o}(s''',s'',t)] \tag{6}$$

Axioms

$$\mathrm{ppt}(s,s',t) \rightarrow \mathrm{pre}(s,t) \wedge \mathrm{pre}(s'',t) \tag{7}$$

$$\neg\mathrm{ppt}(s,s,t) \tag{8}$$

$$\mathrm{ppt}(s,s',t) \wedge \mathrm{ppt}(s',s'',t) \rightarrow \mathrm{ppt}(s,s'',t) \tag{9}$$

$$\mathrm{ppt}(s,s'',t) \wedge \rightarrow \exists s'' \ [\mathrm{ppt}(s'',s',t) \wedge \neg \mathrm{o}(s,s'',t)] \tag{10}$$

$$\mathrm{pre}(s,t) \wedge \mathrm{pre}(s',t) \rightarrow \exists s'' \ \mathrm{Sum}(s'',s,s',t) \tag{11}$$

5.1.3. Location

The axioms characterising the relation of spatial occupancy are taken, with a slight modification due to the lack of the region predicate in the language of the current ontology, from Smith et al. (2012).

$$\mathrm{occ}(r,r,t) \tag{12}$$

$$\mathrm{ppt}(s,r',t) \rightarrow \mathrm{occ}(s,s,t) \tag{13}$$

$$\mathrm{ppt}(s,s',t) \wedge \mathrm{occ}(s',r',t) \rightarrow \exists r[\mathrm{ppt}(r,r',t) \wedge \mathrm{occ}(s,r,t)] \tag{14}$$

5.2. Physical Text Axioms

5.2.1. Existence and Time

Although I neglect here the artefactual aspects of physical texts (i.e., the fact that they were created by some agent(s) for a certain purpose, etc.), let me add an axiom that makes them contingent entities:

$$\mathrm{PhTxt}(x) \rightarrow \exists t \ \neg\mathrm{pre}(x,t) \tag{15}$$

5.2.2. Parthood

This section explains the specific features of the parthood relation among physical texts.

First, let me introduce some auxiliary definitions – most of them are just restriction of the top-level ontological relations to the domain of physical texts.

$$\mathrm{ppt_{ph}}(x,y,t) \triangleq \mathrm{ppt}(x,y,t) \wedge \mathrm{PhTxt}(x) \wedge \mathrm{PhTxt}(y) \tag{16}$$

$$\mathrm{pt_{ph}}(x,y,t) \triangleq \mathrm{pt}(x,y,t) \wedge \mathrm{PhTxt}(x) \wedge \mathrm{PhTxt}(y) \tag{17}$$

$$\mathrm{o_{ph}}(x,y,t) \triangleq \exists z[\mathrm{pt_{ph}}(z,x,t) \wedge \mathrm{pt_{ph}}(z,y,t)] \tag{18}$$

$$\begin{gathered}\mathrm{PhTxt}(y) \wedge \mathrm{PhTxt}(z) \rightarrow \\ \mathrm{Sum_{ph}}(x,y,z,t) \equiv \mathrm{pt_{ph}}(y,x,t) \wedge \mathrm{pt_{ph}}(z,x,t) \wedge \forall v[\mathrm{pt_{ph}}(v,x,t) \rightarrow \mathrm{o_{ph}}(v,y,t) \vee \mathrm{o_{ph}}(v,z,t)]\end{gathered} \tag{19}$$

$$\mathrm{CompPhTxt}(x,t) \triangleq \mathrm{PhTxt}(x) \wedge \neg\exists y \mathrm{ppt_{ph}}(x,y,t) \wedge \mathrm{pre}(x,t) \tag{20}$$

The last predicate, "CompPhTxt", is to represent complete physical texts, i.e., those physical texts that are not parts (in the sense of $\mathrm{ppt_{ph}}$) of any other physical texts. Note that it is a binary predicate, which relativises this property to times. So, being a complete physical text is not rigid. Consider for example a collection of papers that were not part of any other physical text at time t, but at some later time t' they are (physically) bounded together in a single volume.

Here come the axioms for parthood among physical texts. The first axiom below excludes the case when one physical text that was (a proper) part of another later becomes a whole whose (proper) part is the latter. So a physical text can grow, e.g., by new physical texts being added to it, but not in such a way that the whole it used to compose becomes its part. Axiom 22 is a version of the Strong Supplementation Principle for physical texts. The axiom was introduced to establish the extensionality of $\mathrm{ppt_{ph}}$ – see theorem 45 below. Although strong supplementation or extensionality are among the most controversial properties of parthood *simpliciter*, it seems to me that the most common objections (that is, counterexamples) do not hold for our $\mathrm{ppt_{ph}}$ (or $\mathrm{pt_{ph}}$). For instance, the fact that the same letters may make up different words is irrelevant because the same physical texts cannot make up *at a time* different physical texts. Axiom 23 expresses a kind of truism: each physical text has parts that are not physical texts. The next two axioms are to establish that $\mathrm{ppt_{ph}}$ is discrete. The last axiom in this group means that each physical text is built out of the atomic physical texts.

$$\mathrm{ppt_{ph}}(x,y,t) \to \neg\exists t'\,\mathrm{ppt_{ph}}(y,x,t') \tag{21}$$

$$\begin{gathered}\mathrm{pre}(x,t) \wedge \mathrm{pre}(y,t) \wedge \mathrm{PhTxt}(x) \wedge \mathrm{PhTxt}(y) \to \\ \{\neg\mathrm{pt_{ph}}(x,y,t) \to \exists z[\mathrm{pt_{ph}}(z,x,t) \wedge \neg\mathrm{o_{ph}}(z,y,t)]\}\end{gathered} \tag{22}$$

$$\mathrm{PhTxt}(x) \wedge \mathrm{pre}(x,t) \to \exists z[\neg\mathrm{PhTxt}(z) \wedge \mathrm{pt}(z,x,t)] \tag{23}$$

$$\mathrm{ppt_{ph}}(x,y,t) \to \exists z\{\mathrm{ppt_{ph}}(x,z,t) \wedge \neg\exists v[\mathrm{ppt_{ph}}(x,v,t) \wedge \mathrm{ppt_{ph}}(v,z,t)]\} \tag{24}$$

$$\mathrm{ppt_{ph}}(x,y,t) \to \exists z[\mathrm{ppt_{ph}}(z,y,t) \wedge \neg\exists v[\mathrm{ppt_{ph}}(z,v,t) \wedge \mathrm{ppt_{ph}}(v,y,t)]] \tag{25}$$

$$\mathrm{PhTxt}(x) \to \forall y\{\mathrm{pt_{ph}}(y,x,t) \to \exists z[\mathrm{pt_{ph}}(z,y,t) \wedge \forall v \neg\mathrm{ppt_{ph}}(v,z,t)]\} \tag{26}$$

5.2.3. Location

The spatial characterisation of physical texts is rather sparse:

1. Every physical text is located somewhere (at those times at which it exists).
2. There cannot be two physical texts located at the same region in space.[6]

$$\mathrm{pre}(x,t) \wedge \mathrm{PhTxt}(x) \to \exists r\ \mathrm{occ}(x,r,t). \tag{27}$$

$$\mathrm{occ}(x,r,t) \wedge \mathrm{occ}(y,r,t) \to [\mathrm{PhTxt}(x) \wedge \mathrm{PhTxt}(y) \to x = y]. \tag{28}$$

5.2.4. Precedence

First let me spell out the following auxiliary definition for text adjacency:

[6] Note that this does not exclude co-location of entities of other types.

$$\text{adj}_{\text{ph}}(x,y,t) \triangleq \text{prior}_{\text{ph}}(x,y,t) \wedge \neg\exists z[\text{prior}_{\text{ph}}(x,z,t) \wedge \text{prior}_{\text{ph}}(z,y,t)]. \tag{29}$$

The first group of axioms for the relation of precedence characterises the ontological categories of its arguments and its temporal ramifications:

$$\text{prior}_{\text{ph}}(x,y,t) \to \text{PhTxt}(x) \wedge \text{PhTxt}(y) \tag{30}$$

$$\text{prior}_{\text{ph}}(x,y,t) \to \text{pre}(x,t) \wedge \text{pre}(y,t) \tag{31}$$

The second group of axioms describes the basic, and rather obvious, formal properties of prior_{ph}. As in the case of text parthood, the last two axioms in this group guarantee that prior_{ph} is discrete.

$$\neg\text{prior}_{\text{ph}}(x,x,t) \tag{32}$$

$$\text{prior}_{\text{ph}}(x,y,t) \wedge \text{prior}_{\text{ph}}(y,z,t) \to \text{prior}_{\text{ph}}(x,z,t) \tag{33}$$

$$\text{prior}_{\text{ph}}(x,y,t) \to \exists z\{\text{prior}_{\text{ph}}(x,z,t) \wedge \neg\exists v[\text{prior}_{\text{ph}}(x,v,t) \wedge \text{prior}_{\text{ph}}(v,z,t)]\} \tag{34}$$

$$\text{prior}_{\text{ph}}(x,y,t) \to \exists z[\text{prior}_{\text{ph}}(z,y,t) \wedge \neg\exists v[\text{prior}_{\text{ph}}(z,v,t) \wedge \text{prior}_{\text{ph}}(v,y,t)]] \tag{35}$$

Note that I do not assume that prior_{ph} is total – the reason is that some physical texts are not linear - see the introduction to Aarseth (1997). Obviously, you can define the type of linear physical texts:

$$\begin{gathered}\text{LinearPhTxt}(x,t) \triangleq \\ \text{PhTxt}(x) \wedge \\ \forall y_1,y_2[\text{pt}_{\text{ph}}(y_1,x,t) \wedge \text{pt}_{\text{ph}}(y_2,x,t) \wedge \neg\text{o}_{\text{ph}}(y_1,y_2,t) \to \text{prior}_{\text{ph}}(y_1,y_2,t) \vee \text{prior}_{\text{ph}}(y_2,y_1,t)]\end{gathered} \tag{36}$$

The next group collects axioms that specify the formal connections between text precedence and text parthood. The first axiom says that one physical text cannot precede the other if at some point in time one of them was part of the other. The next two axioms imply that the precedence relation among parts of preceding physical texts follows the precedence relation among the texts themselves. Axiom 40 reveals the proper context of precedence: one physical text precedes another only within a unique physical text that contains both of them. Axioms 41 and 42 imply that each physical text has a start and an end – again not that they need not to be unique, e.g., a text with hyperlinks can have more than one end. The last axiom in this group is a restricted version of the mereological principle of unrestricted union. In the case of ppt_{ph} it is claimed that each pair of adjacent texts makes up a whole, which is a physical text as well.

$$\mathrm{prior_{ph}}(x,y,t) \rightarrow \neg\exists t'[\mathrm{pt_{ph}}(x,y,t') \vee \mathrm{pt_{ph}}(y,x,t')] \tag{37}$$

$$\mathrm{prior_{ph}}(x,y,t) \rightarrow \forall z[\mathrm{pt_{ph}}(z,x,t) \rightarrow \mathrm{prior_{ph}}(z,y,t)] \tag{38}$$

$$\mathrm{prior_{ph}}(x,y,t) \rightarrow \forall z[\mathrm{pt_{ph}}(z,y,t) \rightarrow \mathrm{prior_{ph}}(x,z,t)] \tag{39}$$

$$\mathrm{prior_{ph}}(x,y,t) \rightarrow \exists! z[\mathrm{CompPhTxt}(z,t) \wedge \mathrm{pt_{ph}}(x,z,t) \wedge \mathrm{pt_{ph}}(y,z,t)] \tag{40}$$

$$\mathrm{PhTxt}(x) \rightarrow \exists y[\mathrm{pt_{ph}}(y,x,t) \wedge \forall z[\mathrm{pt_{ph}}(z,x,t) \rightarrow \neg\mathrm{prior_{ph}}(z,y,t)]] \tag{41}$$

$$\mathrm{PhTxt}(x) \rightarrow \exists y[\mathrm{pt_{ph}}(y,x,t) \wedge \forall z[\mathrm{pt_{ph}}(z,x,t) \rightarrow \neg\mathrm{prior_{ph}}(y,z,t)]] \tag{42}$$

$$\mathrm{adj_{ph}}(x,y,t) \rightarrow \exists z\ \mathrm{Sum_{ph}}(z,x,y,t) \tag{43}$$

The last group of axioms for precedence concerns its connection to spatial occupancy. Again this characterisation is sparse: two texts such that one precedes another do not occupy overlapping regions.

$$\mathrm{prior_{ph}}(x,y,t) \rightarrow [\mathrm{occ}(x,r,t) \wedge \mathrm{occ}(y,r't) \rightarrow \neg\mathrm{o}(r,r',t)] \tag{44}$$

5.2.5. Identity Criteria

The axioms above imply the following theorems, which can play the role of the identity criteria for physical texts:

$$\exists z[\mathrm{ppt_{ph}}(z,x,t) \vee \mathrm{ppt_{ph}}(z,y,t)] \rightarrow \{x = y \equiv \forall z \exists t[\mathrm{ppt_{ph}}(z,x,t) \equiv \mathrm{ppt_{ph}}(z,y,t)]\} \tag{45}$$

$$\mathrm{PhTxt}(x) \wedge \mathrm{PhTxt}(y) \rightarrow \{x = y \equiv \exists r,t[\mathrm{occ}(x,r,t) \equiv \mathrm{occ}(y,r,t)]\} \tag{46}$$

5.2.6. Intentionality

This section contains the formal characteristic of the intentional aspect of physical texts. As I mentioned above in general physical texts have this idiosyncratic aspect of "pointing out", but not every physical text must be intentional (in this sense). Still, I assume that being intentional is rigid and that it is an essential feature of complete physical texts.

$$\mathrm{int_{ph}}(x,y,t) \rightarrow \mathrm{PhTxt}(x) \tag{47}$$

$$\mathrm{int_{ph}}(x,y,t) \rightarrow \mathrm{pre}(x,t) \tag{48}$$

$$\exists y\ \mathrm{int_{ph}}(x,y,t) \rightarrow \forall t[\mathrm{pre}(x,t) \rightarrow \exists y\ \mathrm{int_{ph}}(x,y,t)] \tag{49}$$

$$\mathrm{CompPhTxt}(x,t) \rightarrow \exists y\ \mathrm{int_{ph}}(x,y,t) \tag{50}$$

Note that I do not assume that a physical text always refers to the same object(s), so the theory makes room for shifts in meaning. Note also that I do not assume, as the above quote about IAO *seems* to assume, that any text is always about something *else*. That is, $\mathrm{int_{ph}}$ is not claimed to be irreflexive with respect to its first two arguments, so statements of the form "$\mathrm{int_{ph}}(x,x,t)$" are consistent with the ontology. The reason for this is obvious – semantic antinomies are the most famous examples of texts that are about themselves.

5.2.7. Equivalence

Finally, let me formally characterise the relation of equivalence between physical texts.

First, we should specify the ontological categories of its arguments and its temporal ramifications:

$$\mathrm{eq_{ph}}(x,y,t) \to \mathrm{PhTxt}(x) \wedge \mathrm{PhTxt}(y) \tag{51}$$

$$\mathrm{eq_{ph}}(x,y,t) \to \mathrm{pre}(x,t) \vee \mathrm{pre}(y,t) \tag{52}$$

Then we need to make it an equivalence relation:

$$\mathrm{pre}(x,t) \wedge \mathrm{PhTxt}(x) \to \mathrm{eq_{ph}}(x,x,t) \tag{53}$$

$$\mathrm{eq_{ph}}(x,y,t) \to \mathrm{eq_{ph}}(y,x,t) \tag{54}$$

$$\mathrm{eq_{ph}}(x,y,t) \wedge \mathrm{eq_{ph}}(y,z,t) \to \mathrm{eq_{ph}}(x,z,t) \tag{55}$$

Now comes two groups of axioms that characterise $\mathrm{eq_{ph}}$ as a kind of congruence for text parthood and text intentionality. The first group, which is focused on parthood, starts from axiom 56 that excludes hat a physical text that is part of another physical text may be equivalent to it. Then there are two axioms whose role is to establish that equivalent physical texts have homomorphic mereological structures.

$$\mathrm{ppt}(x,y,t) \to \neg\mathrm{eq_{ph}}(x,y,t) \tag{56}$$

$$\mathrm{eq_{ph}}(x_1,x_2,t) \wedge \mathrm{ppt_{ph}}(y_1,x_1,t) \to \exists y_2[\mathrm{eq_{ph}}(y_1,y_2,t) \wedge \mathrm{ppt}(y_2,x_2,t)] \tag{57}$$

$$\mathrm{eq_{ph}}(x_1,x_2,t) \wedge \mathrm{ppt_{ph}}(x_1,y_1,t) \to \exists y_2[\mathrm{eq_{ph}}(y_1,y_2,t) \wedge \mathrm{ppt}(x_2,y_2,t)] \tag{58}$$

The group for intentionality has just one axiom, which guarantees that equivalent physical texts refer to the same object(s).

$$\mathrm{eq_{ph}}(x_1,x_2,t) \wedge (\mathrm{pre}(x_1,t) \wedge \mathrm{pre}(x_2,t)) \to [\mathrm{int_{ph}}(x_1,y,t) \equiv \mathrm{int_{ph}}(x_2,y,t)] \tag{59}$$

5.3. Text Axioms

Now I will characterise texts as abstractions over classes of text tokens.

5.3.1. Bridging Principles

A text is understood here as an object that is constantly and generically dependent in its existence on certain physical texts. The exact content of this claim is detailed by the following axioms below.

$$\mathrm{dep_{txt}}(x,y) \to \mathrm{Txt}(x) \wedge \mathrm{PhTxt}(y) \tag{60}$$

$$\mathrm{PhTxt}(x) \to \exists! y[\mathrm{Txt}(y) \wedge \mathrm{dep_{txt}}(y,x)] \tag{61}$$

$$\mathrm{Txt}(x) \to \exists y[\mathrm{PhTxt}(y) \wedge \mathrm{dep_{txt}}(x,y)] \tag{62}$$

$$\mathrm{dep_{txt}}(x,y_1) \wedge \mathrm{dep_{txt}}(x,y_2) \to \forall t[\mathrm{pre}(y_1,t) \vee \mathrm{pre}(y_2,t) \to \mathrm{eq_{ph}}(y_1,y_2,t)] \tag{63}$$

$$\mathrm{dep_{txt}}(x,y) \to \forall t\{\mathrm{pre}(x,t) \to \exists z[\mathrm{eq_{ph}}(z,y,t) \wedge \mathrm{pre}(z,t)]\} \tag{64}$$

$$\mathrm{dep_{txt}}(x,y) \to \forall t[\mathrm{pre}(y,t) \to \mathrm{pre}(y,t)] \tag{65}$$

$$\mathrm{eq_{ph}}(x_1,x_2,t) \to \forall z[\mathrm{dep_{txt}}(z,x_1) \equiv \mathrm{dep_{txt}}(z,x_2)] \tag{66}$$

Note that the axioms above does not exclude that some physical texts are texts, so we need one more principle:

$$\neg\exists x[\mathrm{PhTxt}(x) \wedge \mathrm{Txt}(x)] \tag{67}$$

5.3.2. "Emergent" Definitions

On the basis of these axioms one can introduce the definitions that concern the counterparts of the relations I used to characterise physical texts, e.g., the counterparts of parthood, text precedence, and intentionality. All of the them will follow the same structural pattern:

$$\rho_{\mathrm{txt}}(\alpha_1,\alpha_2,\ldots,\alpha_n,\ldots,\alpha_{n+m})$$
$$\triangleq$$
$$\exists\beta_1,\beta_2,\ldots\beta_n[\mathrm{dep_{txt}}(\alpha_1,\beta_1) \wedge \cdots \wedge \mathrm{dep_{txt}}(\alpha_n,\beta_n) \wedge \rho_{\mathrm{ph}}(\beta_1,\beta_2,\ldots,\beta_n,\ldots,\alpha_{n+m})] \tag{68}$$

where:

1. $n \leq 1$ and $m \leq 0$;
2. "ρ_{ph}" stands for a relation between physical texts;
3. "ρ_{txt}" stands for the counterpart of this relation among abstract texts.

More precisely speaking, ρ_{ph} in 68 ranges over the following set of symbols:
$\{\mathrm{ppt_{ph}},\mathrm{pt_{ph}},\mathrm{Sum_{ph}},\mathrm{o_{ph}},\mathrm{prior_{ph}},\mathrm{adj_{ph}},\mathrm{CompPhTxt},\mathrm{LinearPhTxt},\mathrm{int_{ph}}\}$,
so actually 68 is a schema for nine definitions. For instance, the proper parthood for abstract texts can be defined as below:

$$\mathrm{ppt_{txt}}(x,y,t) \triangleq \exists x_1 \exists y_1[\mathrm{dep_{txt}}(x,x_1) \wedge \mathrm{dep_{txt}}(y,y_1) \wedge \mathrm{ppt_{ph}}(x_1,y_1,t)] \tag{69}$$

5.3.3. "Emergent" Theorems

On the basis of the above axioms (and definitions) one can prove several theorems, which characterise the formal properties of these "emergent" relations. It turns out that almost all properties of the aforementioned relations between physical texts are exhibited by their abstract text counterparts – except for those that involve spatial occupancy and equivalence. That is to say, for each axiom in the range from 15 to 50 – except for 23 – such that

1. it contains one of the symbols from
   $\{\text{ppt}_{\text{ph}}, \text{pt}_{\text{ph}}, \text{Sum}_{\text{ph}}, \text{o}_{\text{ph}}, \text{prior}_{\text{ph}}, \text{adj}_{\text{ph}}, \text{CompPhTxt}, \text{LinearPhTxt}, \text{int}_{\text{ph}}\}$
2. it does not contain occ or eq_{ph}

if we replace the aforementioned symbols by their counterparts introduced by schema 68, then the resulting formula is a theorem. The proof of this claim consists of the subproofs for the respective formulas. The latter can be accessed from the webpage mentioned in Introduction. One can say that relations that are relevant for abstract texts, i.e., temporal existence, identity, parthood, text precedence, and intentionality, are emergent on their counterparts among physical texts.

The fact that the counterpart of axiom 23 for texts is not a theorem should be of little surprise. 23 postulates the existence of non-textual parts of the physical texts and the bridging principles above imply that abstract texts are ontologically grounded in physical texts.

Finally, the above axioms (and definitions) entail an additional identity criterion for abstract texts.

$$\text{Txt}(x) \wedge \text{Txt}(y) \rightarrow \{x = y \equiv \exists z[\text{dep}_{\text{txt}}(x,z) \equiv \text{dep}_{\text{txt}}(y,z)]\} \tag{70}$$

5.4. Metalogical properties

The consistency of the above ontology of text is shown by a Σ-sorted interpretation $< \mathfrak{M}, \beta >$ such that:[7]

1. $\mathfrak{M} =< M, I >$ is a Σ-sorted structure such that
   (a) $M = \{1,2,3\}$
   (b) $I(S_{time}) = \{1\}$
   (c) $I(S_{region}) = \{2\}$
   (d) $I(S_{object}) = \{3\}$
   (e) $I(\text{pre}) = \{< 3,1 >\}$
   (f) $I(\text{ppt}) = \{< 3,3,1 >\}$
   (g) $I(\text{occ}) = \{< 2,2,1 >\}$
   (h) for every other predicate ρ from P, $I(\rho) = \emptyset$
2. β is a Σ-valuation in $\mathfrak{M}$.

Other metalogical properties of this theory, e.g., completeness, are currently unknown.

6. Further Work

As I mentioned in section 3 above the main distinction of this ontology, i.e., the one that distinguish text tokens from texts, can be seen as an impoverished version of the FRBR account of works, expressions, manifestations, and items. Still, the approach taken here can be applied to the FRBR distinction in a relatively straightforward manner. First we need to accommodate the equivalence relation so that it will fit the FRBR perspective. For

[7] For the notion of sorted interpretation relative to (sorted) signature and the related notions, see (Socher-Ambrosius and Johann, 1997, p. 172-173).

example, if we identify text tokens with FRBR items, then the equivalence relation needs to be sufficiently discriminative to draw a distinction between FRBR manifestations and expressions. This adjustment can be implemented in two, ontologically divergent, ways. One option is to add more ontological layers onto the above bipartite ontology. So one can identify text tokens with FRBR items and (abstract) texts with FRBR manifestations. Then one can add a new equivalence relation to segregate the latter so that one can ascend one level up to expressions. The other option is to remain within a two-layer ontology and instead adding equivalence relations between (abstract) texts, one can add the second equivalence relation between text tokens. Then the first equivalence relation will relate just those text tokens that have the same shape and the same material, while the second relation will "ignore" the latter aspect. As a result, the equivalence classes of the former may correspond to FRBR manifestations and the equivalence classes of the former may correspond to FRBR expressions. This extended ontology will remain two-layer with text tokens on one layer and different types of their abstractions on the second layer.

Acknowledgements

This research has been supported by the DEC-2012/07/B/HS1/01938 grant funded by National Science Centre (Poland).

References

Aarseth, E. J. (1997). *Cybertext: perspectives on ergodic literature*. JHU Press.

Arp, R., Smith, B., and Spear, A. D. (2015). *Building Ontologies with Basic Formal Ontology*. MIT Press.

Bekiari, C., Doerr, M., Buf, P. L., and Riva, P. (2015). FRBR Object-Oriented Definition and Mapping From FRBRER, FRAD and FRSAD.

IFLA (1998). *Functional Requirements for Bibliographic Records: Final Report*. K. G. Saur.

Kaneiwa, K. (2004). Order-sorted logic programming with predicate hierarchy. *Artificial Intelligence*, 158(2):155–188.

Smith, B., Bona, M. A. J., Brochhausen, M., Ceusters, W., Courtot, M., Dipert, R., Goldfain, A., Grenon, P., Hastings, J., Hogan, W., Jacuzzo, L., Johansson, I., Mungall, C., Natale, D., Neuhaus, F., Overton, J., Petosa, A., Rovetto, R., Ruttenberg, A., Ressler, M., Rudniki, R., Seppl, S., Schulz, S., and Zheng, J. (2012). Basic formal ontology 2.0.

Socher-Ambrosius, R. and Johann, P. (1997). *Deduction Systems*. Springer.