DODFCorpus: annotated corpus for Brazilian contracts and public bids

Artur Pereira
Department of Computer Science
University of Brasília
Brasília, Brazil
arturhcpereira@gmail.com

***Abstract*—Natural language processing can aid professionals when analysing official documents, but language models based on machine learning rely on larger amounts of data. On top of that, domain specific applications might demand domain specific *corpora* for the model to learn the nuances of that domain's language. With those factor in mind, this paper presents a manually annotated *corpus* for official documents in Brazilian Portuguese. This *corpus* is composed of government gazette's publication of contracts, covenants and public bids. Entities of interest are annotated in each publication as well as the publication's full text. The annotation and validation process are described in detail and the results of agreement metrics are presented. Krippendorff's alpha evaluation showed great label agreement, while the structural similarity metrics revealed a high variation of how each entity was annotated by different annotators.**

***Index Terms*—Annotated *corpus*, *corpus* validation, legal *corpora*, agreement metrics**

I. Introduction

Government gazettes present important information about public bidding processes and the contracts originated therefrom. Being able to analyze that information is relevant for specialists to audit the processes and prevent fraud. Natural language processing (NLP) tasks such as named entity recognition (NER) and text classification can be employed to develop tools to aid that analysis by extracting knowledge from the texts.

Machine learning based NLP models rely heavily on large labeled data to train them. With that in mind, a specific application in the legal domain such as the analysis of government gazettes demands the model to learn the particularities of the legal vernacular and the official document's structure. So the proposal of this paper is the creation and evaluation of a *corpus* of text for Brazilian government gazettes' publications, with special focus on public bidding and contract's excerpts. These publications are retrieved from the *Diário Oficial do Distrito Federal* (DODF) documents, a government gazette from the Federal District.

The annotation process can be approached in three main different ways: automatic, semiautomatic and manual. All of those approaches are reported in the creation of legal domain *corpora*. Filtz et al. [1] present a manually annotated *corpus* for NER and text classification. Petrova et al. [2] use an automatically annotated dataset, while Westermann et al. [3] propose a proof-of-concept system for aiding manual annotation by grouping similar sentences. Besides domain and language specificity, *corpora* need to be specific to the application they are intended to be used for: Luz et al. [4] present a Brazilian Portuguese *corpus* for NER. Their annotation and evaluation processes still differs from the same language, same domain, summarization dataset presented by Vargas and Moreira [5].

Wissler [6] defines a high quality Gold Standard Corpus for NLP as a manually annotated corpus reviewed by experts. The method for the creation of the proposed *corpus* follows the idea of manual annotation with rounds of revision. Even though, the revision was not entirely performed by experts. The annotation process was partially performed by voluntary annotators, while trained researches proceeded with the other part of annotation and performed the revisions consulting experts when necessary to resolve doubts.

The evaluation of *corpora* can be of intrinsic or extrinsic nature, both of which can be important for validating a dataset for practical applications, such as presented by Theijssen et al. [7]. The quality evaluation presented in this paper concerns the intrinsic inter-annotator agreement of the *corpus*. Two types of agreement were measured: label agreement and annotation agreement. The first type concerns annotators labeling for the same annotated entity and the metric adopted was the Krippendorff's Alpha [8]. The second type consists in comparing the structural similarity of the annotations made for the same entity. The metrics adopted for such comparison were the Levenshtein distance [9], the cosine similarity using TF-IDF vectors and the Jaccard's index [10] adapted to character sequences.

The rest of this paper is structured as follows: Section II goes over previous works in the literature, their methodology and results; Section III presents the method used for the creation and evaluation of this work's *corpus*; Section IV shows the experimental results obtained; and Section V briefly summarizes the results of the project and possible further steps.

II. Related works

Luz et al. [4] presents a dataset for name entity recognition (NER) in Brazilian legal documents. The necessity for an annotated *corpus* specific to the legal domain in Portuguese

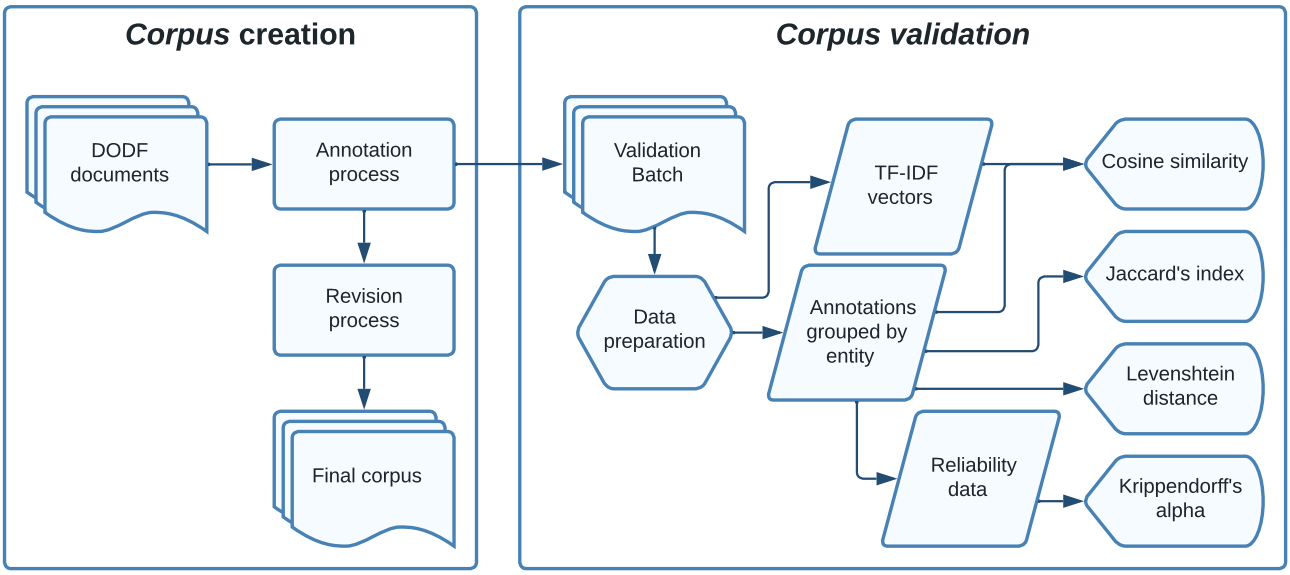

Fig. 1. Proposed method diagram divided in creation and validation processes.

arises from the particularities this kind of text presents, which may not be learned by a model trained on previous datasets composed of other classes of text. The authors first preprocessed the texts for sentence splitting and tokenization and then manually annotated each document with the desired tags. After that, the IOB [11] scheme was used to tag whether a word was the beginning of a named entity (B), was inside of a named entity (I) or was not part of any entity (O). To have baseline results, experiments with the LSTM-CRF model [12] were performed in the Paramopana [13] dataset. They retrained and evaluated the model in their dataset, obtaining F1 scores of 97.04% and 88.82% for the two entity types of interest for the legal field.

Filtz et al. [1] use a manually annotated corpus to evaluate different methods used for event extraction in legal documents. The work is motivated by the necessity of legal professionals to refer to previous cases when working on a current one. They propose using text classification and named entity recognition (NER) to extract information and provide an automated timeline of events for cases. The corpus used consisted of 30 documents from the European Court of Human Rights and was annotated by two experts in a series of iterations before meeting with a third person to resolve disagreements. The authors tested several models based on different approaches, including deep learning, conditional random fields (CRF) and rule-based methods. The best F-scores for the event classification were obtained by a finetuned DistilBERT model, being 92.38% and 79.75% for the two classes. For the NER task, CRF had a F-score of 80.50% for one of the three types of entity while the finetuned BERT obtained 90.22% and 90.44% for the other two.

III. Proposed Method

The proposed method is split in two steps: *corpus* creation and *corpus* validation. Those steps are represented in the diagram of Figure 1 and explained in further detail in the following subsections.

A. Corpus creation

The process for creating the *corpus* starts with the unlabeled DODF documents. The original documents' contents were extracted from the PDF files and loaded in the annotation tool. The tool used for this project was an adapted version of TeamTat [14]. The annotation process was partially performed by voluntary annotators.

There were six publication types of interest for the annotation process: contract excerpt, contractual amendment, covenant excerpt, bidding notice, bidding suspension and bidding annulment/revocation. The full content of the publication was annotated as an entity itself. Table I presents the relation between each kind of publication and its respective entity type.

TABLE I
Publication types and their respective entity types for annotation

Publication type	Entity type
Contract excerpt	EXTRATO_CONTRATO
Contract amendment	EXTRATO_ADITAMENTO_CONTRATUAL
Covenant excerpt	EXTRATO_CONVENIO
Bidding notice	AVISO_LICITACAO
Bidding suspension	AVISO_SUSPENSAO_LICITACAO
Bidding annulment/revocation	AVISO_ANUL_REV_LICITACAO

Besides the publications, 40 other entity types of interest were annotated. Each one of those may appear inside one or more publication types. A more detailed relation between those entities and their parent publication entities will be given in the presentation of results in Section IV.

After the annotation process, each annotated document would be revised by another annotator. During both the annotation and the revision steps, any doubt concerning the annotation of an entity would be compiled in a document. Periodically, the domain experts reviewed those documents and clarified the questions. During the revision process, relations were created to link annotated entities with their parent publication entity. The result of this process was the final version of the *corpus*.

B. Corpus validation

During the annotation process, a group of DODF documents was set aside as a Validation Batch. The Validation Batch was annotated and revised as the rest of the documents to constitute part of the final *corpus*. But the difference was that each document from the Validation Batch was annotated by multiple annotators in such a way the each publication type was annotated by at least two different people. The annotations of the Validation Batch were used, before revision, to evaluate the agreement metrics of the dataset. The steps for obtaining the metrics are detailed below.

1) Data preparation: In order evaluate the quality of the annotations made, it was necessary to determine which annotations referred to the same entity in the text so they could be compared with each other. So the first step of data preparation was to group the annotations.

The adopted approach to use the information about an annotation's length and its offset relative to the beginning of the document as the criteria for grouping. In practice, a base annotation was selected and every other annotation was tested against the grouping criteria. Every annotation that passed the test was added to the annotation group of the base annotation. The process was repeated for every annotation in each annotated document and the criteria were as follows:

1. Offset criterion: check if the annotation's offset is the same as the base annotation's offset with some tolerance based on the base annotation's length;
2. Length criterion: check if the length of the annotation is the same as the base annotation's length with some tolerance based on the base annotation's length.

Given an annotation A_2 and a base annotation A_1, Equations (1) and (2) formalize the grouping criteria respectively. $\text{Off}(A)$ and $\text{Len}(A)$ are respectively the offset and the length of annotation A, Tol is the tolerance for the criterion, P is how much the base length should influence the tolerance and K is a constant to establish a base tolerance. P and K values can be adjusted to address edge cases when, for instance, a small entity has annotations with varying offsets due to annotators including or not an additional word at the beginning.

$$\begin{cases} \text{Off}(A_2) \in]\text{Off}(A_1) - \text{Tol}_{\text{off}}, \text{Off}(A_1) + \text{Tol}_{\text{off}}[\\ \text{Tol}_{\text{off}} = P_{\text{off}} \times \text{Len}(A_1) + K_{\text{off}} \end{cases} \tag{1}$$

$$\begin{cases} \text{Len}(A_2) \in]\text{Len}(A_1) - \text{Tol}_{\text{len}}, \text{Len}(A_1) + \text{Tol}_{\text{len}}[\\ \text{Tol}_{\text{len}} = P_{\text{len}} \times \text{Len}(A_1) + K_{\text{len}} \end{cases} \tag{2}$$

The offset criterion serves the purpose of identifying possible matching annotations taking into account eventual errors in the start of the annotation. The length criterion is important to handle the cases of nested annotations that may start at near offsets, but shouldn't actually be compared with one another since they refer to different entities. This relies on the knowledge that nested annotations greatly differ in size. Figure 2 exemplifies the effects of the criteria when taking two annotations A_1 and A_2 made by different annotators in the case that A_2 is nested in A_1.

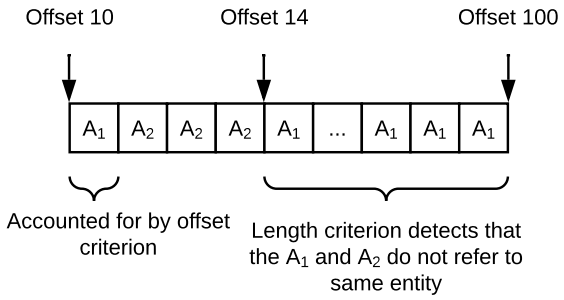

Fig. 2. The effects for the grouping criteria when comparing nested annotations of different entity types that often differ in length. The two annotations would be grouped by the offset criterion alone, but are not due to the length criterion.

A third criterion could be added to check if the annotation's were labeled with the same entity type, so annotations with different labels would not be compared. In that case, the evaluation would be considering that the annotators that disagreed in the labeling of the same entity did so consciously. In this paper, we considered that the mislabeling of annotations that passed the two first criteria was done by mistake and not by error of judgement. For that reason, the label criterion was not used for the purposes of grouping, but could be taken into account in future work to verify how the metrics are affected by it.

Besides the annotations grouped by entity, another data structure created in data preparation were the TF-IDF vectors for each annotation, that were used to compute cosine similarity. The TF-IDF vectors were created considering all the words in all the annotations without any text preprocessing in order to preserve particularities of the legal vernacular.

2) Agreement metrics: To evaluate the Krippendorff's alpha of the Validation Batch, a reliability data table was derived from the grouped annotations. The reliability data has one row per annotator and one column per entity, as exemplified in Table II. Each value of the table corresponds to the label the annotator gave to the entity.

TABLE II
EXAMPLE OF RELIABILITY DATA WITH THREE ANNOTATORS a, FOUR ENTITIES e AND THREE LABELS l

	e_1	e_2	e_3	e_4
a_1		l_1	l_2	
a_2			l_1	l_3
a_3	l_1	l_1		l_3

The other metrics compared two annotations at a time. So, for each annotation in a group, that annotation was assigned the average of the metrics' application over each annotation pair in that group. To handle cases of single annotation groups, the annotation was assigned a standard value for each metric. For the Jaccard's index and the cosine similarity that value was 0. For the Levenshtein distance that value was the length of the annotation, as it would be equivalent to compare the annotation's text with an empty text. Jaccard's index and Levenshtein distance compared the annotation's text directly, while cosine similarity used the text's TF-IDF vector representation.

IV. EXPERIMENTAL RESULTS

The experiments conducted used the values of P_{off} and P_{len} of 10% and the values $K_{\text{off}} = 13$ and $K_{\text{len}} = 23$. Those values were determined empirically to address cases short entity annotations varying much in length and offset due to inclusion or omission of words. The Krippendorff's alpha value obtained with those values for the grouping criteria was $\alpha_k = 0.97$.

For each entity type, the average and standard deviation of each metric were calculated and the number of annotations labeled with each type was recorded. Table III present those results for the entity types common to contracts and contract amendment, as well as for the publication types followed by their publication specific entity. Table IV present the results for covenant specific entity types. Table V present the results for entity types common to bidding notices, suspensions and annulment/revocations, as well as for the respective publication types followed by their publication specific entities. The results for the remaining entity types are presented in Table VI, with a column indicating to which publication types they belong to.

V. CONCLUSION

The experiments performed to extract the agreement metrics showed a high label agreement of $\alpha_k = 0.97$ and high average results for the structural agreement of most entity types. Even though, the results for structural agreement presented high standard deviations, which may reflect the difficulty to ensure higher standards of annotation when working with many voluntary annotators.

The results of structural metrics show that the publication type with the lowest agreement was EXTRATO_CONVENIO. That was an expected outcome considering that annotators would have difficulty understanding the structure of a publication that was so scarce in the dataset, as shown by the number of annotations in Table IV.

One unexpected result was the high standard deviation for Jaccard index, 0.4, for some purely numeric entity types, namely fonte_recurso, natureza_despesa and codigo_siggo. Other numeric entities, such as processo_gdf and codigo_licitacao_sistema_compras presented half of that deviation, indicating a greater consistency that was expected for numeric codes.

In light of the results, the importance of the revision process and the aid of specialists on the field is made clear. For future work, it might be interesting comparing the agreement obtained in the *corpus* after revision and see if the variance of annotation structure is improved. Another possibility for future work is to further validate the *corpus* with extrinsic evaluation.

REFERENCES

[1] E. Filtz, M. Navas-Loro, C. Santos, A. Polleres, and S. Kirrane, "Events matter: Extraction of events from court decisions," 2020.

[2] A. Petrova, J. Armour, and T. Lukasiewicz, "Extracting outcomes from appellate decisions in us state courts." in *JURIX*, 2020, pp. 133–142.

[3] H. Westermann, J. Šavelka, V. R. Walker, K. D. Ashley, and K. Benyekhlef, "Sentence embeddings and high-speed similarity search for fast computer assisted annotation of legal documents." in *JURIX*, 2020, pp. 164–173.

[4] P. H. Luz de Araujo, T. E. d. Campos, R. R. de Oliveira, M. Stauffer, S. Couto, and P. Bermejo, "Lener-br: a dataset for named entity recognition in brazilian legal text," in *International Conference on Computational Processing of the Portuguese Language*. Springer, 2018, pp. 313–323.

[5] D. d. Vargas Feijó and V. P. Moreira, "Rulingbr: A summarization dataset for legal texts," in *International Conference on Computational Processing of the Portuguese Language*. Springer, 2018, pp. 255–264.

[6] L. Wissler, M. Almashraee, D. M. Díaz, and A. Paschke, "The gold standard in corpus annotation." *IEEE GSC*, vol. 21, 2014.

[7] D. Theijssen, L. Boves, H. van Halteren, and N. Oostdijk, "Evaluating automatic annotation: automatically detecting and enriching instances of the dative alternation," *Language resources and evaluation*, vol. 46, no. 4, pp. 565–600, 2012.

[8] K. Krippendorff, "Computing krippendorff's alpha-reliability," 2011.

[9] V. I. Levenshtein *et al.*, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8. Soviet Union, 1966, pp. 707–710.

[10] P. Jaccard, "The distribution of the flora in the alpine zone. 1," *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[11] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," in *Natural language processing using very large corpora*. Springer, 1999, pp. 157–176.

[12] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *arXiv preprint arXiv:1603.01360*, 2016.

[13] C. M. Júnior, H. Macedo, T. Bispo, F. Santos, N. Silva, and L. Barbosa, "Paramopama: a brazilian-portuguese corpus for named entity recognition," *Encontro Nac. de Int. Artificial e Computacional*, 2015.

[14] R. Islamaj, D. Kwon, S. Kim, and Z. Lu, "Teamtat: a collaborative text annotation tool," *Nucleic acids research*, vol. 48, no. W1, pp. W5–W11, 2020.

TABLE III
AGREEMENT METRICS FOR CONTRACT EXCERPT AND AMENDMENT SPECIFIC ENTITY TYPES

Entity type	Levenshtein distance	Jaccard index	Cosine similarity with TF-IDF	Annotations
numero_contrato	2 ± 3	0.8 ± 0.3	0.8 ± 0.3	1724
orgao_contratante	8 ± 17	0.7 ± 0.4	0.7 ± 0.4	1972
codigo_siggo	$\mathbf{1 \pm 2}$	0.8 ± 0.4	0.8 ± 0.4	141
EXTRATO_CONTRATO	148 ± 486	0.9 ± 0.3	0.9 ± 0.3	1083
data_assinatura_contrato	1 ± 3	0.9 ± 0.3	0.9 ± 0.3	726
entidade_contratada	6 ± 13	0.8 ± 0.4	0.8 ± 0.3	1084
cnpj_entidade_contratada	1 ± 4	$\mathbf{0.9 \pm 0.2}$	$\mathbf{0.9 \pm 0.2}$	225
objeto_contrato	38 ± 125	0.9 ± 0.3	0.9 ± 0.3	1039
valor_contrato	1 ± 3	0.9 ± 0.3	0.9 ± 0.3	800
nota_empenho	1 ± 4	0.9 ± 0.3	0.9 ± 0.3	680
vigencia_contrato	17 ± 44	0.8 ± 0.4	0.8 ± 0.4	938
cnpj_orgao_contratante	2 ± 7	0.9 ± 0.3	0.9 ± 0.3	71
EXTRATO_ADITAMENTO_CONTRATUAL	65 ± 293	0.9 ± 0.2	0.9 ± 0.2	607
numero_termo_aditivo	1 ± 5	$\mathbf{0.9 \pm 0.2}$	$\mathbf{0.9 \pm 0.2}$	628
objeto_aditamento_contratual	97 ± 266	0.8 ± 0.4	0.8 ± 0.4	606

TABLE IV
AGREEMENT METRICS FOR COVENANT EXCERPT SPECIFIC ENTITY TYPES

Entity type	Levenshtein distance	Jaccard index	Cosine similarity with TF-IDF	Annotations
EXTRATO_CONVENIO	647 ± 591	0.3 ± 0.5	0.3 ± 0.5	9
entidade_convenente	21 ± 18	0.2 ± 0.4	0.2 ± 0.4	10
cnpj_entidade_convenente	14 ± 8	0.2 ± 0.4	0.3 ± 0.4	4
orgao_concedente	34 ± 11	0.0 ± 0.0	0.0 ± 0.0	7
cnpj_orgao_concedente	18 ± 0	0.0 ± 0.0	0.0 ± 0.0	1
objeto_convenio	111 ± 179	0.5 ± 0.5	0.5 ± 0.5	11
vigencia_convenio	25 ± 23	0.4 ± 0.5	0.4 ± 0.5	9
valor_convenio	4 ± 6	$\mathbf{0.7 \pm 0.5}$	$\mathbf{0.7 \pm 0.5}$	3
data_assinatura_convenio	10 ± 11	0.5 ± 0.5	0.5 ± 0.5	8
numero_convenio	$\mathbf{7 \pm 5}$	0.2 ± 0.4	0.2 ± 0.4	9

TABLE V
AGREEMENT METRICS FOR BIDDING NOTICE, SUSPENSION AND ANNULMENT/REVOCATION SPECIFIC ENTITY TYPES

Entity type	Levenshtein distance	Jaccard index	Cosine similarity with TF-IDF	Annotations
modalidade_licitacao	2 ± 4	0.9 ± 0.3	0.9 ± 0.3	830
numero_licitacao	3 ± 51	0.9 ± 0.3	0.9 ± 0.3	1104
orgao_licitante	6 ± 17	0.7 ± 0.4	0.7 ± 0.4	774
AVISO_LICITACAO	29 ± 169	$\mathbf{1.0 \pm 0.1}$	$\mathbf{1.0 \pm 0.1}$	873
tipo_objeto	4 ± 6	0.7 ± 0.5	0.7 ± 0.5	873
valor_estimado_contratacao	0 ± 2	1.0 ± 0.2	1.0 ± 0.2	560
data_abertura_licitacao	1 ± 3	0.9 ± 0.2	0.9 ± 0.2	758
sistema_compras	2 ± 7	0.9 ± 0.3	0.9 ± 0.3	900
codigo_licitacao_sistema_compras	$\mathbf{0 \pm 2}$	0.9 ± 0.2	0.9 ± 0.2	611
AVISO_SUSPENSAO_LICITACAO	27 ± 204	$\mathbf{1.0 \pm 0.2}$	$\mathbf{1.0 \pm 0.2}$	84
decisao_tcdf	$\mathbf{23 \pm 37}$	0.2 ± 0.3	0.4 ± 0.4	19
AVISO_ANUL_REV_LICITACAO	34 ± 165	$\mathbf{1.0 \pm 0.2}$	$\mathbf{1.0 \pm 0.2}$	30
identificacao_ocorrencia	$\mathbf{2 \pm 5}$	0.8 ± 0.3	0.8 ± 0.3	33

TABLE VI
AGREEMENT METRICS FOR ENTITY TYPES COMMON TO TWO OR MORE PUBLICATION TYPES

Entity type	Levenshtein distance	Jaccard index	Cos. similarity with TF-IDF	Annotations	Publication types
processo_gdf	1 ± 5	0.9 ± 0.2	$\mathbf{0.9 \pm 0.2}$	2570	All
nome_responsavel	1 ± 5	$\mathbf{1.0 \pm 0.2}$	0.9 ± 0.2	1183	All
unidade_orcamentaria	1 ± 3	0.9 ± 0.3	0.9 ± 0.3	558	Contracts and covenants
programa_trabalho	3 ± 6	0.8 ± 0.3	0.9 ± 0.3	712	Contracts and covenants
fonte_recurso	4 ± 10	0.7 ± 0.4	0.7 ± 0.4	801	Contracts and covenants
natureza_despesa	$\mathbf{1 \pm 2}$	0.8 ± 0.4	0.8 ± 0.4	564	Contracts and covenants
objeto_licitacao	31 ± 90	0.9 ± 0.3	0.9 ± 0.3	951	Bidding notices and suspensions
data_escrito	1 ± 4	0.9 ± 0.3	0.9 ± 0.3	560	Amendments and suspensions