An annotation scheme for financial news in Portuguese

António Leal
University of Macau
University of Porto
CLUP
antonioleal@um.edu.mo

Purificação Silvano
University of Porto
CLUP
INESC TEC
msilvano@letras.up.pt

Zuo Qinren
University of Porto
up202202310@edu.letras.up.pt

Evelin Amorim
University of Porto
INESC TEC
evelin.f.amorim@inesctec.pt

Alípio Jorge
University of Porto
INESC TEC
amjorge@fcc.up.pt

Abstract

We present an annotation scheme designed to capture information related to the maintenance or change in the price of some goods (fuels, water, and vehicles) in news articles in Portuguese. The methodology we used involved adapting an existing annotation scheme, the Text2Story scheme (Silvano et al., 2021; Leal et al., 2022), which is based on different parts of ISO 24617 to capture the essential information for this project. Adaptations were needed to accommodate specific information, namely, information related to quantitative data and comparative relations that are abundant in this type of news. In this paper, we provide an overview of the annotation scheme, highlighting attributes and values of the entity and link structures specifically designed to capture financial information, as well as some problems we had to overcome in the process of building it and the rationale of some decisions behind its overall architecture.

1 Introduction

Corpora annotation is fundamental for theoretical linguistics research and for faster progress in improving Natural Language Processing and Information Extraction tasks by providing training material and gold standards for model evaluation (e.g., Levi and Shenhav, 2022). In recent years, projects have been carried out to build annotation schemes and annotated corpora to capture the content of texts with more or less generic themes (e.g., Groningen Meaning Bank (GMB) (Basile et al., 2012; Bos et al., 2017); Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008); Georgetown University Multilayer Corpus (GUM) (Zeldes and Simonson, 2016; Zeldes, 2017); for Portuguese, Nunes et al., 2024).

However, several studies have also emerged that seek to capture the content of texts from specific domains, such as the medical domain (cf., e.g., Sun et al., 2013 and Campillos et al., 2018). These studies seek to overcome the particular difficulties of these domains, such as the specialized lexicon, which requires a more detailed ontological categorization of entities or specific links between these entities. The financial domain is one of these domains where efforts to build resources are scarce (cf., e.g., Lee et al., 2022). In fact, there is a significant lack of annotated corpora to train models in this domain, which is not unrelated to the lack of explicitly designed annotation schemes to capture financial information.

The project we present in this article aims to contribute to filling this gap. Our general objective was to create an annotation scheme that captured information related to the maintenance or change in the price of some goods (fuels, water, and vehicles) in European Portuguese (EP) news.

The methodology we used involved adapting an existing annotation scheme, the Text2Story (T2S) scheme, to capture the essential information for this project. The reasons for choosing this methodology are related to the characteristics of the T2S scheme and the objectives of this annotation project, which are to capture the maintenance and variation of prices of some goods. On the one hand, the T2S scheme is a scheme created based on the ISO 24617 standard (Silvano et al., 2021; Leal et al., 2022), which guarantees interoperability, particularly if you want to add in the future distinct semantic layers to the annotation scheme, such as, for example, discourse relations. On the other hand, the T2S scheme was created to capture the narrative structure of news of generalist themes in EP and has already proven suitable for this purpose (Silvano et al., 2023, 2024). Thus, it was predictable that, despite having to make a certain number of changes to the T2S scheme to adapt it to the objectives of this specific project, the T2S scheme should be

Proceedings of the 21st Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-21), pages 63–75

suitable, given that the type of text (news) and the language (EP) would be the same.

In the following section, we present some previous work on annotating financial information in texts and, in particular, on annotating quantities. In the third section, we present the annotation scheme we built for annotating financial news and describe some problems we have encountered during the process. We also present some justifications for specific choices we have made while constructing this annotation scheme. We conclude with some final remarks and future work.

2 Related work

Reasoning based on numerical information is common in all human domains (Thawani et al., 2021), particularly in scientific domains, in which one seeks to deduce scientific facts from premises that, in empirical sciences, are constructed based on quantitative data. For these reasons, in NLP, research on measurement expressions has focused on specific domains (Göpfert et al., 2022). For example, the medical domain is particularly interesting, as electronic health records contain much quantitative information (e.g., blood test results) relevant, for instance, to including patients in clinical trials. However, there are other interesting domains, such as the financial domain. In this field, research has focused more on the value of company shares and crude oil and, to a lesser extent, on the factors underlying the price variation of certain commodities (Lee et al., 2022).

Extracting these quantification expressions from texts in natural languages faces several challenges, starting with the format of these expressions (Göpfert et al., 2022). They are typically formed with a numerical value and a unit (which may or may not correspond to a standardized value). However, the numerical value can be presented in different ways (with letters or numbers); it can correspond to some vague form of quantification (e.g., 'a couple') or denote a numerical range (e.g., '3-5 g.'). The numerical value can correspond, for example, to a percentage relative to another quantification expression, which may be implicit in the text (e.g., 'The price of gasoline increased by 4%').

Quantification expressions may also contain some modification, which alters their overall meaning (cf. '3 g.' vs. 'approx. 3 g.'). One problem found in some works related to the annotation of numerical expressions (e.g., Ning et al., 2022) is the fact that they only annotate cardinal quantifiers, leaving aside modifiers of these quantifiers: for example, in "at most 3%", they only annotate "3%" leaving aside the modifier "at most"; the same happens in "at least 2%". Leaving aside the modifiers, they lose part of the nominal expressions' semantics. In this specific case, the distinction between right monotone increasing determiners ('at least') and right monotone decreasing determiners ('at most')[1] is lost (cf., e.g., Barwise and Cooper, 1981; Partee et al., 2012).

It is not just the identification of quantification expressions that is problematic. Indeed, identifying how these quantification expressions measure entities or properties of entities (explicit or implicit in the texts) is also a problem in Information Extraction. For example, in 'gasoline costs 1.5 euros', the quantification expression is locating a characteristic of the entity denoted by "gasoline", its price per liter, on a numerical scale associated with a particular monetary unit. Another problem is identifying price changes over time, as in 'Gasoline costs 4 cents more than last week'. All these problems are further aggravated by a lack of uniformity in the terms used: in different projects, the same concept may correspond to different terms; conversely, the same term may have non-equivalent definitions in other projects (cf. Göpfert et al., 2022). In this regard, the work carried out under ISO 24617, parts 11 and 12, seeks to answer some of these questions. ISO 24617-12 (ISO-24617-12, 2024) is a proposal to deal with quantification issues mainly related to the quantificational information of nominal expressions but also of eventualities expressions and the scope relations in which these expressions are involved. As for ISO 24617-11 (ISO-24617-11, 2021), it deals with aspects related to measurable quantitative information, that is, with the standardization of expressions involving a quantity n and a unit u, as in 'two meters'. Also noteworthy is the work of Abzianidze and Bos (2017), who propose a semantic tagset for use in a new NLP task to encode information that better characterizes lexical semantics than POS tags[2]. These tags describe the

[1] The following examples illustrate the distinct inferences of the two semantic types:
(i) right monotone increasing determiners (e.g., 'at least n')
If at least two men walk fast, then at least two men walk
(ii) right monotone decreasing determiners (e.g., 'at most n')
If at most two men walk, then at most two men walk fast.

[2] We thank one of the anonymous reviewers for bringing

semantic contribution of particular words or punctuation marks concerning the meaning of the whole expression and are grouped into 13 meta-tags. For example, the ATT (attribute) meta-tag includes the QUC (quantity of concrete) tag, which applies to words such as 'two', and the NAM (named entity) meta-tag includes the UOM (unit of measurement) tag, which applies to words such as 'euro' or 'percent'.

Datasets available to train models specifically with quantitative information are not abundant, particularly full-text datasets. Quantitative information is particularly relevant in texts from specialized domains, so these domains, such as medicine or finance, are necessary for this type of annotation. However, the question of finding suitable annotators for this type of task is also pertinent: annotators must have grammatical knowledge and, in some cases, knowledge of the specific domain (for instance, in cases of quantities given relative to a standard; cf. Göpfert et al., 2022). Projects have sought to use LLMs as annotators to overcome this difficulty, given that these models have proven effective in annotating general domain datasets. Aguda et al. (2024) show that LLMs are a possible alternative to non-expert crowd workers for domain-specific tasks. However, they do not surpass domain-specific human experts.

Another problem with datasets in the financial domain is that they often consist only of news headlines, and the annotation aims to perform sentiment analysis on stock prices. There are a few publicly available datasets with information about commodity news. Sinha and Khandait (2021) present one of these datasets, which is a dataset consisting of 11,412 news headlines about gold. This dataset was manually annotated with various information, namely whether the price is attributed to the past or the future ("Past/Future Price Information") and whether this value corresponds to an increasing, decreasing, or maintenance trend ("Price Up/Constant/Down").

Regarding news annotation work on commodities, Lee et al. (2022) report that datasets of this type are scarce, and the authors assume that their dataset may be the only annotated corpus. This is a dataset of 425 news articles about crude oil in English, manually annotated by undergraduate students from a School of Business. This dataset has some similarities to the one we developed in our project. The dataset from Lee et al. (2022) contains information about Entities (divided into 21 types) and Events (triggers, argument roles, and properties such as polarity, modality, and intensity). However, this annotation scheme has several problems. For example, in the list of Entity types, we can find ontologically very distinct entities, such as "commodities" (oil), "dates" (1998), "locations" (Europe), "Money" (USD 50), "percent" (25%), "Price unit" (USD 58 per barrel), "Quantity" (18 million tonnes) and "Production Unit" (29 million barrels per day). Furthermore, there is no unified treatment in the annotation of quantification expressions. Regarding Events, these authors identify 18 types, ranging from events that describe the change in price (e.g., CAUSED-MOVEMENT DOWN-LOSS; MOVEMENT-DOWN-LOSS; MOVEMENT-FLAT) to events that indicate the cause for the price change (e.g., CRISIS; EMBARGO). However, the strictly linguistic annotation is significantly reduced: polarity (POSITIVE and NEGATIVE); modality (ASSERTED and OTHER); and intensity (NEUTRAL, INTENSIFIED, and EASED). All in all, this annotation scheme is similar to our scheme in that it contains events and participants. However, the scheme we have developed is richer from the point of view of the grammatical information provided and the information contained in the annotation of quantification expressions, in addition to including different types of links, namely, semantic function links.

To sum up, identifying quantification expressions is a complex task, but one that is fundamental in several areas of research (cf. Göpfert et al., 2022 for a survey on measurement extraction in NLP tasks). In our project, we are particularly interested in extracting the value attributed to specific products over time: fuels (gasoline, diesel), motor vehicles, and essential goods (water, electricity). We will present a corpus of newspaper articles about finances, given that these texts often present numeric information representing different aspects of the situations they describe: in the case that interests us, the price of commodities (Roy et al., 2015).

3 Description of the scheme for annotating commodities price information

The scheme we created includes the following:

(i) a series of entity structures with information

this proposal to our attention.

about events, the participants in these events, and temporal and measurement expressions;

(ii) a series of link structures: between entities of a temporal nature, such as events and temporal expressions (TLinks); between participants, quantification expressions, and events (semantic roles); between participants (OLinks); and between quantification expressions and participants. All (entity and link) structures are composed of several attributes and values. The scheme also includes a function associating each document with its publication time. See the Appendix A for an overview of the annotation scheme.

Most of the content of this annotation scheme is based on parts of the ISO standard ISO24617, Language Resource Management - Semantic annotation framework: Part 1: Time and events (ISO-24617-1, 2012); Part 4: Semantic roles (ISO-24617-4, 2014); Part 7: Spatial information (ISO-24617-7, 2020); and Part 9: Reference annotation framework (ISO-24617-9, 2019). However, it was necessary to create some attributes and links to capture the information relevant to this annotation project, which will be detailed below.

Although a part of ISO deals with quantification, ISO 24617 - Part 12: Quantification (ISO-24617-12, 2024), we decided not to include it in this project. There are several reasons for this. For example, the expressive power of ISO 24617 - 12, notably in the annotation of quantificational information of nominal expressions and in the scope relations between nominal expressions and between these expressions and eventualities or other operators, such as negation, would not be used because this kind of information does not appear in the news that constitute our corpus. So, although ISO 24671-12 is quite expressive, our project did not require the level of detail proposed in this part of the ISO. We argue that the simplicity of the annotation scheme is crucial if one wants to recruit annotators without specialized linguistic knowledge.

Using ISO 24617-12 also poses some problems. One is that this part of the standard does not cover some of the expressions we had to annotate, namely, the quantification expressions of the "non-exact" type. Another problem is that many of the eventualities in the annotated news are expressed by nominal expressions with adjectives expressing quantification. Neither ISO 24617-12, for quantification, nor ISO 24617-9, for referential annotation, suggests ways to deal with these adjectival expressions that express some form of quantification over eventualities expressed by nouns. Examples (1) and (2) illustrate these cases. These sentences contain nominal expressions that, semantically, are equivalent to the sentences in (3) and (4), with explicit quantification within the verb phrase.

(1) *A gasolina registou um aumento expressivo.*
'Gasoline registered a significant increase.'

(2) *A gasolina registou uma descida ligeira.*
'Gasoline registered a slight drop.'

(3) *A gasolina aumentou expressivamente.*
'Gasoline prices have increased significantly.'

(4) *A gasolina desceu ligeiramente.*
'Gasoline prices have fallen slightly.'

Although ISO 24617-12 proposes a scheme for annotating quantified nominal expressions that could potentially apply to the annotation we intend to make (cf., for example, the @involvement attribute of entities, i.e., the information about the number of entities or part of the entity involved in a particular eventuality[3]), this proposal is not effectively applicable directly to the data we had to annotate, as in (5), but to constructions of another type, as in (6), which do not exist in the news that makes up our corpus.

(5) *A gasolina custa 1,5 euros.*
'Gasoline costs 1.5 euros.'

(6) *Comprei 3 euros/ dois litros de gasolina.*
'I bought 3 euros/two liters of gasoline.'

Finally, ISO 24617-11 (ISO-24617-11, 2021) was also assessed in this annotation project. However, we considered that its use could be dispensable, which is why we chose not to include it. For example, ISO 24617-11 proposes using the measure link and the comparison link. As for the measure link, which connects an entity to a measure, its use would unnecessarily complicate our scheme, given that it overlaps with the semantic roles. The comparison link, which is similar to the ARG1 and

[3]The value of the @involvement attribute indicates how many/much or which fraction of the reference domain is contained in the participant set (ISO-24617-12, 2024).

ARG2 that we propose, is a link that is established between two measures (according to ISO 24617-11). This type of link is irrelevant to our annotation, given that the comparison relations in the news are not between measures but between entities.

The news annotations were performed in BRAT (Stenetorp et al., 2012). The manual was built incrementally, following the MAMA cycle (Pustejovsky et al., 2017), taking the Text2Story scheme as a starting point. A PhD student in linguistics annotated a set of news items and identified problems in the annotation process. In a meeting with two senior linguistics researchers, these problems were discussed, and solutions were proposed. The student revised the annotation according to these solutions, and in the following meeting, the new results were analyzed. If the problems persisted, new solutions were proposed. If the problems were solved, the student annotated a new set of news items. This process was repeated cyclically until the entire set of news items was annotated, and no problems were left in the annotation. Finally, the news annotations (N=98) were reviewed by both senior linguistics investigators to identify and correct any lapses or inconsistencies that might persist.

3.1 Overview of the annotation scheme

1. Events

The markables of events follow the same rules defined in ISO 24617-1, the basis for the T2S annotation scheme. Regarding the attributes, it was decided to simplify them. Thus, the @class attribute was eliminated, as it was not relevant to the objectives of this annotation project. The @aspect, @vform, and @mood attributes were included in the @tense attribute. This attribute was adapted to capture the verb forms in Portuguese texts, thus making the annotation task less complex and easier to implement. Table 2 in the appendix presents the values used in the @tense attribute. The @pos attribute was limited to just two values (noun and verb). The @polarity attribute was maintained. Finally, in the case of the @movement attribute, it was decided to simplify the annotation to two values, "upward" and "downward," to capture the rise or fall of prices in the case of events or progressive states. Example (7) illustrates the case of price increase (upward movement).

(7) *A gasolina subiu esta semana. / A gasolina está a subir esta semana.*

'Gasoline rose this week. / Gasoline is rising this week.'

2. Time expressions

The time expressions in this annotation project follow the annotation proposed in T2S, which in turn abides by the rules of ISO-24617-1 (2012) proposal. The tag spans are the same, as are the @type attributes (with the values "date," "time," "duration," and "set") and the "publication time".

3. Participants

Regarding participants, this annotation project follows the T2S proposal (based on ISO-24617-9, 2019), with few adaptations, only those necessary to capture some specificities of news about price variations (recall that the T2S scheme was created to capture the structure of news about narratives of generalist themes, which is why it lacks specific tags for entities in the financial domain). Thus, the tag spans and the annotation of @lexical-head, @individuation, and @involvement remained the same, as did the objectal relations between participants. In the case of the @type attribute, some values were added to capture specific information. Therefore, the value "relation-price/unit" was included for referents corresponding to an abstract numerical relation between a measuring unit and a value in some monetary system. This value is subdivided into "average-value" (when the relation corresponds to an average) and "precise-value" (when the relation corresponds to an exact value). The relevant measuring unit is indicated in the text box (in Brat). Examples (8) and (9) illustrate these two possibilities.

(8) *O preço de referência do litro de gasolina em Portugal é actualmente de 1,741 euros enquanto o do gasóleo vale 1,521 euros.*

'The reference price of a liter of gasoline in Portugal is currently 1.741 euros, while that of diesel is 1.521 euros.'

(relation-price/unit-precise-value; unit=liter)

(9) *O preço médio do gasóleo na quarta-feira era de 1.410 euros.*

'The average price of diesel on Wednesday was 1,410 euros.'

(relation-price/unit-average-value; unit=liter)

The value "TARIFA" (TAR) is also included for cases where the referent is a table (or a range of that table) of rates charged for a given service. Example (10) is one of those cases.

(10) *O primeiro escalão dos consumidores de água teve uma descida de 3%. (TAR)*

'The first tier of water consumers saw a 3% drop.'

Table 3 in the appendix summarizes all of the @type values used in the annotation.

4. Semantic roles

The semantic roles used in this annotation project are a subset of those proposed in ISO-24617-4 (2014), with definitions explicitly oriented to the financial domain (cf. Table 4 in the appendix). Examples (11) and (12) are instances of this annotation layer.

(11) *Ao mesmo tempo, o crude apreciava 1,31% para 92,87 euros.*

'At the same time, crude oil appreciated by 1.31% to 92.87 euros.'

(92,87 euros — Goal — apreciava; O crude – Theme — apreciava; 1,31% — Amount — apreciava)

(12) *O gasóleo permanece em 1,22 euros por litro.*

'Diesel remains at R$1.22 per liter.'

(O gasóleo — Pivot — permanece; 1,22 euros — Attribute — permanece)

5. Quantification expressions

One of the objectives of this project is to capture information related to changes in the prices of goods. Therefore, annotating each expression that conveys quantitative information related to prices is essential (see Figure 2 for an overview of the occurrence of entity structures in the annotated corpus). In this annotation scheme, these expressions are annotated in entity structures called "quantification expressions". These expressions appear in the news in very different formats. However, they can be grouped into two groups: those of the "exact" type, providing quantitative information that is expressed, in some way, by a numerical value, and those of the "non-exact" type, whose quantitative information corresponds to a non-numerical value and is typically expressed by an adjective or an adverb (see Table 1 for the distribution of quantification expressions in the annotated corpus).

From a linguistic point of view, expressions of the type "exact", often referred to as "measurement phrases", correspond, in most cases, to argumental nominal expressions in the oblique case (cf. Gonçalves and Raposo, 2013). The verbs with which these expressions occur are varied: (i) verbs such as *custar* 'to cost' (which, like *durar* 'to last', *medir* 'to measure' or *pesar* 'to weigh', express the value of physical or abstract entities on a quantitative scale); (ii) verbs of movement (which, in this case, have fictive readings; cf. Talmy, 1996), such as *subir* 'to go up' and *descer* 'to go down' and their synonyms; (iii) verbs that denote variation of properties in general, such as *aumentar* 'to increase' or *reduzir* 'to decrease'; (iv) verbs lexically specialized in price variation, such as *encarecer* 'to make more expensive' or *custar* 'to cost'; (v) eventive verbs that are not lexically specified as to the nature of the variation, such as *evoluir* 'to evolve' or *encerrar* 'to close', which lexically encode only some type of change. These measurement phrases can also occur as predicatives with copula verbs or verbs that proceed to some static location in cases where no variation in price is expressed, but an indication is given of the value associated with the good in a given time interval (e.g., *ser* 'to be,' *manter-se* 'to maintain,' and *situar-se* 'to be located'). Since these "exact" quantification expressions are included in a more general group of expressions associated with measurement verbs, such as *medir* 'to measure,' we annotated them according to the ISO-24617-7 (2020) proposal, using the "measure" tag. Thus, quantification expressions have the attributes @value, @unit, and @modifier. The @value attribute provides the indication of the quantity of the entity, denoted by cardinal numeral quantifiers or other quantifiers; the @unit corresponds to the monetary unit used in the quantification expression; the @modifier corresponds to an expression that modifies the quantification in terms of quantitative, circumstantial or temporal information. Example (13) illustrates this part of the annotation.

(13) *Os combustíveis já aumentaram, entre 1 de janeiro e 1 de março, cerca de nove cêntimos.*

'Fuel prices have already increased by around nine cents between January 1st and March 1st.'

(@modifier - cerca de; @value - 0,09; @unit – euro)

Moving on to "non-exact" quantification expressions, they correspond to seven subtypes, which are syntactically very diverse and distinct from the

type described previously. Therefore, creating specific attributes for these expressions was necessary, which are not provided in any part of ISO 24617. In what follows, we describe those attributes.

- *value-minimum*: when it corresponds to a minimum price value in the time interval considered in the sentence. Typically, it corresponds to a nominal expression with an adjective in the superlative degree.

  (14) *O preço do barril de Brent, o petróleo que serve de referência ao mercado português, desvalorizou 4,57% para 37,93 dólares, o que representa* ***o valor mais baixo*** *desde dezembro de 2008.*
  'The price of a barrel of Brent, the oil used as a reference for the Portuguese market, fell 4.57% to 37.93 dollars, representing the lowest value since December 2008.'

- *value-maximum*: when it corresponds to a maximum price value in the time interval considered in the sentence. Typically, it corresponds to a nominal expression with an adjective in the superlative degree.

  (15) *Em maio, a matéria-prima superou a fasquia dos 80 dólares por barril,* ***o valor mais alto*** *desde novembro de 2014.*
  'In May, the raw material surpassed the US$ 80 per barrel mark, the highest value since November 2014.'

- *value-much*: when it corresponds to a high differential price value in the time interval considered in the sentence. Typically, it corresponds to an adjective that denotes an interval in a scale whose extension is greater than a reference value (which, in most cases, is implicit).

  (16) *A gasolina registou um aumento* ***expressivo***.
  'Gasoline registered a significant increase.'

- *value-low*: when it corresponds to a low differential price value in the time interval considered in the sentence. Typically, it corresponds to an adjective that denotes an interval in a scale whose extension is smaller than a reference value (which, in most cases, is implicit).

  (17) *A gasolina registou uma descida* ***ligeira***.
  'Gasoline registered a slight drop.'

- *value-comparative-superiority*: when a differential value of superiority of the price is expressed relative to the price of another entity referred to in the sentence. It corresponds to an adjective in the comparative degree.

  (18) *A gasolina é* ***mais cara do que*** *o gasóleo.*
  'Gasoline is more expensive than diesel.'

- *value-comparative-inferiority*: when a differential value of inferiority of the price is expressed relative to the price of another entity referred to in the sentence. It corresponds to an adjective in the comparative degree.

  (19) *A gasolina é* ***menos cara do que*** *o gasóleo.*
  'Gasoline is less expensive than diesel.'

- *value-comparative-equality*: when a value of equality of the price is expressed relative to the price of another entity referred to in the sentence. It corresponds to an adjective in the comparative degree.

  (20) *A gasolina é* ***tão cara como*** *o gasóleo.*
  'Gasoline is as expensive as diesel.'

As with "exact" quantification expressions, "non-exact" quantification expressions may be subject to some modification. For this reason, these expressions can also be annotated with the @modifier attribute, as exemplified in (21).

(21) *Este combustível continuará a ter um preço de venda média muito perto de máximos de agosto de 2015.*

'This fuel will continue to have an average selling price very close to the highs of August 2015.'

(máximos – quantification expression; non-exact; value-maximum; @modifier - muito perto de)

Quantification expressions that express non-exact values of the Value-Comparative type (superiority/inferiority/equality) establish a relationship between two entities that share the same scalar property, in this case, the price. Therefore, these three quantification expressions trigger comparison relationships. Since these relations are not provided

for in ISO 24617, they were created specifically for this work, seeking to follow the same general principles of this standard. Thus, these comparison relations link quantification expressions of Value-comparative and the respective compared entities (identified through the links "argument 1" (Arg1) and "argument 2" (Arg2), which correspond to the first and second terms of comparison, respectively). The relation is established from the quantification expression to the entity that plays the role of Arg1 or Arg2. Example (22) shows how this link is used.

(22) *A gasolina é mais cara do que o gasóleo.*

'Gasoline is as expensive as diesel.'

mais cara do que – quantification expression; value-comparative-superiority

a gasolina – Argument 1 – mais cara do que

o gasóleo – Argument 2 – mais cara do que

Example (23) illustrates how this annotation scheme was implemented (see also Figure 1).

(23) *Segundo o presidente da Anarec, Augusto Cymbron, o preço do gasóleo passará dos actuais 1,339 euros para os 1,369 euros.*

'According to the president of Anarec, Augusto Cymbron, the price of diesel will increase from the current R$1.339 to R$1.369'

In this example, the change in the price of diesel is captured. The participant of the "relation price-unit" type is expressed by the noun phrase *o preço* (the price), which denotes a count entity. The substance is represented by the noun phrase *o diesel*, a participant that is a mass-type "object." The relationship between these two participants is captured by the objectal link "partOf." The price change event is annotated in the markable *passará* (will pass): it is an event expressed by a verb in the future tense. It encodes a price increase (an upward movement on the numerical scale associated with the price). The expressions *1.339 euros* and *1.369 euros* correspond to participants of the "quantification expression" type: both correspond to exact information, and their unit of measurement is "euro," varying only in their respective measurement values. Semantic roles link these entity structures. Thus, *o preço* (the price) is connected to *passará* (will pass) by the SR Theme, indicating that it is the entity that suffers a change of state during the event (in this case, it is the entity whose position on the numerical scale associated with the price changes). In turn, the quantification expressions *1.339 euros* and *1.369 euros* are linked to *passará* by the SR Source and Goal, respectively: the first expression denotes the starting point of the price change, while the second denotes the end point of the price change.

Category	Count	Perc. (%)
Exact	422	78.00
Non-Exact	119	22.00
Max	40	7.39
Much	32	5.91
Min	19	3.51
Low	10	1.85
Comparative-superiority	8	1.48
Comparative-equality	6	1.11
Comparative-inferiority	3	0.55
Total	541	100.00

Table 1: Distribution of Quantification Categories

3.2 Building the annotation scheme: some problems

Throughout the process of building this annotation scheme, we encountered several problems that had to be overcome, from the simplest problems found in any annotation project (for example, the correct identification of the @type of a nominal expression, which is referentially ambiguous), to the more complex problems this particular type of text poses. In fact, the variation in expressions that indicate numerical values and the variation in expressions of eventualities that denote price variation or price maintenance are challenging if one wants to develop a comprehensive but straightforward annotation scheme. We will discuss some of those cases in this section.

The expression denoting a fuel's "price/liter" ratio, such as gasoline, may appear in the following formats (non-exhaustive list).

(24) *O preço do litro da gasolina / O preço da gasolina / O litro da gasolina / O preço do litro no caso da gasolina / A gasolina / O valor da unidade de gasolina está a 1,60 euros.*

'The price of a liter of gasoline / The price of gasoline / The liter of gasoline / The price of a liter in the case of gasoline / Gasoline / The value of the gasoline unit is R$1.60.'

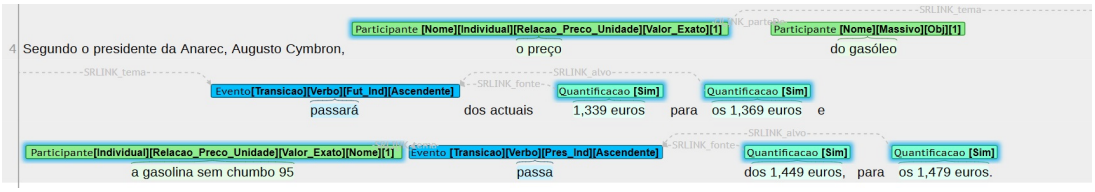

Figure 1: (23): an example of annotation in BRAT (fuel, news 22)

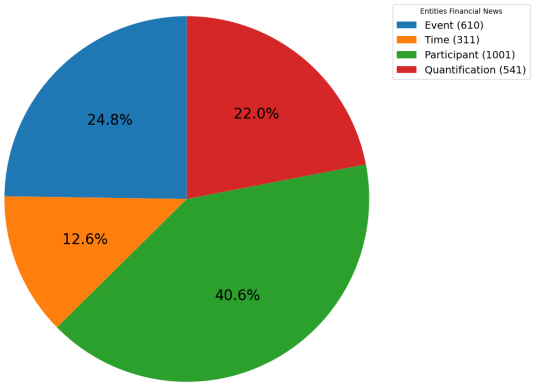

Figure 2: Proportion of Entities in Dataset

(25) *A inflação colocou o custo do litro da gasolina nos 1,60 euros/ colocou o custo do litro nos 1,60 euros no caso da gasolina.*

'Inflation put the cost of a liter of gasoline at R$1.60/ put the cost of a liter at R$1.60 in the case of gasoline.'

Another problem related to the different manners in which the price variation can be expressed. In all of the examples (26)-(29), the event corresponds to a decrease in price, but in example (26), the markable is of the nominal type (*descida*), while in the others ((27)-(29)), the markables are of the verbal type. In this second set of examples, the decrease can be encoded lexically by the verb (such as *cair* 'to fall,' in (27), in a case of fictive motion), by a semi-lexicalized expression (such as *evoluir em baixa* 'to evolve downwards', in (28)), or compositionally, as in (29), through the combination of the verb *atingir* 'to reach' and the direct object that contains a word that corresponds to a quantification expression of the type "non-exact-value-minimum" (*mínimos deste ano* 'minimums of this year').

(26) *A variação do euro face ao dólar deverá determinar uma descida de 1,5 cêntimos no preço de venda da gasolina e de 2 cêntimos no gasóleo.*

'The variation of the euro against the dollar should determine a drop of 1.5 cents in the sale price of gasoline and 2 cents in diesel.'

(27) *O preço do gasóleo deverá cair 2 cêntimos, colocando o custo do litro nos 1,399 euros.*

'The diesel price is expected to fall by 2 cents, putting the cost per liter at 1.399 euros.

(28) *O brent evoluiu hoje em baixa, depois de ter estado a subir durante várias sessões.*

'Brent fell today after rising for several sessions.'

(29) *O brent atingiu mínimos deste ano.*

'Brent reached its lowest price this year.'

4 Concluding remarks

The annotation of information about quantities is a topic that still requires further study, both in terms of annotation schemes that adequately capture the various nuances that quantification expressions can have in natural languages and in terms of annotated corpora that can be used to train models or as gold standards in evaluation tasks. In this article, we seek to contribute to the discussion on annotation schemes and present a new scheme to capture information related to the rise, fall, or maintenance of commodities prices in news articles. To that end, we propose an extension of the Text2Story scheme, which, in turn, was built by combining four parts of ISO 24617 (parts 1, 4, 7, and 9). The new scheme contains new attributes and values specially designed to capture the quantificational information typical of news in the financial domain.

This new scheme can also be extended to capture other types of information. In the future, we plan to include other ISO resources, such as the entire semantic role array of ISO 24617-4, and other parts, such as ISO 24617–8 (ISO, 2016), using the discourse relations framework to analyze the underlying reasons for price changes. Another line of research is this scheme's application (with the necessary adaptations) to other domains with

abundant quantificational information, such as electronic health records. Finally, formalizing the semantics of the proposed annotation scheme is still missing.

References

Abzianidze, L. and Bos, J. (2017). Towards universal semantic tagging. In Gardent, C. and Retoré, C., editors, *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Short papers*.

Aguda, T., Siddagangappa, S., Kochkina, E., Kaur, S., Wang, D., Smiley, C., and Shah, S. (2024). Large language models as financial data annotators: A study on effectiveness and efficiency. *arXiv preprint arXiv:2403.18152*.

Barwise, J. and Cooper, R. (1981). Generalized quantifiers and natural language. In *Philosophy, language, and artificial intelligence: Resources for processing natural language*, pages 241–301. Springer.

Basile, V., Bos, J., Evang, K., and Venhuizen, N. (2012). Developing a large semantically annotated corpus. In *LREC 2012, Eighth International Conference on Language Resources and Evaluation*.

Bos, J., Basile, V., Evang, K., Venhuizen, N. J., and Bjerva, J. (2017). The groningen meaning bank. *Handbook of linguistic annotation*, pages 463–496.

Campillos, L., Deléger, L., Grouin, C., Hamon, T., Ligozat, A.-L., and Névéol, A. (2018). A french clinical corpus with comprehensive semantic annotations: development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources and Evaluation*, 52:571–601.

Gonçalves, A. and Raposo, E. P. (2013). Verbo e sintagma verbal. In Raposo, E. P., Mota, M. A., Segura, L., and Mendes, A., editors, *Gramática do Português*, pages 1155–1220. FCG, Lisboa.

Göpfert, J., Kuckertz, P., Weinand, J., Kotzur, L., and Stolten, D. (2022). Measurement extraction with natural language processing: a review. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2191–2215.

Ide, N., Baker, C., Fellbaum, C., Fillmore, C., and Passonneau, R. (2008). Masc: The manually annotated sub-corpus of american english. In *6th international conference on language resources and evaluation, LREC 2008*, pages 2455–2460. European Language Resources Association (ELRA).

ISO (2016). ISO 24617-8. 2016. Language resource management, part 8: Semantic relations in discourse (DR-Core). Standard, International Organization for Standardization, Geneva, CH.

ISO-24617-1 (2012). Language resource management - semantic annotation framework (semaf) - part 1: Time and events (semaf-time, iso-timeml). Standard, Geneva, CH.

ISO-24617-11 (2021). Language resource management-semantic annotation framework (semaf) - part 11: Measurable quantitative information. Standard, Geneva, CH.

ISO-24617-12 (2024). Language resource management-semantic annotation framework (semaf) - part 12: Quantification. Standard, Geneva, CH.

ISO-24617-4 (2014). Language resource management-semantic annotation framework (semaf) - part 4: Semantic roles (semaf-sr). Standard, Geneva, CH.

ISO-24617-7 (2020). Language resource management-semantic annotation framework (semaf) - part 7: Spatial information. Standard, Geneva, CH.

ISO-24617-9 (2019). Language resource management-semantic annotation framework (semaf) - part 9: Reference annotation framework (raf). Standard, Geneva, CH.

Leal, A., Silvano, P., Amorim, E., Cantante, I., Silva, F., Jorge, A. M., and Campos, R. (2022). The place of iso-space in text2story multilayer annotation scheme. In *Proceedings of the 18th Joint ACL-ISO Workshop on Interoperable Semantic Annotation within LREC2022*, pages 61–70.

Lee, M., Soon, L.-K., Siew, E.-G., and Sugianto, L. F. (2022). Crudeoilnews: An annotated crude oil news corpus for event extraction. *arXiv preprint arXiv:2204.03871*.

Levi, E. and Shenhav, S. R. (2022). A decomposition-based approach for evaluating inter-annotator disagreement in narrative analysis. *arXiv preprint arXiv:2206.05446*.

Ning, Q., Zhou, B., Wu, H., Peng, H., Fan, C., and Gardner, M. (2022). A meta-framework for spatiotemporal quantity extraction from text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2736–2749.

Nunes, S., Jorge, A. M., Amorim, E., Sousa, H., Leal, A., Silvano, P. M., Cantante, I., and Campos, R. (2024). Text2story lusa: A dataset for narrative analysis in european portuguese news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15773–15782.

Partee, B. B., Ter Meulen, A. G., and Wall, R. (2012). *Mathematical methods in linguistics*, volume 30. Springer Science & Business Media.

Pustejovsky, J., Bunt, H., and Zaenen, A. (2017). Designing annotation schemes: From theory to model. In Ide, N. and Pustejovsky, J., editors, *Handbook of Linguistic Annotation*, pages 31–63. Springer, Dordrecht.

Roy, S., Vieira, T., and Roth, D. (2015). Reasoning about quantities in natural language. *Transactions of the Association for Computational Linguistics*, 3:1–13.

Silvano, M. d. P., Amorim, E., Leal, A., Cantante, I., Jorge, A., Campos, R., and Yu, N. (2024). Untangling a web of temporal relations in news articles. In *Proceedings of Text2Story 2024-seventh workshop on narrative extraction from texts*.

Silvano, M. d. P., Amorim, E., Leal, A., Cantante, I., Silva, M. d. F. H. d., Jorge, A., Campos, R., and Nunes, S. S. (2023). Annotation and visualisation of reporting events in textual narratives. In *Proceedings of Text2Story 2023: Sixth Workshop on Narrative Extraction From Texts*.

Silvano, P., Leal, A., Silva, F., Cantante, I., Oliveira, F., and Jorge, A. M. (2021). Developing a multilayer semantic annotation scheme based on iso standards for the visualization of a newswire corpus. In *Proceedings of the 17th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 1–13.

Sinha, A. and Khandait, T. (2021). Impact of news on the commodity market: Dataset and results. In *Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 2*, pages 589–601. Springer.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (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*, pages 102–107.

Sun, W., Rumshisky, A., and Uzuner, O. (2013). Annotating temporal information in clinical narratives. *Journal of biomedical informatics*, 46:S5–S12.

Talmy, L. (1996). Fictive motion in language and ception.

Thawani, A., Pujara, J., Szekely, P. A., and Ilievski, F. (2021). Representing numbers in nlp: a survey and a vision. *arXiv preprint arXiv:2103.13136*.

Zeldes, A. (2017). The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Zeldes, A. and Simonson, D. (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)*, pages 68–78.

A Annotation Scheme and Descriptions

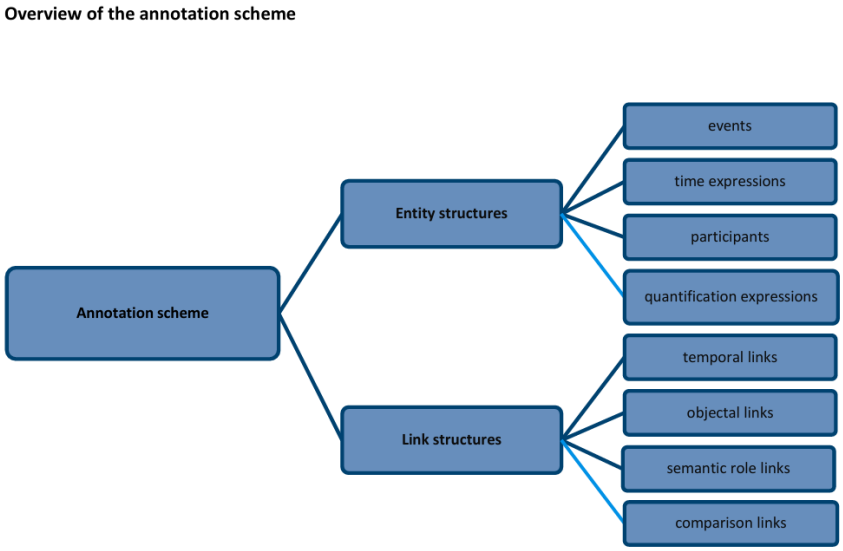

Figure 3: Overview of the Annotation Scheme for Finance news

Indicativo	Conjuntivo	Progressivo	Gerúndio	Infinitivo	IR + Infinitivo
Presente do Indicativo – Pres-Ind	Presente do Conjuntivo – Pres-Conj	Presente progressivo (está a + inf) – PresPro	Gerúndio Simples – GS	Infinitivo Simples – INF-S	IR (presente) + infinitivo (futuro) – ir(Pres)+INF-S
Pretérito Perfeito Simples – PP-Ind	Pretérito Imperfeito – PIMP-Conj	Passado progressivo (esteve/estava a + inf) – PstPro	Gerúndio Composto – GC	Infinitivo Composto – INF-C	IR (futuro) + infinitivo (futuro) – ir(Fut)+INF-S
Pretérito Imperfeito – PIMP-Ind	Pretérito Perfeito – PPC-Conj				
Pretérito Perfeito Composto – PPC-Ind	Pretérito mais-que-perfeito – PMP-Conj				
Pretérito mais-que-perfeito – PMP-Ind	Futuro Simples – Fut-Conj				
Futuro Simples – Fut-Ind	Futuro Composto – Fut-C-Conj				
Futuro Composto – Fut-C-Ind					

Table 2: Verb tense attributes

Type	Definition
PES	The referent is a person.
ORG	The referent is an organization.
OBJ	The referent is a tangible object, whether or not made by a human being.
relation_price/unit	The referent is an abstract numerical relationship between a unit of measurement and a value in some monetary system.
LOC	The referent is a concrete or abstract location.
TAR	The referent is a table (or a range of that table) of rates that are charged for a specific service performed.

Table 3: Type values and definitions

Semantic Role	Definition
Agent	Entity that intentionally causes a price change.
Cause	Entity that causes a price change unintentionally.
Theme	Entity that changes price, in an event.
Pivot	Entity that maintains the price, in a state.
Quantity	Quantification expression that indicates the amount of change operated in the price, in events.
Attribute	Quantification expression that indicates the quantity at which the price is maintained, in states.
Goal	Quantification expression indicating the end point of the price change.
Source	Quantification expression that indicates the starting point of the price change.
Locative	Expression that represents the place (concrete or abstract) where an entity is located or the event is held.

Table 4: Semantic roles and definitions