Developing a multilayer semantic annotation scheme based on ISO standards for the visualization of a newswire corpus

Purificação Silvano[1], António Leal[2], Fátima Silva[3], Inês Cantante[4], Fátima Oliveira[5] & Alípio Mário Jorge[6]
[1,2,3,4,5]University of Porto/ Centre of Linguistics [6]University of Porto/ INESC
[1]msilvano@letras.up.pt, [2]jleal@letras.up.pt,[3]mhenri@letras.up.pt, [4]cantante.ines@gmail.com, [5]moliv@letras.up.pt, [6]amjorge@fc.up.pt

Abstract

In this paper, we describe the process of developing a multilayer semantic annotation scheme designed for extracting information from a European Portuguese corpus of news articles, at three levels, temporal, referential and semantic role labelling. The novelty of this scheme is the harmonization of parts 1, 4 and 9 of the ISO 24617 *Language resource management - Semantic annotation framework.* This annotation framework includes a set of entity structures (participants, events, times) and a set of links (temporal, aspectual, subordination, objectal and semantic roles) with several tags and attribute values that ensure adequate semantic and visual representations of news stories.

1 Introduction

The development of an annotation framework can be an overwhelming task, even more when its purpose is to account for different linguistic phenomena. However, as challenging as it may be, designing an annotation scheme is an indispensable step to generate language resources that can be the starting point of fundamental corpus-based linguistic research.

When deciding on an annotation framework, one has to take into consideration several factors (Pustejovsky et al., 2017), such as main objectives of the annotation, the linguistic phenomena under analysis, the corpus genre, and the nature of the annotation, and weigh in the advantages and disadvantages of adapting/ adopting an existing model, or of creating one. Ideally, the model is custom designed to deal with all the specificities of a particular project, but also broad enough so that it can be applied to other datasets. In fact, with the growth of the Semantic Web and Linguistic Linked Data (Chiarcos et al., 2020), interoperability is key to read and to interpret linguistic resources (Ide and Pustejovsky, 2010).

With all the above-mentioned provisos in mind, we developed a multilayer semantic annotation scheme by combining three standards from the *Language resource management-Semantic annotation framework*: *Part 1- Time and events* (ISO-24617-1), *Part 4- Semantic roles* (ISO-24617-4) and Part 9- *Referential annotation framework* (ISO-24617-9). In addition to promoting interoperability, our model has proven to be able to markup manually the relevant features of the genre news to generate visual representations of their narratives. Moreover, our proposal operationalizes the integration of three different standards in the same framework, which is, to the best of our knowledge, a novelty.

This multilayer semantic annotation scheme was designed to annotate a European Portuguese corpus of news articles in three different, but complementary, levels, temporal, referential and thematic, within the Text2Story project[1], which aims to extract narratives from news, represent them in intermediate data structures, and make these available to subsequent media production processes, i.e., visualizations such as message sequence charts (MSC) and knowledge graphs (KG). In this paper, we document the decision-

[1] https://text2story.inesctec.pt/

Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation, pages 1–13
June 16–17, 2021. ©2021 Association for Computational Linguistics

making process about which annotation format to adopt, what adjustments to make, and how to harmonize the three layers into an integrated and wide-ranging model.

2 Background and Motivation

News may frequently assume the format of a story that reports on current events involving one or more entities in given time and place. In addition to the main event, however, news stories typically present contextual content that allows connecting it to others, explaining the circumstances and consequences of its occurrence. It may also include other complementary information that frames, comments, clarifies, or evaluates the reported events (Caswell and Dörr, 2019; Choubey et al., 2020; cf. also van Dijk, 1985; Bell, 1991). A complete story usually answers six questions: what, who, where, when, why, and how, that is, 5W1H (a.o. Bonet-Jover et al., 2021), following a top-down organization, corresponding to an inverted pyramid discourse structure (cf. Rabe 2008), in which information flows in decreasing order of importance. A news organization structure usually features a title, a lead, and the body. In many cases, the lead or introductory paragraph condenses the answers to the above six questions and is followed by complementary information (a.o. Thomson et al., 2008; Norambuena et al., 2020). Sometimes, the answer to some of the questions is distributed throughout the text (Bonet-Jover et al., 2021). Because of this organization, events frequently follow a non-chronological order, presenting a complex time structure regarding other kinds of narratives (Zahid et al., 2019). Besides, the narrative may return to previous data, as well as adding information (a.o. van Dijk, 1985; Thomson et al., 2008; Choubey et al., 2020).

Establishing the temporal sequencing of events, their participants, and interrelations is crucial to understand the news story, and ultimately to extract the narratives to be represented graphically by means of MSC (Harel and Thiagarajan, 2003) or KG (Ehrlinger and Wöß, 2016), which is our project's main objective. These visualizations by portraying the narratives more schematically can be of great interest to news agencies, for example. The more overarching and rigorous the annotation the more informative is the visualization, and, in the case of news articles, this requires featuring participants, events and times, as well the relationships between them. For these reasons, the annotation scheme that we designed encompasses three intertwined semantic layers: temporal, referential and thematic. Since our aim was to adopt a coherent and interoperable annotation scheme with these three layers, and because none of the existing proposals satisfied these requisites, we designed an annotation scheme which compatibilizes three ISO.

3 Related work

Over the last years, there has been a proliferation of multilayer corpora, that is, corpora that "contain mutually independent forms of information, which cannot be derived from one another reliably" (Zeldes, 2019: 4). These layers can be defined in an independent way and they "are explicitly analyzed using multiple, independent annotation schemes" (Zeldes, 2019: 7), or resorting to one unique scheme that integrates all the layers. In fact, an in-depth analysis of the relevant literature reveals that there are many different types of multilayer annotation schemes. In the remainder of this section, we will only present a brief overview of some of those proposals.

One of the most well accomplished and far-reaching multilayer annotation schemes is the one developed within the *Groningen Meaning Bank* (GMB) (Basile et al., 2012; Bos et al., 2017). Besides morphological and syntactic annotation, it comprises different semantic annotation levels, such as named entity recognition, temporal features, and thematic roles. The adopted semantic formalism is an extension of *Discourse Representation Theory* (Kamp and Reyle, 1993), which renders a semantic representation (*discourse representation structures*) that unifies the various layers. Another important feature of this scheme is that it was designed to analyze linguistic phenomena in texts, instead of only sentences, and it has been used quite successfully in 10,000 texts from different genres, namely news and fables. Its implementation requires a human-aided machine annotation insofar as it employs NLP software such as an automatic tagger for named entity recognition, VerbNet (Schuler, 2005) for semantic role labelling, a semantic analyzer for coreference, and then a module Boxer (Bos, 2005, 2008; Curran et al. 2007), responsible for the overall semantic analysis, but also relies on the input of experts and general public. Although, in terms of semantic annotation, it is one of the most complete, this scheme lacks information about more referential relations. Moreover, since the temporal annotation is based on DRT-language, it does not integrate tags about lexical and contextual meaning with bearing on temporal interpretation, namely a more diversified class of events, and other link types between events.

Other multilayer annotation schemes have been developed for *Manually Annotated Sub-Corpus* (MASC) (Ide et al., 2008), *Georgetown University Multilayer Corpus* (GUM) (Zeldes and Simonson, 2016; Zeldes, 2017), *OntoNotes* (Hovy et al., 2006), for *AMALGUM* (Gessler et al., 2020), or *SenSem* (Fernández and Vázquez, 2014), just to name a few, but none of those provide a comprehensive and harmonized semantic framework suitable to handle the linguistic phenomena that we need to address.

For European Portuguese (EP), one can point out the scheme used in CINTIL DeepBank (Branco et al., 2010), which is a corpus of Portuguese news and novels that is annotated with several grammatical information (morphological, syntactic, and semantic) for each sentence. Currently, there are 32497 sentences, mainly from news, which were semi-automatically annotated with Treebank, DependencyBank, Propbank, and LogicalFormBank (with formal representations of the sentences meanings using Minimal Recursion Semantics). However, the CINTIL DeepBank's scheme does not include a level for referential annotation, nor for temporal annotation. The fact that only the sentences that the grammar can parse are included in the corpus is a downside. Additionally, though each level of annotation can be accessed separately, a unifying formalism that combines all the layers is missing.

Regarding schemes aimed exclusively at semantic annotation, some are intended to handle a specific phenomenon, resort to non-standardized markup language, and are not widely known (cf. for an overview (Gries and Berez, 2017). Moreover, the majority deals with lexical problems, such as word disambiguation, and less with compositional semantics. The scarcity of proposals within this branch of semantics can be explained by the complexity underlying the process of annotating semantically data at a sentential and textual level. This task requires not only a great amount of time, but also a wide variety and substantial number of resources. Nonetheless, semantic schemes to represent the meaning of texts are of utmost relevance to the development of different applications.

4 The Annotation scheme

4.1 The process

Building a bootstrapping annotation scheme is a very complex and time-consuming endeavor involving different phases. After the literature review, we started by defining the tags and their attributes first for the temporal layer, then for the referential level, and finally for the semantic role labelling. To create a model, we followed the MATTER (Pustejovsky and Stubs, 2012) sub-cycle, MAMA, with four steps, (1) model, (2) annotate, (3) evaluate and (4) revise. This process allowed us to identify and resolve the scheme's inconsistencies, gaps and incompatibilities, and to gradually improve it so that it could properly account for the linguistic data, and to deliver the necessary input for the visualization task. This cycle was repeated several times until we were satisfied with the model. The annotation tool that we used, BRAT (brat rapid annotation tool) (Stenetorp et al., 2012), enabled the updates of the annotation scheme without having to rebuild the whole scheme.

4.1.1 Temporal Layer

Temporal interpretation plays a crucial part in understanding how the events are organized in natural language texts. For this reason, extraction of temporal information has been receiving a lot of attention within NLP during the past few years. One approach to extract temporal features, and eventually to rebuild chronological sequences of events, is designing a suitable annotation scheme. In this field, research has started with the extraction of time expressions in message understanding conferences (MUCs) and progressed to relating events to times (eg. Filatova and Hovy, 2001; Katz and Arosio, 2001; Song et al., 2016). From the growing investment on temporal extraction, on the one hand, and from its usefulness, on the other hand, ensued not only a significant number of corpora annotated according to different schemes, but also annotation standards. One of these standards is TimeML (Pustejovsky et al., 2003a, 2003b), based on the work of Setzer (2001), Setzer & Gaizauskas (2000a, 2000b, 2001) and Ferro et al. (2003), from which ISO-TimeML (ISO 24617-1) stemmed.

ISO-TimeML, a model grounded on linguistic approaches (eg. Reichenbach, 1947; Comrie, 1985), defines a full-fledged markup language that permits a fine-grained annotation of time expressions, events, and temporal relations between events and between events and time expressions. Its efficacy and productivity in capturing the text's temporal structure is evidenced by corpora such as TIDES Temporal Corpus (Gerber et al., 2002), TimeBank (Pustejovsky et al., 2003b), composed of news articles, or Sun et al. (2013)'s corpus with clinical narratives. Costa (2012) and Costa and Branco (2010, 2012) use TimeML to annotate for the first time a EP corpus with temporal information, TimeBankPT. This corpus, nonetheless, only comprises the translations of texts from the original TimeBank, as well as the same annotations with some adaptations required by language specificities.

Compared to the scheme employed by TimeBankPT, the temporal tagset and linkset that we subscribe follow more closely ISO-24617-1. As expected, bearing in mind the project's main aim, that is, visualization of news narratives, and the necessity of not overloading the scheme with unnecessary information, some tags and links were excluded. Thus, for the temporal layer, our scheme incorporates two tags, event and times, and three links, temporal link (TLink), aspectual link (ALink) and subordination link (SLink).

The tag event marks eventualities (Bach, 1985), represented by tensed or untensed verbs, nominalizations, adjectives, predicative constructions or prepositional complements. The combination of all the required attributes, class, part of speech, tense, aspect, verb form, mood, modality and polarity, provides the necessary information about temporal, aspectual and modal features of events. With respect to the values for each attribute, we maintained the ones established by ISO-24617-1, namely for Italian, but added in the attribute mood the value *future* to account for its modal uses, and the modality values *dever* ('must'), *poder* ('can'), *ter de* ('have to') and *ser capaz de* ('be able to').

Regarding the tag times, we adopted a very simple scheme, which meets the needs of our project. The attributes that incorporate our annotation scheme are the required ones, according to ISO-24617-1, that is, type (date, time, duration and set) and value (the specific value of the type). We have also integrated two optional attributes: temporal function with the value publication time and anchor time, which are pertinent to process time expressions common in news articles, like *hoje* 'today', *na sexta-feira* 'Friday'.

The sequencing of the events, that is, their ordering, is essential to depict the way the narrative evolves in time. ISO-24617-1 specifies the adequate manner to establish the events timeline, as well as the relations between events and time expressions by postulating TLinks, which we integrated in our scheme. In turn, the ALink, by specifying the relation between aspectual verbs and their event arguments, gives crucial input to create the visualizations of the events. The relevance of the SLink derives from the fact that the news articles frequently include contexts of subordinating relationships between events. We omitted the measuring link (MLink) because the information it conveys is already captured to a certain extent by the value duration for the attribute type of tag times. The values for the three links of our model are the ones proposed by ISO-24617-1.

4.1.2 Referential Layer

Pointing out to the referring expressions in a text, identifying the discourse entities denoted by those expressions, and establishing the links between them are key tasks to reference annotation, and underly referential phenomena in discourse, such as anaphora.

In our corpus, those referring expressions correspond to named entities, or participants that play an important role in the story. Therefore, we needed a framework to deal with named entities recognition and their relation throughout the news texts. ISO-24617-9 met these needs, as it is a meta-model of referential annotation that articulates the discourse domain with the linguistic domain, contributing to a comprehensive representation of the discourse entities, the referring expressions that denote them, and their relations.

Despite following the standard in its overall guidelines, we did not annotate all its categories, and both discourse entity structures and referential expression structures were kept as simple as possible, to avoid overloading the process of annotation: the former include only information concerning the lexical head (noun, pronoun), whereas the latter include information concerning domain (individuation and types) and involvement. The individuation attribute, with the values set, individual and mass, follows ISO-24617-9 definitions, while for involvement we defined the values: *0* (the empty set); *1* (a set with only one entity); *>1* (a set with more than one entity, but less than the totality of entities in the domain); *all* (the totality of entities in the domain = universal quantification); *undef* (undefined involvement).

As for types, since ISO 24617-9 does not provide a typology of named entities, we selected, considering our corpus text genre and the purpose of the project, a tagset of six named entities: PER, ORG, LOC, OBJ, NAT, OTHER. In fact, the definition of named entities is neither easy nor consensual, and there are several typologies for their classification, being the number and types of entities influenced by factors, such as the domain from which they are extracted or the purpose of its classification (for a survey on this topic, see, a.o. Nouvel et al., 2016; Goyal et al., 2018). This tagset is an adaptation of general categories depicted in the named entity classification typologies used in many other corpora, including multilayer ones. The first three named entities are common to all the annotated corpora while the others may vary.

In what concerns the relations included in ISO-24617-9, we did not include in our specifications the lexical relational links between entity structures and referring expressions (eg. synonym, antonym, hyponym, meronym), the referential status of referring expressions (old/new), and the properties of discourse entities (abstractness, animacy, alienability, natural gender and cardinality), because they were not necessary for the visual representations of news. As a matter of fact, it is more useful for visualization to mark two linguistic expressions as referring to the same participant. Thus, our analysis only considers the proposed objectal links (objectalIdentity, partOf, subset, memberOf and referentialDisjunction) between discourse entities, which allows to represent nominal anaphora's mechanisms. Unlike many studies that focus on anaphora resolution and depict only coreferential mechanisms, leaving out other types of relations, the adopted framework allows for the marking of different types of anaphoric linkage between entities, namely direct and indirect anaphora.[2]

4.1.3 Semantic Role Layer

The task of semantic role labelling for English texts usually uses one of the following frameworks (see also ISO 24617-4, Annex B): FrameNet (Baker et al., 1998), VerbNet (Schuler, 2005), PropBank (Palmer et al., 2005), EngVallex (Cinková, 2006), and LIRICS (Petukhova and Bunt, 2008).

As for EP data, there are some proposals that approach the issue of semantic role labelling, typically using the methodology of PropBank and VerbNet. However, these proposals have a very narrow scope, working with small datasets and small lists of (typically) verbs. Some examples of these works are PropBankPT (Branco et al., 2012), a corpus of 3406 sentences translated from the *Wall Street Journal*, and annotated with information concerning constituency structure (phrase constituency and grammatical relations) and semantic roles; and CINTIL-PropBank (Branco et al., 2012), a corpus of 10039 sentences extracted from news and novels, and annotated with information concerning constituency structure and semantic roles. There is also ViPer (Talhadas et al., 2013), a verbal lexical database with information about the verb's arguments semantic roles (using PropBank approach) manually annotated. However, there are some aspects of the semantic roles list that is used that can be problematic for our project (for instance, event-denoting nouns are treated as arguments of the "occurrence" type, instead of being treated as events, like in ISO-24617-1).

So, the semantic role labelling task in our project could not be based on previous work done for EP, and it had to be done from scratch. The easier way to do so was to use some established framework and adapt it to EP, but the methodology typically used in frameworks designed for English (eg. FrameNet) requires that, for each verb, a frame be

[2] The Universal Anaphora initiative (https://universalanaphora.github.io/UniversalAnaphora/) has been working towards a proposal markup scheme compatible with Universal Dependencies, and that codifies different aspects of the anaphoric phenomena.

constructed, and the construction of each frame entails many examples with the same verb and their analysis (to identify all the meanings the verb can have and all the constructions in which it can occur), to determine its semantic selection. This work would be colossal, and impracticable taking into account the time frame and objectives of the project. Therefore, we needed a framework that would allow semantic annotation to be limited to the analysis of concrete examples of the news to be annotated. We started working with the framework provided by LIRICS, which was the most appropriate for the task. Furthermore, as LIRICS was the basis for the construction of the ISO standard for thematic annotation, there would be fewer potential problems when integrating semantic role annotation with referential and temporal annotation.

Consequently, in our project, we annotate semantic roles following ISO-24617-4 specifications in what concerns semantic roles. We do not construct entity structures, nor event structures in this level of annotation. Instead, we use the entity structures constructed in the referential annotation to deal with non-event discourse entities, and the entity structures constructed in the temporal annotation to deal with event discourse entities. The semantic role annotation consists in establishing the thematic relation between predicates and their arguments and modifiers.

4.2 Harmonizing Different Layers

The foregoing describes how the markup language used in each layer of our annotation scheme was extracted from three different standards. Although they comply with the principles for semantic annotation (ISO-24617-6), in fact, they were elaborated separately and assynchronically, and they lack information about how to combine them with each other. ISO-24617-6, in addition to defining some overall guidelines for the semantic annotation framework (SemAF), attempts at tackling some overlaps and inconsistencies between the different parts of the SemAF, but its coverage is limited. This means that, when combining different parts of the SemAF, as we did, it is expected that not only some incompatibilities may arise, but also some loose-ends and gaps may be left unsolved. Proposals such as Bunt (2019) improve the absence of some information in one particular part of SemAF by resorting to some notations from other parts of the ISO. Gaizauskas and Alrashid (2019), for instance, put forward a scheme with some annotations from ISO-24617-1/7, but do not refer to issues related to incompatibilities. Therefore, in the process of constructing our model, we had to overcome these difficulties in order to obtain a fully integrated scheme.

We began by modelling the types of structures as entity structures and link structures, and defined subtypes for each type, as described in Figure 1.

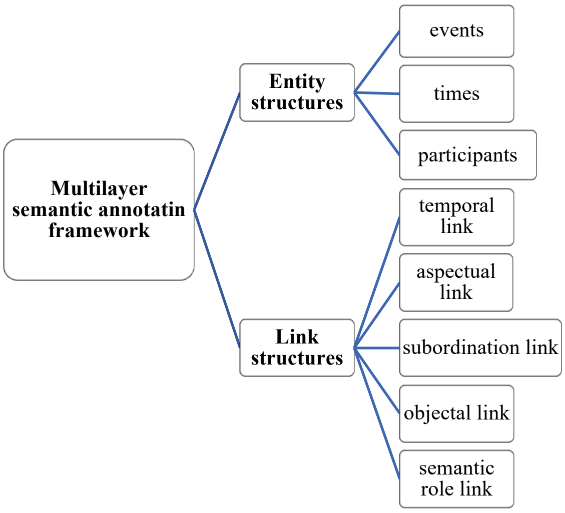

Figure 1: Text2story semantic annotation framework

This annotation structure is the first step to guarantee that all the layers are combined into a coherent annotation scheme. The entity structures, regardless of the layer to which they are associated, are available to be related among them by different types of link structures. Such unifying approach facilitates a uniform semantic representation in discourse representation structures (DRS).

The next step was to decide on the attributes and their respective values, so the information they codified would be compatible and not repetitive, as explained in the previous sections. The final annotation scheme is presented in Table 1.

ENTITY STRUCTURES		
EVENTS	class	occurrence, state, reporting, perception, aspectual, I-action, I-state
	type	state, process, transition
	pos	verb, noun, adjective, preposition
	tense	present, past, future, imperfect, none
	aspect	progressive, perfective, imperfective, imperfective-progressive, perfective-progressive, none
	vform	none, gerundive, infinitive, participle
	mood	none, subjunctive, conditional, future, imperative
	modality	*dever, poder, ter de, ser capaz de*
	polarity	negative, positive
TIME	type	date, time, duration, set
	value	specific value
	temporal function	publication_time
	anchortime	time ID (select relevant time)
PARTICIPANTS	lexical head	noun, pronoun
	domain	individuation: set, individual, mass
		types: per, org, loc, obj, nat, other
	involvement	0,1, >1, all, undefined
LINK STRUCTURES		
Temporal links		before, after, includes. is_included, during, simultaneous, identity, begins, ends, begun_by, ended_by
Aspectual links		initiates, culminates, terminates, continues, reinitiates
Subordination links		intensional, evidential, neg_evidential, factive, counter_factive, conditional
Objectal links		objectalIdentity, partof, subset, memberOf, referentialDisjunction
Semantic role links		agent, source, location, path, goal, time, theme, instrument, partner, patient, pivot, cause, beneficiary, result, reason, purpose, manner, medium, means, setting, initialLocation, finalLocation, distance, amount, attribute

Table 1: Text2story annotation scheme

The harmonization of the different annotation layers using ISO-standards presented us with some mismatches between the three ISOs, which had to be addressed and solved. As an illustration, we present two of those issues.

Concerning markables, while the thematic annotation specifications in ISO 24617-4 foretold that a clause may receive a semantic role, the referential ISO does not stipulate any entity structure for clauses. Our solution to this problem was to mark the event structure corresponding to the verbal predicate of the subordinated clause so that the semantic role link can be set up. Accordingly, in a sentence like *John said that Mary went to Porto* the chunk that is linked to "said" by the semantic role theme is not the whole clause, but only the verb "went", because it has been already associated to an entity structure, more precisely to an event structure, in the temporal layer, contrary to the clause. This solution adopts a Neo-Davidsonian perspective of the relation between events and their arguments and considers that all entities with an event structure annotated in the temporal level correspond to an event argument of a predicate. So, in a Neo-Davidsonian version, the sentence above would have the following logical form: $\exists e^1$ [SAY (e^1) & AGENT (e^1, John) & $\exists e^2$ [GO ($\exists e^2$) & AGENT ($\exists e^2$, Mary) & TO ($\exists e^2$, Porto) & THEME (e^1, e^2)]].

However, some problems are of more difficult resolution. ISO-24617-4 envisages that some adverbial phrases may be attributed the semantic role of manner, like "tightly" in the sentence *The tiny stick was fastened tightly to his wrist* (ISO-24617: 23). Nonetheless, "tightly" in our framework (and in the relevant ISO-standards, for that matter) cannot be marked as any kind of entity structure. We could simply disregard it because it is a modifier, but in some cases manner adverbial phrases are complements (*The child behaved badly*), conveying pertinent information to the story, and, hence, they should be annotated. At this moment, we still have no means to come to grips with this conundrum.

Despite the above-mentioned hurdles, we have been able to conciliate three ISO-standards and produce a consistent and complete multilayer semantic annotation scheme, which not only adequately serves the purpose of our project, but may also contribute to other annotations' schemes.

5 An Annotated Example

In our model, the annotation procedure consists of three stages. Example (1) will serve to illustrate the three stages.

(1) 20/03/2021
Cientistas que estudavam a erupção de um vulcão da Islândia decidiram esta sexta-feira usar a lava expelida da cratera para assar salsichas.
Scientists that were studying the eruption of a volcano of Iceland decided this Friday to use the lava expelled from the crater to roast sausages.

In the first stage, the annotator marks the entity structures of events and times, and, then, the temporal, aspectual and subordination links are established.

EVENTS
e1=*estudavam* class=occurrence type=process pos=verb tense=past aspect=imperfective polarity=pos vform=none mood=none
e2=*erupção* class=occurrence type=process pos=noun tense= none aspect= none polarity= pos vform=none mood=none
e3=*decidiram* class=occurrence type=transition pos=verb tense= past aspect=perfective polarity= pos vform= none mood= none
e4= *usar* class=occurrence type= process pos=verb tense=none aspect=none polarity=pos vform=infinitive mood= none
e5=*expelida* class=occurrence type=transition pos=verb tense=past aspect= perfective polarity= pos vform=participle mood=none
e6=*assar* class=occurrence type=process pos=verb tense=none aspect=none polarity=pos vform= infinitive mood=none

TIME EXPRESSIONS
t1=*20/03/2021* type=date value=20-03-2021 FunctionInDocument= publication time
t2=*esta sexta-feira* type=date value=19-03-2021 AnchorTimeID=t1

TLINK
e2 before e1
e3 is_included e1
e3 is_included t1
e4 after e3
e5 before e3
e6 simultaneous e4

SLINK
e4 intensional e3
e6 intensional e4

In the second stage, the participants are identified, and they are related to each other by objectal links.

PARTICIPANTS
p1=*cientistas que estudavam a erupção de um vulcão na Islândia* lexical head=noun individuation=individual type =per involvement= >1
p2=*que* head=pronoun individuation=individual type =per involvement= >1
p3=*um vulcão da Islândia* head=noun individuation=individual type =per involvement=1
p4=*a lava expelida da cratera* head=noun individuation= mass type =nat involvement=1
p5= *a lava* head=noun individuation=mass type=nat involvement= undef
p6=*a cratera* head=noun individuation= individual type =nat involvement=1
p7=*salsichas* head=noun individuation= individual type =obj involvement=>1

OBJECTAL LINKS
p2 ObjIdentity p1
p5 partOf p3
p6 partOf p3

In the third stage, the annotator connects participants to events by semantic role links.

SEM_ROLE_LINK
p1=agent (e3)
p2=agent (e1)
p3=patient (e2)
p4=instrument (e4)
p5=theme (e5)
p6=initial location (e5)
p7=patient (e6)
e6=purpose (e4)
p1=agent (e4)
p1=agent (e6)
e2=theme (e1)
e4=theme (e3)

After carrying out this manual annotation in the annotation tool BRAT[3], our project's pipeline includes two more modules: the Brat2DRS, which takes the annotation file generated by Brat, parses it, and creates a DRS representation; and the BRAT2Viz, which takes as input the DRS representation, and deploys a web application that produces the visualizations in the form of MSC or KG (Amorim et al., 2021).

6 Conclusion

In this paper, we present an annotation framework for news articles in EP that aims to provide the input for visualization processes. First, we determined what type of information was necessary to account for events and participants in the narratives, and decided that three annotation layers - temporal, referential and thematic - were required. The next step was to decide which tags and links should be used in each layer to fulfill the annotation purposes. Since interoperability is crucial when we talk about semantic resources, three standards ISO 24617-1/4/9 were utilized to create a multilayer semantic annotation scheme. Notwithstanding the fact that these standards are, in fact, themselves three parts of the same standard, when combined, some inconsistencies arise. So, we had to harmonize the three layers, to attain a cohesive annotation framework. Additionally, we sought to balance the amount of information needed to capture the news stories and the load of the annotation process.

Although this model was built to capture the structure of stories in news in EP, its scope is not limited to news nor to EP, as it can be extended to other narrative texts and other languages with some adaptations to deal with genre and language specificities. Moreover, the integration of three different layers in a single annotation framework enables formal semantic representation with DRS, which acts as an intermediate language to generate visualizations in the form of knowledge graphs, for instance.

In the future, we intend to endow our annotation scheme with more granularity. To this end, ISO standard for spatial information (ISO 24617-7) will be added to our framework. For now, spatial annotation has relied on the tags, attributes and links available in the referential and thematic layers. Likewise, a more detailed information regarding quantification of participants and of events is a component to be improved in the future. At this moment, this kind of information has a very simplified representation solely in the referential layer, which does not fully represent the different possibilities of quantification over entities.

Acknowledgments

The authors wish to thank the reviewers for their constructive comments. This research is financed by the ERDF – European Regional Development Fund through the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project PTDC/CCI-COM/31857/2017 (NORTE-01-0145-FEDER-03185). The usual disclaimers apply.

References

Amorim, Evelin; Ribeiro, Alexandre; Cantante, Inês; Jorge, Alípio; Santana, Brenda; Nunes, Sérgio; Silvano, Purificação; Leal, António; & Campos, Ricardo (2021). Brat2Viz: a Tool and Pipeline for Visualizing Narratives from Annotated Texts. In *Text2Story 2021. Fourth International Workshop on Narrative Extraction from Texts*. (pp. 49-56). Lucca, Italy: CEUR Workshop Proceedings, CEUR-WS.org.

Bach, Emmon (1985). The algebra of events. *Linguistics and Philosophy*, *9*, 5–16.

Baker, Collin; Fillmore, Charles; & Lowe, John (1998). The Berkeley FrameNet project. In *Proceedings of the Conference on 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. (pp. 86–90). Montréal, Quebec: Université de Montréal. Retrieved from https://www.aclweb.org/anthology/P98-1013/

Basile, Valerio; Bos, Johan; Evang, Kilian; & Venhuizen, Noortje J. (2012). Developing a large semantically annotated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. (pp. 3196–3200). Istanbul, Turkey: ELRA. Retrieved from https://www.aclweb.org/anthology/L12-1299/

[3] https://nabu.dcc.fc.up.pt/brat/#/examples_demos/paper_ISA-17

Branco, António; Costa, Francisco; Silva, João; Silveira, Sara; Castro, Sérgio; Avelãs, Mariana; Pinto, Clara; & Graça, João (2010). Developing a Deep Linguistic Databank Supporting a Collection of Treebanks: the CINTIL DeepGramBank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (pp. 1810–1815). Valletta, Malta: ELRA. Retrieved from http://www.di.fc.ul.pt/~ahb/pubs/2010BrancoCostaSilvaEtAl.pdf

Branco, António; Carvalheiro, Catarina; Pereira, Sílvia; Avelãs, Mariana; Pinto, Clara; Silveira, Sara; Costa, Francisco; Silva, João; Castro, Sérgio; & Graça, João (2012). A PropBank for Portuguese: The CINTIL-PropBank. In *Proceedings of the Eight International Conference on Language Resources and Evaluation* (pp. 1516–1521). Istanbul, Turkey: ELRA. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/summaries/373.html

Bell, Allan (1991). *The Language of News Media*. Oxford: Blackwell.

Bonet-Jover, Alba; Piad-Morffis, Alejandro; Saquete, Estela; Martínez-Barco, Patricio; & García-Cumbreras, Miguel Ángel (2021). Exploiting discourse structure of traditional digital media to enhance automatic fake news detection. *Expert Systems with Applications, 169,* 1–19. doi: 10.1016/j.eswa.2020.114340

Bos, Johan (2005). Towards wide-coverage semantic interpretation. In *Proceedings of IWCS-6*. (pp. 42–53). Tilburg, The Netherlands. Retrieved from https://www.let.rug.nl/bos/pubs/Bos2005IWCS.pdf

Bos, Johan (2008). Wide-Coverage Semantic Analysis with Boxer. In Johan Bos & Rodolfo Delmonte (eds.). *Semantics in Text Processing. STEP 2008 Conference Proceedings, volume 1 of Research in Computational Semantics*. (pp. 277–286). College Publications.

Bos, Johan; Basile, Valerio; Evang, Kilian; Venhuizen, Noortje J.; & Bjerva, Johannes (2017). The Groningen Meaning Bank. In Nancy Ide & James Pustejovsky (eds.), *Handbook of Linguistic Annotation* (pp. 463–496). USA: Springer. ISBN 978-94-024-0879-9. doi.org/10.1007/978-94-024-0881-2_18

Bunt, Harry (2019). *Plug-ins for content annotation of dialogue acts* . In *Proceedings of the 15th Joint ACL - ISO Workshop on Interoperable Semantic Annotation* (ISA-15) (pp.33–45). Gothenburg, Sweden. Retrieved from https://sigsem.uvt.nl/isa15/ISA-15_proceedings.pdf

Caswell, David; & Dörr, Konstantin (2019). Automating Complex News Stories by Capturing News Events as Data. *Journalism Practice, 13(8),* 951–955. doi.org/10.1080/17512786.2019.1643251

Chiarcos, Christian; Klimek, Bettina; Fäth, Christian; Declerck, Thierry; & McCrae, John P. (2020). On the Linguistic Linked Open Data Infrastructure. In *Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020)* (pp. 8–15). Language Resources and Evaluation Conference (LREC). Marseille, France. Retrieved from https://www.aclweb.org/anthology/2020.iwltp-1.2/

Choubey, Prafulla Kumar; Lee, Aron; Huang, Ruihong; & Wang, Lu (2020). Discourse as a Function of Event: Profiling Discourse Structure in News. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. (pp. 5374–5386). Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.478/

Cinková, Silvie (2006). From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)* (pp. 2170–2175). Genova, Italy: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L06-1058/

Comrie, Bernard (1985). *Tense*. Cambridge: Cambridge University Press.

Costa, Francisco (2012). *Processing Temporal Information in Unstructured Documents*. (Doctoral dissertation, Universidade de Lisboa). Retrieved from https://repositorio.ul.pt/handle/10451/8639

Costa, Francisco; & Branco, António (2010). Temporal information processing of a new language: Fast porting with minimal resources. In *ACL2010—Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 671–677). Uppsala, Sweden. Retrieved from https://www.aclweb.org/anthology/P10-1069/

Costa, Francisco; & Branco, António (2012). Extracting temporal information from Portuguese texts. In Helena Caseli; Aline. Villavicencio; António Teixeira; & Fernando Perdigão (Eds). *Computational Processing of the Portuguese Language. PROPOR 2012. Lecture Notes in Computer Science, vol 7243* (pp. 99–105). Springer, Berlin, Heidelberg. doi.org/10.1007/978-3-642-28885-2_11

Curran, James; Clark, Stephen; & Bos, Johan (2007). Linguistically Motivated Large-Scale NLP with CandC and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* (pp. 33–36). Prague, Czech Republic. Retrieved from https://www.aclweb.org/anthology/P07-2009/

Ehrlinger, Lisa; & Wöß, Wolfram (2016). Towards a definition of knowledge graphs. In: SEMANTiCS (Posters, Demos, SuCCESS), 48, 1-4. http://ceur-ws.org/Vol-1695/paper4.pdf

Fernández-Montraveta, Ana; & Vázquez, Gloria (2014). The SenSem Corpus: an annotated corpus for Spanish and Catalan with information about aspectuality, modality, polarity and factuality. *Corpus Linguistics and Linguistic Theory*, *10 (2),* 273–288. doi.org/10.1515/cllt-2013-0026

Ferro, Lisa; Gerber, Laurie; Mani, Inderjeet; & Wilson, George (2003). *TIDES 2003 standard for the annotation of temporal expressions* (technical report). The MITRE Corporation. Retrieved from https://www.mitre.org/sites/default/files/pdf/ferro_tides.pdf

Filatova, Elena; & Hovy, Eduard (2001). Assigning Time-Stamps to Event-Clauses. In *Proceedings of the ACL-EACL 2001 Workshop on Temporal and Spatial Information Processing* (pp. 88–95). Toulouse: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W01-1313/

Gaizauskas, Robert, & Alrashid, Tarfah. (2019) SceneML: A Proposal for Annotating Scenes in Narrative Text, In *Proceedings of the 15th Joint ACL - ISO Workshop on Interoperable Semantic Annotation* (ISA-15) (pp.13–21), Gothenburg, Sweden. Retrieved from https://sigsem.uvt.nl/isa15/ISA-15_proceedings.pdf

Gerber, Laurie; Ferro, Lisa; Mani, Inderjeet; Sundheim, Beth; Wilson, George; & Kozierok, Robyn (2002). Annotating Temporal Information: From Theory to Practice. In *Proceedings of the 2nd international conference on Human Language Technology Research* (pp. 226–230). San Francisco, CA: Morgan Kaufmann Publishers. Retrieved from https://dl.acm.org/doi/10.5555/1289189.1289202

Gessler, Luke; Peng, Siyao Logan; Liu, Yang; Zhu, Yilun; Behzad, Shabnam; & Zeldes, Amir (2020). AMALGUM - A free, balanced, multilayer English web corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020).* (pp. 5267–5275). Marseille: European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.648/

Goyal, Archana; Vishal, Gupta; & Kumar, Manish (2018). Recent Named Entity Recognition and Classification techniques: A systematic review. *Computer Science Review, 29,* 21–43. doi.org/10.1016/j.cosrev.2018.06.001

Gries, Stefan Th.; & Berez, Andrea L. (2017). Linguistic Annotation in for Corpus Linguistics. The Groningen Meaning Bank. In Nancy Ide & James Pustejovsky (Eds.). *Handbook of Linguistic Annotation* (pp. 379–410). USA: Springer. ISBN 978-94-024-0879-9.

Harel, David; & Thiagarajan, P.S. (2003). Message Sequence Charts. In Luciano Lavagno; Martin Grant; & Bran Selic (Eds.). *UML for Real: Design of Embedded Real-Time Systems* (pp. 77–105). USA: Springer. ISBN 978-0-306-48738-5. https://doi.org/10.1007/0-306-48738-1_4

Hovy, Eduard; Marcus, Mitchell; Palmer, Martha; Ramshaw, Lance; & Weischedel, Ralph (2006) OntoNotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers.* (pp. 57–60). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N06-2015/

Ide, Nancy; Baker, Collin; Fellbaum, Christiane; Fillmore, Charles; & Passonneau, Rebecca (2008). MASC: The manually annotated Sub-Corpus of American English. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008* (pp. 2455-2460). European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/pdf/617_paper.pdf

Ide, Nancy; & Pustejovsky, James (2010). What does interoperability mean, anyway? Toward an operational definition of interoperability. In *Proc. of the 2nd International Conference on Global Interoperability for Language Resources (ICGL)*. Hong Kong, China. Retrieved from https://www.cs.vassar.edu/~ide/papers/ICGL10.pdf

ISO24617-1:2012, Language resource management- Semantic annotation framework (SemAF) - Part 1: Time and events (SemAF-Time, ISO-TimeML)

ISO-24617-4: 2014, Language resource management- Semantic annotation framework (SemAF) - Part 4: Semantic roles (SemAF-SR)

ISO 24617-6: 2016, Language resource management- Semantic annotation framework (SemAF) - Part 6:

Principles of semantic annotation (SemAF Principles)

ISO 24617-7: 2019, Language resource management-Spatial information (SemAF) - Part 7: Reference annotation framework (ISO-Space)

ISO 24617-9: 2019, Language resource management-Semantic annotation framework (SemAF) - Part 9: Reference annotation framework (RAF)

Kamp, Hans; & Uwe Reyle (1993). *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: Kluwer Academic Publishers.

Katz, Graham; & Arosio, Fabrizio (2001). The Annotation of Temporal Information in Natural Language Sentences. In *Proceedings of ACL-EACL 2001, Workshop for Temporal and Spatial Information Processing* (pp. 104–111). Association for Computational Linguistics. Toulouse. Retrieved from https://www.aclweb.org/anthology/W01-1315/

Norambuena, Brian Keith; Horning, Michael; & Mitra, Tanushree (2020). Evaluating the Inverted Pyramid Structure through Automatic 5W1H Extraction and Summarization. *Computation Journalism Symposium,* 1–7. Retrieved from https://par.nsf.gov/biblio/10168974

Nouvel, Damien; Ehrmann, Maud; & Rosset, Sophie (2016). *Named Entities for Computational Linguistics.* ISTE/Wiley, UK/USA.

Palmer, Martha; Gildea, Daniel; & Kingsbury, Paul (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistis, 31 (1),* 71–106. Retrieved from https://www.aclweb.org/anthology/J05-1004/

Petukhova, Volha; & Bunt, Harry (2008). LIRICS semantic role annotation: design and evaluation of a set of data categories. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)* (pp. 39–45). Marrakech, Morocco: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L08-1428/

Pustejovsky, James; Bunt, Harry; & Zaenen, Annie (2017). Designing Annotation Schemes: From Theory to Model. The Groningen Meaning Bank. In Nancy Ide & James Pustejovsky (Eds.). *Handbook of Linguistic Annotation* (pp. 21–72). USA: Springer. ISBN 978-94-024-0879-9.

Pustejovsky, James; & Stubbs, Amber (2012). *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc., USA.

Pustejovsky, James; Castaño, José; Ingria, Robert; Saurí, Roser; Gaizauskas, Robert; Setzer, Andrea; & Katz, Graham (2003a). TimeML: robust specification of event and temporal expressions in text. In *IWCS-5, Fifth International Workshop on Computational Semantics* (pp. 28–34). Retrieved from https://www.aaai.org/Papers/Symposia/Spring/2003/SS-03-07/SS03-07-005.pdf

Pustejovsky, James; Hans, Patrick; Saurí, Roser; See, Andrew; Gaizauskas, Robert; Setzer, Andrea; Radev, Dragomir; Sundheim, Beth; Day, David; Ferro, Lisa; & Lazo, Marcia (2003b). The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics, Lancaster* (pp. 647–656). Retrieved from https://www.researchgate.net/publication/228559081_The_TimeBank_corpus

Rabe, Robert (2008). Inverted Pyramid. In Stephen L. Vaughn (Ed.). *Encyclopedia of American Journalism*. (pp. 223–225). New York: Routledge.

Reichenbach, Hans (1947). *Elements of Symbolic Logic*. New York: Macmillan.

Schuler, Karin (2005). *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon* (Doctoral dissertation, University of Pennsylvania). Retrieved from https://verbs.colorado.edu/~kipper/Papers/dissertation.pdf

Setzer, Andrea (2001). *Temporal Information in Newswire Articles: an Annotation Scheme and Corpus Study* (Doctoral dissertation, University of Sheffield). Retrieved from http://etheses.whiterose.ac.uk/14436/

Setzer, Andrea; & Gaizauskas, Robert (2000a). Annotating events and temporal information in newswire text. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)* (pp. 1287–1293). Athens, Greece: European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/L00-1241/

Setzer, Andrea; & Gaizauskas, Robert (2000b). Building a temporally annotated corpus for information extraction. In *Proceedings of the Information Extraction Meets Corpus Linguistics Workshop at the 2nd International Conference on Language Resources and Evaluation (LREC 2000)* (pp. 9–14). Athens, Greece: European Language Resources Association (ELRA). Retrieved from http://staffwww.dcs.shef.ac.uk/people/R.Gaizauskas/research/papers/lrec00-ie-meets-cl-ter.pdf

Setzer, Andrea; & Gaizauskas, Robert (2001). A pilot study on annotating temporal relations in text. In *ACL 2001 Workshop on Temporal and Spatial Information Processing* (pp. 73–80). Toulouse, France. Retrieved from https://www.aclweb.org/anthology/W01-1311.pdf

Song, Zhiyi; Bies, Ann; Strassel, Stephanie; Ellis, Joe; Mitamura, Teruko; Dang, Hoa Trang; Yamakawa, Yukari; & Holm, Sue (2016). Event Nugget and Event Coreference Annotation. In *Proceedings of the Fourth Workshop on Events*. (pp. 37–45). San Diego, CA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W16-1005/

Stenetorp, Pontus; Pyysalo, Sampo; Topić, Goran: Ohta, Tomoko; Ananiadou, Sophia; & Tsujii, Junı́chi (2012). BRAT: a web-based tool for NLP-assisted text annotation. In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. (pp. 102–107). Avignon, France: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/E12-2021.pdf

Sun, Weiyi; Rumshisky, Anna; & Uzuner, Ozlem (2013). Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics, 46, Supplement,* 5–12. doi: 10.1016/j.jbi.2013.07.004

Talhadas, Rui; Mamede, Nuno; & Baptista, Jorge (2013). Semantic Roles for Portuguese Verbs. In *32nd International Conference on Lexis and Grammar* (pp. 127–132). Faro: Universidade do Algarve. Retrieved from https://www.inesc-id.pt/ficheiros/publicacoes/11312.pdf

Thomson, Elizabeth A.; White, Peter R.R; & Kitley, Philip (2008). "Objectivity" and "hard news" reporting across cultures: comparing the news report in English, French, Japanese and Indonesian journalism. *Journalism Studies*, *9(2),* 212–228. doi.org/10.1080/14616700701848261

Van Dijk, Teun A. (1985). Structures of news in the press. In Teun A. Van Dijk (Ed). *Discourse and Communication: New Approaches to the Analysis of Mass Media Discourse and Communication*. (pp. 69–93). Berlin/New York: Walter de Gruyter.

Zahid, Iqra; Zhang, Hao; Boons, Frank; & Batista-Navarro, Riza (2019). Towards the Automatic Analysis of the Structure of News Stories. In Alípio Jorge; Ricardo Campos; Adam Jatowt & Sumit Bhatia (Eds.). *Proceedings of the Text2StoryIR'19 Workshop* (pp. 71–79). Cologne, Germany. Retrieved from http://ceur-ws.org/Vol-2342/paper9.pdf

Zeldes, Amir (2017). The GUM Corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation, 51(3),* 581–612. doi: 10.1007/s10579-016-9343-x

Zeldes, Amir (2019). Multilayer Corpus Studies. New York and London: Routledge.

Zeldes, Amir; & Simonson, Dan (2016). Different flavors of GUM: Evaluating genre and sentence type effects on multilayer corpus annotation quality. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW X 2016)* (pp. 68–78). Berlin: Association for Computational Linguistics. doi: 10.18653/v1/W16-1709