Base de Normas Jurídicas Brasileiras: uma iniciativa de *Open Government Data*

Hudson de Martim [I]

João Alberto de Oliveira Lima [II]

Lauro César Araujo [III]

[I] *Instituto Legislativo Brasileiro do Senado Federal, Brasília, DF, Brasil.*
Mestre e Graduado em Ciência da Computação pela UFMG.
Pesquisador do Instituto Legislativo Brasileiro do Senado Federal.

[II] *Doutor em Ciência da Informação pela UnB.*
Pesquisador do Instituto Legislativo Brasileiro do Senado Federal.

[III] *Doutor em Ciência da Informação pela UnB.*
Pesquisador do Instituto Legislativo Brasileiro do Senado Federal.

http://dx.doi.org/10.1590/1981-5344/3567

As normas jurídicas, produzidas por meio do Processo Legislativo, são a base formal de regulação da convivência em sociedade. Por isso, são naturalmente redigidas de forma técnica com objetivo de serem interpretadas juridicamente. Neste trabalho, porém, apresenta-se uma série de transformações automáticas aplicadas ao arcabouço de leis federais de modo a estruturar a informação descrita nesses documentos com intuito de prepará-las para diferentes tipos de interpretações automáticas, como identificação de entidades nomeadas, definições, remissões, eventos de criação, alteração e encerramento de instituições jurídicas, recuperação da versão vigente de uma lei no tempo. Isso visa auxiliar atividades de informação que vão além da própria interpretação jurídica, e vão ao encontro da Open Government Data. O artigo descreve uma série de datasets contendo os resultados de transformações da base de normas jurídicas brasileira, que contemplam os textos articulados das normas em representação LexML, CoNLL-U, representações sintáticas de sentenças obtidas com a Google Natural Language Processing API, entre outras.

Palavras-chave: *Norma Jurídica. Processamento de Linguagem Natural. LexML. CoNLL-U. Google Natural Language Processing API.*

Base of Brazilian Legal Norms: an Open Government Data Initiative

The legal norms, produced through the Legislative Process, are the formal basis for regulating coexistence in society. Therefore, they are naturally drafted in a technical way with the aim of being interpreted legally. In this paper, however, we present a series of automatic transformations applied to the framework of federal laws in order to structure the information described in these documents with the aim of preparing them for different types of automatic interpretations, such as identification of named entities, definitions, references, creation, alteration and closure events of legal institutions, recovery of the current version of a law in time. This aims to support information activities that go beyond legal interpretation itself, aligned with Open Government Data. The article describes a series of datasets containing the results of transformations of the Brazilian legal norms base, which include the articulated texts of the norms in LexML, CoNLL-U, syntactic representations of sentences obtained with the Google Natural Language Processing API, among others.

Keywords: *Legal norms. Natural language processing. LexML. CoNLL-U. Google Natural Language Processing API.*

Recebido em 18.06.2018 Aceito em 11.09.2018

1 Introdução

No contexto contemporâneo de produção e disponibilidade de grande quantidade de dados processados por máquina, um dos principais desafios enfrentados por cientistas da informação e pesquisadores em geral é o acesso organizado a esses dados. O movimento *Open Data* surgiu nos anos 2000 para, dentre outros objetivos, ajudar a suprir essas necessidades. Sendo um braço do *Open Science*, o *Open Data* prega a ideia dos dados abertos, compartilhados, livres para serem acessados, usados e redistribuídos por qualquer pessoa, para qualquer propósito, sem qualquer restrição (AUER et al., 2007). Uma especialização do *Open Data*

é o *Open Government Data* (OGD) (GRAY, 2014), que promove a ideia de que, tornando públicos os dados do governo, pode-se alcançar maior transparência das ações governamentais, estimulando um aumento da participação da sociedade na vida pública e contribuindo com o progresso em pesquisas baseadas nessas informações, como em Santos Neto et al. (2013), que propõe uma abordagem para interligar dados abertos de bibliotecas, arquivos e museus baseada em tecnologias e princípios para publicação de dados abertos estruturados na Web.

O Poder Legislativo coleta e produz diariamente uma grande quantidade de dados que são utilizados na execução de suas funções constitucionais de legislar, fiscalizar o governo e representar a sociedade. Dentre esses dados, as normas jurídicas, produzidas através do Processo Legislativo, constituem uma importante fonte de informação para a sociedade, por definir direitos, deveres, competências e imunidades, regulando as relações entre as pessoas e entre as pessoas e o Estado (HOHFELD, 2008). As normas jurídicas são naturalmente utilizadas como textos para interpretação jurídica no caso concreto. Porém, é desejável estruturar a informação descrita nesses documentos de modo que possa ser processada por máquinas para auxílio a diferentes tipos de atividades além da própria interpretação jurídica, a exemplo da obtenção do texto vigente para determinada data, da compreensão das alterações de ordenamento jurídico, da navegabilidade entre remissões expressas, entre outros.

O conjunto de normas jurídicas historicamente produzidas pelo Legislativo Federal representa um volume expressivo de textos que descrevem a evolução do ordenamento jurídico federal do país, sendo uma base de dados valiosa para cientistas da informação e pesquisadores em geral. Diante disso, o objetivo do presente trabalho é a construção de *datasets* abertos contendo os dados de normas jurídicas brasileiras de forma bruta, textual e estruturada.

Como resultado deste trabalho, foram produzidos oito *datasets*, conforme a seguir:

a) *dataset* "Textos Articulados das Normas" (subseção 2.1): contempla os textos articulados[1] da publicação original de cada norma jurídica produzida pelo Processo Legislativo Federal a partir de um marco histórico definido;

b) *dataset* "Representação LexML dos Textos Articulados das Normas" (subseção 2.2): contém os textos articulados de cada norma do *dataset* anterior estruturados em formato LexML;

c) *dataset* "Sentenças da Epígrafe, Ementa, Preâmbulo, Dispositivos e Fecho das Normas" (subseção 2.3): separa em arquivos distintos a epígrafe, a ementa, o preâmbulo,

[1] Textos estruturados com base na técnica legislativa brasileira, advindo da tradição do Império (ordenações portuguesas Filipinas, Manuelinas e Afonsinas) e da Lei Complementar n. 95, de 26 de fevereiro de 1998 (BRASIL. Presidência da República, 1998).

os dispositivos[2] e o fecho de cada norma do *dataset* anterior;

d) *dataset* "Sentenças dos Dispositivos das Normas com Enumerações Agrupadas" (subseção 2.4): agrupa os incisos e alíneas como enumerações da sentença de cada dispositivo do *dataset* anterior;

e) *dataset* "Representação CoNLL-U das sentenças dos Dispositivos das Normas" (subseção 2.5): contempla a representação CoNLL-U dos dispositivos do terceiro *dataset*;

f) *dataset* "Representação Sintática das sentenças dos Dispositivos das Normas" (subseção 2.6): contém o resultado da análise sintática de cada dispositivo do quarto *dataset* realizada por processamento de linguagem natural por meio da *Google Natural Language Processing API*;

g) *dataset* "Textos da Articulação e da Ementa das Normas" (subseção 2.7): agrupa as sentenças dos dispositivos de cada norma criando um arquivo com o texto completo da articulação da norma. É ainda oferecido, para cada norma, um arquivo com o texto da sua ementa;

h) *dataset* "Representação Sintática dos Textos da Articulação e da Ementa das Normas" (subseção 2.8): contém o resultado da análise sintática do texto da articulação e do texto da ementa de cada norma realizada por processamento de linguagem natural por meio da *Google Natural Language Processing API*.

A seção 2 contempla a descrição de cada um dos *datasets*. A seção 3 apresenta os métodos utilizados para produzi-los e as limitações encontradas. Considerações finais são apresentadas na seção 4.

2 Resultados

As normas jurídicas representam não apenas a saída, mas também a entrada do Processo Legislativo, pois as normas vigentes são constantemente alteradas. Para apoiar o Processo Legislativo Federal, o Senado Federal mantém uma base de dados de normas jurídicas com textos articulados da publicação original, republicações e retificações obtidas do Diário Oficial da União (DOU). As normas publicadas no DOU contemplam os textos vigentes na época de sua publicação, ou seja, são os textos com validade jurídica. Os *datasets* produzidos nesta pesquisa foram construídos a partir das bases do Senado Federal. Portanto, referem-se aos textos originais válidos das normas jurídicas. Entretanto, como ressaltado em Lima (2013, p. 80), "os textos publicados nas bases de dados de legislação na internet no Brasil, mesmo que armazenadas em

2 Exemplos de dispositivos são Capítulos, Artigos, Incisos, entre outros.

sítios oficiais das instituições do Estado, não possuem nenhum caráter oficial". Por isso, a construção de defesas jurídicas a partir de documentos derivados da publicação original devem observar a necessidade de consulta ao DOU.

Os *datasets* são compostos de textos de Leis Ordinárias e Leis Complementares federais do período entre 4 de outubro de 1946 e 12 de abril de 2017. Procurou-se um recorte metodológico que selecionasse uma quantidade razoável de normas com maior probabilidade de estarem vigentes. Isso não exclui normas expressamente revogadas, pois o histórico dos textos é importante para a obtenção do texto vigente em uma determinada data[3]. Não estão incluídas as Emendas Constitucionais, Decretos-Leis, Decretos, entre outros atos normativos, tão pouco os textos das proposições legislativas.

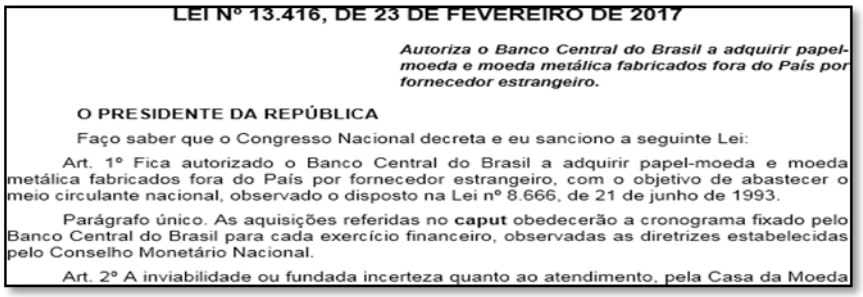
As subseções a seguir descrevem cada *dataset* oferecido e apresentam, como exemplo, o resultado do processamento da Lei 13.416 de 23 de fevereiro de 2017 (BRASIL. Presidência da República, 2017).

2.1 Dataset 1: textos articulados das normas

O *dataset* "Textos Articulados das Normas" contém, para cada lei ordinária e lei complementar do período definido, um arquivo *Rich Text Format* (RTF) com o texto articulado original extraído do DOU. Há 13.567 arquivos de normas, cujo nome é formatado como <tipo_da_norma>-<ano_da_norma>-<número da norma>.rtf.

A Figura 1 apresenta o texto da Lei 13.416, de 2017 (BRASIL. Presidência da República, 2017), composto pela epígrafe, ementa, preâmbulo e parte articulada.

Figura 1 – Exemplo de conteúdo de arquivo do *Dataset 1* (arquivo LEI-2017-13416.rtf)



LEI Nº 13.416, DE 23 DE FEVEREIRO DE 2017

Autoriza o Banco Central do Brasil a adquirir papel-moeda e moeda metálica fabricados fora do País por fornecedor estrangeiro.

O PRESIDENTE DA REPÚBLICA

Faço saber que o Congresso Nacional decreta e eu sanciono a seguinte Lei:

Art. 1º Fica autorizado o Banco Central do Brasil a adquirir papel-moeda e moeda metálica fabricados fora do País por fornecedor estrangeiro, com o objetivo de abastecer o meio circulante nacional, observado o disposto na Lei nº 8.666, de 21 de junho de 1993.

Parágrafo único. As aquisições referidas no **caput** obedecerão a cronograma fixado pelo Banco Central do Brasil para cada exercício financeiro, observadas as diretrizes estabelecidas pelo Conselho Monetário Nacional.

Art. 2º A inviabilidade ou fundada incerteza quanto ao atendimento, pela Casa da Moeda

Fonte: Dados da pesquisa.

2.2 Dataset 2: representação LexML dos textos articulados das normas

O LexML é um portal administrado pelo Senado Federal especializado em informação jurídica e legislativa que pretende reunir leis, decretos, acórdãos, súmulas, projetos de leis entre outros documentos

[3] O princípio do direito civil "*Tempus regit actum*" define que o ato jurídico se rege pela lei da época, daí a importância de se ter o histórico de textos.

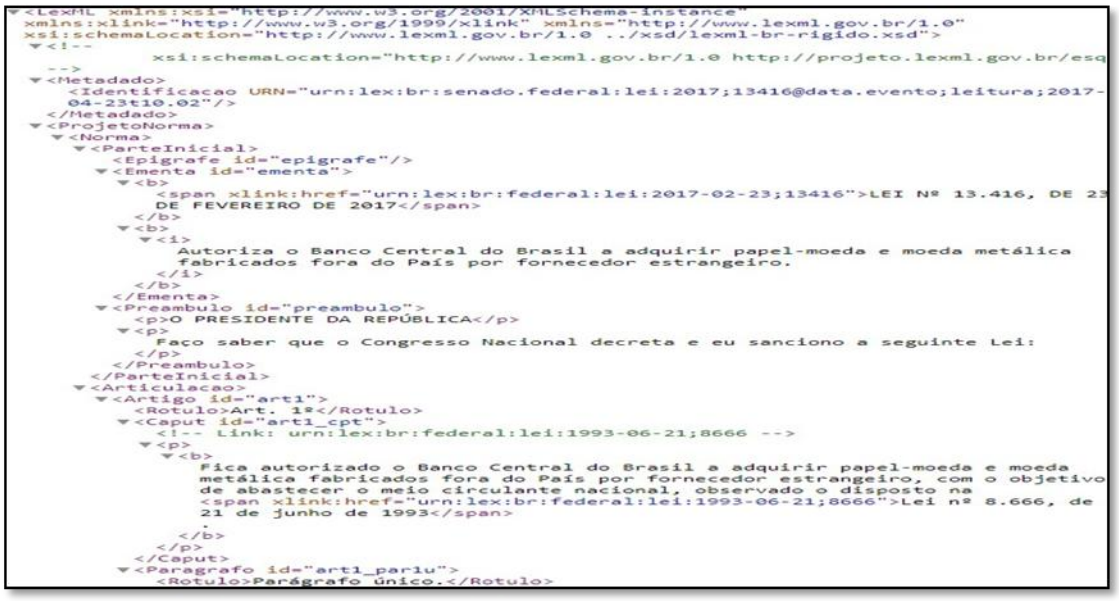
das esferas federal, estadual e municipal dos Poderes Executivo, Legislativo e Judiciário de todo o Brasil. O LexML oferece um esquema XML[4] para a estruturação dos textos de normas, julgados e projetos de normas através de vocabulário unificado.

O *dataset* "Representação LexML dos Textos Articulados das Normas" é o resultado da execução do *Parser* LexML sobre o *dataset* 1 e contém um arquivo no formato LexML para cada norma da entrada, em caso de sucesso na conversão. Há 12.569 arquivos, cujo nome é formatado como <tipo_da_norma>-<ano_da_norma>-<número_da_norma>.xml – na subseção "Limitações Encontradas" da seção "Métodos", há uma descrição sobre as causas de o número de arquivos de saída (*dataset* 2) ser menor que o número de arquivos de entrada (*dataset* 1).

A Figura 2 apresenta o texto da Lei n.13416/2017 estruturado em XML, esquema LexML. Nota-se que os elementos epígrafe, ementa, preâmbulo e parte articulada estão precisamente delimitados por *tags* (rótulos) do vocabulário LexML. O vocabulário incorporado denota elementos semânticos de organização da informação normativa estruturada. O uso padrão LexML vai ao encontro do que Ribeiro e Pereira (2015, p. 75) defendem:

> Para que os princípios de dados abertos sejam completamente atendidos, é preciso que haja uma organização semântica dos dados publicados [...], por exemplo, através do uso de padrões abertos e de vocabulários estruturados que possibilitem a padronização terminológica, visando à comunicação e compreensão dos dados descritos por máquinas.

Figura 2 – Exemplo de conteúdo de arquivo do *Dataset 2* (LEI-2017-13416.xml)



Fonte: Dados da pesquisa.

4 A documentação do esquema LexML encontra-se em: <http://projeto.lexml.gov.br/documentacao/Parte-3-XML-Schema.pdf>. Acesso em: 11 mar. 2018.

2.3 Dataset 3: sentenças da epígrafe, ementa, preâmbulo, dispositivos e fecho das normas

Em trabalhos de extração semântica de normas, pode ser necessário analisar o texto da norma como um todo, por exemplo, para extrair entidades nomeadas, ou analisar individualmente a sentença de cada elemento (epígrafe, ementa, preâmbulo, dispositivos, fecho) para, por exemplo, tipificar remissões ou extrair definições.

O *dataset* "Sentenças da Epígrafe, Ementa, Preâmbulo, Dispositivos e Fecho das Normas" é o resultado do processamento do *dataset* 2 e contém um diretório para cada norma de entrada, que contém um arquivo separado para cada elemento da norma. Há 362.030 arquivos nesse *dataset*.

Na Figura 3, é apresentada a sentença do *caput* do artigo 1 da Lei n. 13.416, de 2017 (BRASIL. Presidência da República, 2017).

Figura 3 – Exemplo de conteúdo de arquivo do *Dataset 3* (arquivo LEI-2017-13416/0005-art1_cpt.txt)

```
Fica autorizado o Banco Central do Brasil a adquirir papel-moeda e moeda metálica fabricados fora do País por
fornecedor estrangeiro, com o objetivo de abastecer o meio circulante nacional, observado o disposto na Lei nº
8.666, de 21 de junho de 1993.
```

Fonte: Dados da pesquisa.

2.4 Dataset 4: sentenças dos dispositivos das normas com enumerações agrupadas

O *dataset 2* apresenta os incisos e alíneas, nos textos articulados das normas, estruturados como dispositivos distintos, apesar de serem enumerações de um dispositivo agregador. Os incisos 1 e 2 do parágrafo 1 do artigo 2 da Lei n. 13.416, de 2017 (BRASIL. Presidência da República, 2017), apresentados na Figura 4, são um exemplo dessa situação.

Figura 4 – Conteúdo LexML do parágrafo 1 do artigo 2 da Lei 13.416/2017

```
<Paragrafo id="art2_par1">
  <Rotulo>§ 1º</Rotulo>
  <p>
    <b>
      Caracterizam a inviabilidade ou fundada incerteza de que trata o caput :
    </b>
  </p>
  <Inciso id="art2_par1_inc1">
    <Rotulo>I -</Rotulo>
    <p>
      <b>
        o atraso acumulado de 15% (quinze por cento) das quantidades contratadas, por denominação,
        de papel-moeda ou de moeda metálica; e
      </b>
    </p>
  </Inciso>
  <Inciso id="art2_par1_inc2">
    <Rotulo>II -</Rotulo>
    <p>
      <b>
        outras hipóteses de descumprimento de cláusula contratual, devidamente justificadas, que
        tornem inviável o atendimento da demanda por meio circulante ou do cronograma para seu
        abastecimento.
      </b>
    </p>
  </Inciso>
</Paragrafo>
<Paragrafo id="art2_par2">...</Paragrafo>
</Artigo>
```

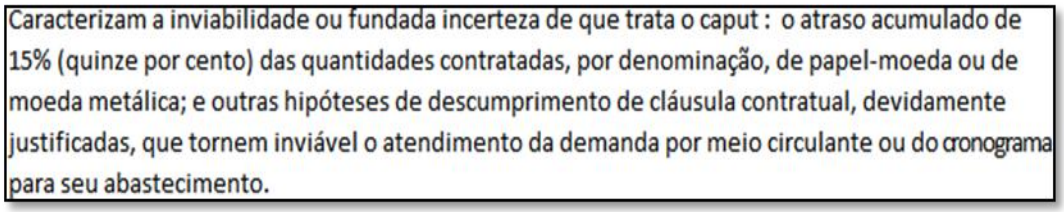
Fonte: Dados da pesquisa.

Com essa separação, as sentenças no *dataset* 3 de dispositivos que possuem incisos e alíneas ficam malformadas, incompletas, geralmente terminadas com um ":" (dois pontos). Exemplo desse tipo de sentença é apresentado no conteúdo do parágrafo 1 do artigo 2 da Lei n. 13.416, de 2017 (BRASIL. Presidência da República, 2017) da Figura 4.

Assim, a partir do *dataset* 3, foi produzido o *dataset* "Sentenças dos Dispositivos das Normas com Enumerações Agrupadas", que possui arquivos apenas para os dispositivos agregadores, tendo acrescentadas ao final de suas sentenças, como enumerações, as sentenças de seus incisos e alíneas. Há 128.547 arquivos nesse *dataset*.

Na Figura 5, é apresentada a sentença do parágrafo 1 do artigo 2 da Lei n. 13.416, de 2017 (BRASIL. Presidência da República, 2017) com as sentenças dos seus incisos 1 e 2 concatenadas como enumerações, conforme o arquivo gerado no *dataset* 4 para o parágrafo em questão -- não foram gerados arquivos para os incisos 1 e 2 no *dataset* 4.

Figura 5 – Sentença do parágrafo 1 do artigo 2 da Lei 13.416/2017 com seus incisos concatenados como enumerações


Caracterizam a inviabilidade ou fundada incerteza de que trata o caput : o atraso acumulado de 15% (quinze por cento) das quantidades contratadas, por denominação, de papel-moeda ou de moeda metálica; e outras hipóteses de descumprimento de cláusula contratual, devidamente justificadas, que tornem inviável o atendimento da demanda por meio circulante ou do cronograma para seu abastecimento.

Fonte: Dados da pesquisa.

2.5 Dataset 5: representação CoNLL-U das sentenças dos dispositivos das normas

O processamento de textos por máquina com objetivo de indexação, de extração de definições, de tipificação de remissões, por exemplo, não traz resultados satisfatórios em geral quando realizado diretamente sobre sentenças em linguagem natural. Esse tipo de processamento, que busca o reconhecimento da semântica das sentenças, exige uma abordagem mais avançada do que a de processamento textual simples. Estudos de Batista et al. (2011), Nakamura, Ogawa e Toyama (2013) demonstram que, em um processamento de extração da semântica, um passo intermediário de análise sintática das sentenças contribui de maneira relevante para a acurácia dos resultados, contribuindo também na simplificação da implementação do processamento.

O formato CoNLL-U (BUCHHOLZ; MARSI, 2006), utilizado no projeto *Universal Dependences* (UD) (NIVRE et al., 2016), estrutura cada palavra/*token* de uma sentença em uma linha com colunas separadas por *tab*. Cada coluna, em termos gerais, representa uma aspecto da palavra/*token*, como a forma, o lema, a *tag "part-of-speech"*, características morfológicas etc. Essa estruturação de características e dependências sintáticas/morfológicas tem como objetivo facilitar o processamento multi-linguagem de linguagem natural, além de, por

exemplo, suportar aprendizado e avaliações comparativas através de línguas diferentes.

Tendo o *dataset* 4 como entrada, foi produzido o *dataset* "Representação CoNLL-U das sentenças dos Dispositivos das Normas", que contém um diretório para cada norma de entrada e, dentro do diretório de cada norma, um arquivo texto (TXT) para cada dispositivo, cujo conteúdo é a sentença do dispositivo em formato CoNLL-U. Há 337.520 arquivos nesse *dataset*.

Na Figura 6, é apresentada a sentença do *caput* do artigo 1 da Lei n.13.416, de 2017 (BRASIL. Presidência da República, 2017) estruturada em CoNLL-U.

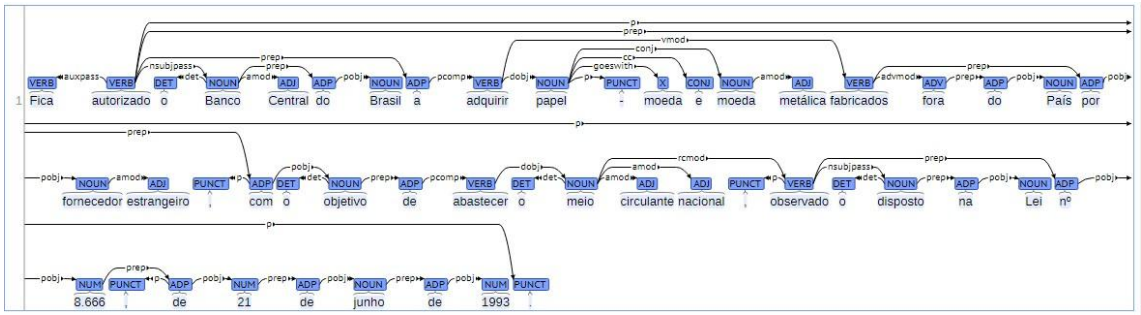
Figura 6 – Estrutura CoNLL-U da sentença do *caput* do artigo 1 da Lei 13.416/2017

1	Fica	_	VERB	VERB	_	2	nsubj	_	_
2	autorizado	_	VERB	VERB	_	0	ROOT	_	_
3	o	_	DET	DET	_	4	det	_	_
4	Banco	_	PROPN	PNOUN	_	2	obj	_	_
5	Central	_	PROPN	PNOUN	_	4	amod	_	_
6	do	_	CCONJ	CONJ	_	7	cc	_	_
7	Brasil	_	PROPN	PNOUN	_	4	conj	_	_
8	a	_	ADP	ADP	_	9	mark	_	_
9	adquirir	_	VERB	VERB	_	2	advcl	_	_
10	papel-moeda	_	ADJ	ADJ	_	9	xcomp:adj	_	_
11	e	_	CCONJ	CONJ	_	12	cc	_	_
12	moeda	_	NOUN	NOUN	_	10	conj	_	_
13	metálica	_	ADJ	ADJ	_	12	amod	_	_
14	fabricados	_	VERB	VERB	_	9	acl:part	_	_
15	fora	_	ADV	ADV	_	14	advmod	_	_
16	do	_	X	ADPPRON	_	17	case	_	_
17	País	_	PROPN	PNOUN	_	15	nmod	_	_
18	por	_	ADP	ADP	_	19	case	_	_
19	fornecedor	_	NOUN	NOUN	_	14	nmod	_	_
20	estrangeiro	_	ADJ	ADJ	_	19	amod	_	_
21	,	_	PUNCT	.	_	26	punct	_	_
22	com	_	ADP	ADP	_	26	mark	_	_
23	o	_	DET	DET	_	22	fixed	_	_
24	objetivo	_	NOUN	NOUN	_	22	fixed	_	_
25	de	_	ADP	ADP	_	22	fixed	_	_
26	abastecer	_	VERB	VERB	_	14	advcl	_	_
27	o	_	DET	DET	_	28	det	_	_
28	meio	_	NOUN	NOUN	_	26	obj	_	_
29	circulante	_	NOUN	NOUN	_	28	appos	_	_
30	nacional	_	ADJ	ADJ	_	29	amod	_	_
31	,	_	PUNCT	.	_	32	punct	_	_
32	observado	_	VERB	VERB	_	26	acl:part	_	_
33	o	_	DET	DET	_	34	det	_	_
34	disposto	_	NOUN	NOUN	_	32	obj	_	_
35	na	_	ADJ	ADJ	_	34	amod	_	_
36	Lei	_	PROPN	PNOUN	_	34	appos	_	_
37	nº	_	NUM	NUM	_	38	nummod	_	_

Fonte: Dados da pesquisa.

Na Figura 7 é apresentada a sentença do *caput* do artigo 1 da Lei n. 13.416, de 2017 (BRASIL. Presidência da República, 2017) no grafo da sua árvore sintática, renderizado a partir de sua estrutura CoNLL-U.

Figura 7 – Renderização da árvore sintática da sentença do *caput* do artigo 1 da Lei 13.416/2017



Fonte: Dados da pesquisa.

2.6 Dataset 6: representação sintática das sentenças dos dispositivos das normas

A Google oferece uma *API Cloud*[5] para processamento de linguagem natural (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011), que é apresentada como um conjunto de serviços web/rest que utilizam modelos e estratégias de aprendizado de máquina para realizar análise avançada de textos. Dentre os serviços oferecidos por essa *API*, há o serviço de "Análise Sintática", que extrai do texto *tokens* e frases, identifica classes gramaticais (*Parts of Speech* - PoS), e cria uma árvore de análise sintática para cada frase. Além dele, há o serviço de "Análise de Entidades" que identifica entidades que aparecem no texto e as classifica como pessoa, organização, local, eventos, produtos, mídia.

Com base no *dataset* 4, foi produzido o *dataset* "Representação Sintática das sentenças dos Dispositivos das Normas", que contém um diretório para cada norma de entrada e, dentro do diretório de cada norma, um arquivo *json* para cada dispositivo, cujo conteúdo é o resultado do processamento da sentença do dispositivo pela *API* da Google, utilizando-se os serviços "Análise Sintática" e "Análise de Entidades". Há 128.547 arquivos nesse *dataset*.

Na Figura 8, é apresentada a sentença do *caput* do artigo 1 da Lei 13.416, de 2017 (BRASIL. Presidência da República, 2017) estruturada em elementos *json* da *API* de processamento de linguagem natural da Google.

Figura 8 – Exemplo de conteúdo de arquivo do *Dataset 6* (arquivo LEI-2017-13416/0005-art1_cpt.json)


```
{
  "sentences": [
    {
      "text": {
        "content": "Fica autorizado o Banco Central do Brasil a adquirir papel-moeda e moeda metÃ¡lica
fabricados fora do PaÃs por fornecedor estrangeiro, com o objetivo de abastecer o meio circulante
nacional, observado o disposto na Lei nÂº 8.666, de 21 de junho de 1993.",
        "beginOffset": 0
      }
    }
  ],
  "tokens": [
    {
      "text": {
        "content": "Fica",
        "beginOffset": 0
      },
      "partOfSpeech": {
        "tag": "VERB",
        "aspect": "IMPERFECTIVE",
        "case": "CASE_UNKNOWN",
        "form": "FORM_UNKNOWN",
        "gender": "GENDER_UNKNOWN",
        "mood": "INDICATIVE",
        "number": "SINGULAR",
        "person": "THIRD",
        "proper": "NOT_PROPER",
        "reciprocity": "RECIPROCITY_UNKNOWN",
        "tense": "PRESENT",
        "voice": "VOICE_UNKNOWN"
      },
      "dependencyEdge": {
        "headTokenIndex": 1,
        "label": "AUXPASS"
      },
      "lemma": "Ficar"
    },
    {
      "text": {
        "content": "autorizado",
        "beginOffset": 5
      },
      "partOfSpeech": {
        "tag": "VERB",
        "aspect": "PERFECTIVE",
        "case": "CASE_UNKNOWN",
```


Fonte: Dados da pesquisa.

[5] A API de Processamento de Linguagem Natural da Google para português-br encontra-se em: <https://cloud.google.com/natural-language/?hl=pt-br>. Acesso em: 11 mar. 2018.

2.7 Dataset 7: textos da articulação e da ementa das normas

Como consta na descrição do *dataset* 3, "em trabalhos de extração semântica de normas, pode ser necessário analisar o texto da norma como um todo, por exemplo, para extrair entidades nomeadas, ou analisar individualmente a sentença de cada elemento (epígrafe, ementa, preâmbulo, dispositivos, fecho), para, por exemplo, tipificar remissões ou extrair definições".

A construção dos *datasets* 3 a 6, por exemplo, é focada em trabalhos que precisam analisar individualmente a sentença de cada elemento da norma. Já a construção dos *datasets* 7 e 8 tem como objetivo contribuir com trabalhos em que é necessário analisar o texto da norma como um todo, mais especificamente, o texto da articulação da norma. Por isso, com base no *dataset* 4, foi produzido o *dataset* "Textos da Articulação e da Ementa das Normas", que agrupa as sentenças dos dispositivos de cada norma criando, para cada norma, um arquivo com o texto completo da sua articulação e um arquivo com o texto da sua ementa. Há 25.944 arquivos nesse *dataset*, sendo metade com arquivos das articulações das normas, metade com arquivos das ementas das normas.

Na Figura 9, é apresentado o texto da articulação da Lei n. 13416, de 2017 (BRASIL. Presidência da República, 2017).

Figura 9 – Texto da articulação da Lei *LEI−2017−13416*

Fica autorizado o Banco Central do Brasil a adquirir papel-moeda e moeda metálica fabricados fora do País por fornecedor estrangeiro, com o objetivo de abastecer o meio circulante nacional, observado o disposto na Lei nº 8.666, de 21 de junho de 1993.
As aquisições referidas no caput obedecerão a cronograma fixado pelo Banco Central do Brasil para cada exercício financeiro, observadas as diretrizes estabelecidas pelo Conselho Monetário Nacional.
A inviabilidade ou fundada incerteza quanto ao atendimento, pela Casa da Moeda do Brasil, da demanda por meio circulante ou do cronograma para seu abastecimento, em cada exercÃcio financeiro, caracteriza situação de emergência, para efeito de aquisição de papel-moeda e de moeda metálica de fabricantes estrangeiros, na forma do inciso IV do caput do art. 24 da Lei nº 8.666, de 21 de junho de 1993.
Caracterizam a inviabilidade ou fundada incerteza de que trata o caput :
o atraso acumulado de 15% (quinze por cento) das quantidades contratadas, por denominação, de papel-moeda ou de moeda metálica; e
outras hipóteses de descumprimento de cláusula contratual, devidamente justificadas, que tornem inviável o atendimento da demanda por meio circulante ou do cronograma para seu abastecimento.
Para fins da caracterização da situação de emergência de que trata este artigo, o Banco Central do Brasil fica obrigado a enviar o Programa Anual de Produção à Casa da Moeda do Brasil, até 31 de agosto de cada ano, no qual serão indicadas as projeções de demandas de papel-moeda e de moeda metálica para o exercício financeiro seguinte.
Esta Lei entra em vigor na data de sua publicação.

Fonte: Dados da pesquisa.

Dataset 8: representação sintática dos textos da articulação e da ementa das normas

Com base no *dataset* 7, foi produzido o *dataset* "Representação Sintática dos Textos da Articulação e da Ementa das Normas", que contém dois arquivos *json* para cada norma, um para sua articulação e outro para sua ementa, cujos conteúdos são o resultado do processamento do texto correspondente pela *API* de processamento de linguagem natural da Google, utilizando-se os serviços "Análise Sintática" e "Análise de Entidades" – conforme apresentados na descrição do *dataset* 6. Há 25.944 arquivos nesse *dataset*, sendo metade com arquivos das articulações processadas das normas, metade com arquivos das ementas processadas das normas.

Na Figura 10 é apresentado o texto da articulação da Lei n. 13.416, de 2017 (BRASIL. Presidência da República, 2017) estruturada em elementos *json* da *API* de processamento de linguagem natural da Google.

Figura 10 – Exemplo de conteúdo de arquivo do *Dataset 8* (arquivo LEI-2017-13416-dispositivos.json)


```
{
  "sentences": [
    {
      "text": {
        "content": "Fica autorizado o Banco Central do Brasil a adquirir papel-moeda e moeda metálica
fabricados fora do País por fornecedor estrangeiro, com o objetivo de abastecer o meio circulante nacional,
observado o disposto na Lei nº 8.666, de 21 de junho de 1993.",
        "beginOffset": 1
      }
    },
    {
      "text": {
        "content": "As aquisições referidas no caput obedecerão a cronograma fixado pelo Banco Central do
Brasil para cada exercício financeiro, observadas as diretrizes estabelecidas pelo Conselho Monetário
Nacional.",
        "beginOffset": 257
      }
    },
    {
      "text": {
        "content": "A inviabilidade ou fundada incerteza quanto ao atendimento, pela Casa da Moeda do
Brasil, da demanda por meio circulante ou do cronograma para seu abastecimento, em cada exercício
financeiro, caracteriza situação de emergência, para efeito de aquisição de papel-moeda e de moeda metálica
de fabricantes estrangeiros, na forma do inciso IV do caput do art. 24 da Lei nº 8.666, de 21 de junho de
1993.",
        "beginOffset": 461
      }
    },
    {
```


Fonte: Dados da pesquisa.

3 Métodos

Para gerar o texto em formato XML baseado no *schema* LexML, a partir do texto articulado de uma norma em formato RTF, foi utilizado o *Parser* LexML disponibilizado[6] pelo projeto LexML do Senado Federal.

6 Os fontes do parser LexML encontram-se em: <https://github.com/lexml/lexml-parser-projeto-lei>. Acesso em: 11 mar. 2018.

Para realizar o processamento da análise sintática das sentenças dos dispositivos das normas, optou-se por utilizar *framework* ou *API Cloud* oferecidos de forma gratuita.

Iniciou-se a pesquisa por *frameworks* e foi identificado que um dos mais lembrados *Parsers* sintáticos para a língua portuguesa é o Palavras (BICK, 2000), um analisador automático (*tagger-parser*) para português baseado em regras gramaticais, que foi desenvolvido por Eckhard Bick no contexto de um projeto de doutoramento (1994-2000) na Universidade de Århus (Dinamarca) e utilizado pelo projeto Floresta Sintática (FREITAS; ROCHA; BICK, 2008). O que impediu a utilização do Palavras neste trabalho foi o seu caráter não-gratuito.

Depois de mais pesquisas, optou-se por utilizar o SyntaxNet, um *framework* aberto e gratuito, oferecido pela Google, baseado em redes neurais e implementado sobre o TensorFlow[7]. O SyntaxNet fornece uma fundação para sistemas *Natural Language Understanding* (NLU) e provê um *Parser* sintático que pode ser treinado com corpus em CoNLL-U[8].

A Google oferecia um modelo pré-treinado para português brasileiro baseado no *corpus* da UD. O SyntaxNet, como qualquer framework baseado em *Deep Learning*, precisa de uma base de treinamento expressiva para alcançar níveis de acurácia satisfatórios. O *corpus* português-Br oferecido pela UD tinha em torno de 9.000 sentenças na época dos nossos processamentos. Os resultados (*dataset* 5) da análise sintática sobre as sentenças do *dataset* 4 usando o SyntaxNet com o *corpus* de treinamento UD não se mostraram muito promissores, em avaliação por amostragem de sentenças com características utilizadas com frequência em textos legislativos. Isso se deve basicamente à quantidade pequena de sentenças da base da UD utilizada no treinamento.

Durante o período de testes com o SyntaxNet, a Google lançou uma versão beta da sua *Cloud Natural Language API* para o idioma português brasileiro (anteriormente, só estava disponível para inglês, francês, japonês e espanhol). A Google oferecia um crédito por conta de usuário para testar sua *API Cloud*. Foi calculado que, com o volume de sentenças do *dataset* 4, o limite de crédito de apenas uma conta seria suficiente para o processamento de todas as sentenças. Mesmo tendo saldo suficiente para todas as sentenças, para economizar nas chamadas à *API* da Google, foi criada uma estrutura de *cache*, aplicando-se um *Message-Digest algorithm 5* (MD5) em cada sentença com o objetivo de não submeter sentenças repetidas à *API*. Foram encontradas 26.795 repetições de sentenças, uma economia de 21% no total das potenciais chamadas.

Os resultados (*dataset* 6) da análise sintática sobre as sentenças do *dataset* 4 usando a *API Cloud* da Google se mostraram promissores,

7 O TensorFlow é uma biblioteca de código aberto desenvolvida pela Google para computação numérica usando grafos de dados e que vêm sendo amplamente utilizada na área de aprendizado de máquina.

8 Uma visão geral do SyntaxNet encontra-se em: <https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html>. Acesso em: 11 mar. 2018.

baseando-se na mesma avaliação por amostragem realizada em relação ao processamento com o SyntaxNet usando o *corpus* da UD.

4 Limitações encontradas

A massa de dados de entrada nunca foi formatada tendo em vista o processamento automático da estrutura via *Parser*. Por isso, durante os processamentos, foram encontradas situações em normas específicas que impediram a geração do texto estruturado em XML, segundo *schema* LexML.

Na produção do *dataset* 2 (formato LexML), 998 normas de entrada do *dataset* 1 (em torno de 7%) não tiveram arquivos produzidos no *dataset* 2, pelo fato de o *Parser* LexML não ter reconhecido de forma correta a articulação do texto da norma de entrada. Um desses casos é a Lei 10.406, de 2002 ("Novo Código Civil") (BRASIL. Presidência da República, 2002), que, apesar de recente, apresenta técnica legislativa incompatível com a Lei Complementar 95, de 1997 (BRASIL. Presidência da República, 1997), na qual o *Parser* LexML é baseado.

5 Conclusão

Seguindo a ideia promovida pela OGD de tornar públicos os dados do governo para promover transparência e contribuir com o progresso em pesquisas baseadas nessas informações (VICTORINO et al., 2017), este trabalho permitiu oferecer oito *datasets* de forma aberta contendo textos originais e processados com análise sintática de normas federais brasileiras – os *datasets* estão publicados na plataforma *Figshare* de repositórios digitais especializados em pesquisas acadêmicas[9].

No intuito de contribuir com cientistas da informação e pesquisadores em geral, esses *datasets* podem ser entendidos como uma plataforma de dados para trabalhos futuros, como por exemplo na construção de um extrator de definições ou de tipificação de remissões, que, em vez de utilizar como entrada o texto original de cada norma, poderia utilizar por exemplo o *dataset* 6 como entrada, que contém a análise sintática da sentença de cada dispositivo.

Conforme apresentado no seção 2, adotou-se como recorte metodológico o período entre 4 de outubro de 1946 e 12 de abril de 2017. Uma possibilidade de novas pesquisas seria atualizar os *datasets* com normas de um período de publicação ampliado, como por exemplo, a partir de 1824, ano em que foi outorgada a Primeira Constituição Brasileira.

Outro filtro metodológico utilizado foi o de considerar apenas as normas federais do tipo Lei (ordinária) ou Lei Complementar. Trabalhos futuros poderiam expandir os *datasets* para contemplarem mais tipos de norma.

[9] Disponível em: <https://doi.org/10.6084/m9.figshare.c.4029253.v1>. Acesso em: 11 mar. 2018.

Gerar um *dataset* das sentenças dos dispositivos em formato CoNLL-U a partir do *dataset* 6, que contém um *json* de análise sintática da Google para cada dispositivo, poderia ser uma contribuição para a formação de um *corpus* UD com foco em textos legislativos. Seria necessária uma revisão, por pessoas com experiência em linguística, das estruturas sintáticas geradas.

Como são produzidas continuamente novas normas jurídicas através do Processo legislativo, trabalhos futuros poderiam implementar um mecanismo de atualização periódica ou contínua dos *datasets* produzidos neste trabalho.

O Processamento de linguagem natural tem evoluído rapidamente nos últimos anos, muito pela aplicação de técnicas de inteligência artificial. Por exemplo, a extração da semântica de textos vem se utilizando até o momento de passos intermediários de análise sintática para alcançar uma acurácia satisfatória. Por outro lado, recentemente a Google publicou o trabalho de criação do SLING (RINGGA- ARD; GUPTA; PEREIRA, 2017), um *Parser* para anotação semântica em textos baseado em redes neurais que, partindo apenas dos *tokens* textuais de entrada, produz grafos de *frames* semânticos sem qualquer representação simbólica interveniente. Ainda é cedo para conclusões definitivas, mas a análise sintática de textos e sentenças para fins de extração semântica pode se tornar um passo desnecessário. Assim, trabalhos futuros poderiam acompanhar a evolução dessas técnicas para avaliar se alguns dos *datasets* produzidos neste trabalho ainda serão relevantes e se outros precisarão ser produzidos.

6 Agradecimentos

Agradecemos as contribuições substanciais de Daniel Viero, Fabrício Santanna, Flávio Heringer, Jideão Vieira Filho e Wagner Teixeira. Trabalho produzido pelo Grupo de Estudos e Pesquisas Acadêmicas do Instituto Legislativo Brasileiro do Senado Federal (ILB/SF), edital COESUP 004/2016.

Referências

AUER, S. et al. Dbpedia: a nucleus for a web of open data. In: INT'L SEMANTIC WEB CONFERENCE, 16., 2007*,* Busan, Korea. *Procedings…* [S.l.]: Springer, 2007. p. 11-15.

BATISTA, A. H. et al. *Extração automática de definições*: um estudo de caso em textos legislativos. Brasilia: Universidade Católica de Brasília, 2011.

BICK, E. The parsing system palavras. *Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, University of Arhus, 2000.

BICK, E. The parsing system palavras. *Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, 2000. 505f. Tese

(Doutorado em Linguística) - Department of Linguistics, University of Arhus, Arhus, Dinamarca, 2000.

BRASIL. Lei Complementar nº 95, de 26 de fevereiro de 1998. Dispõe sobre a elaboração, a redação, a alteração e a consolidação das leis, conforme determina o parágrafo único do art. 59 da Constituição Federal, e estabelece normas para a consolidação dos atos normativos que menciona. *Diário Oficial da República Federativa do Brasil*, Brasília, 27 fev. 1998.

BRASIL. Lei nº 10.406, de 10 de janeiro de 2002. Institui o Código Civil. *Diário Oficial da República Federativa do Brasil*, Brasília, 11 jan. 2002.

BRASIL. Lei nº 13.416, de 23 de fevereiro de 2017. Autoriza o Banco Central do Brasil a adquirir papel-moeda e moeda metálica fabricados fora do País por fornecedor estrangeiro. *Diário Oficial da República Federativa do Brasil*, Brasília, 24 fev. 2017.

BUCHHOLZ, S.; MARSI, E. Conll-x shared task on multilingual dependency parsing. In: CONFERENCE ON COMPUTATIONAL NATURAL LANGUAGE LEARNING, 10., Nova York, Estados Unidos. 2006. *Procedings…* [S.l.]: Association for Computational Linguistics, 2006. p. 149-164.

FREITAS, C.; ROCHA, P.; BICK, E. Um mundo novo na floresta sintática: o treebank do português. *Calidoscópio*, v. 6, n. 3, p. 142-148, 2008.

GRAY, J. *Towards a genealogy of open data*. 2014. Disponível em: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2605828>. Acesso em: 6 de dez. de 2018.

HOHFELD, W. N. *Os conceitos jurídicos fundamentais aplicados na argumentação judicial*. Tradução de Margarida Lima Rego. Lisboa: Fundação Calouste Gulbenkian, 2008. 192 p.

LIMA, J. A. d. O. Apuração do texto original da lei geral de orçamento (Lei n. 4.320/64): um estudo de caso sobre a acurácia de bases de dados de legislação federal. *Boletim de Direito Administrativo*, São Paulo, v. 29, n. 1, p. 1-12, 2013.

NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, London, v. 18, n. 5, p. 544-551, 2011.

NAKAMURA, M.; OGAWA, Y.; TOYAMA, K. Extraction of legal definitions from a japanese statutory corpus–toward construction of a legal term ontology. In: LAW VIA THE INTERNET CONFERENCE, 2013, Jersey, Reino Unido. *Proceedings …*. [S.l.: s.n.], 2013. p. 1-11.

NIVRE, J. *et al*. Universal dependencies: a multilingual treebank collection. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION LREC 2016, 10., 2016, Portoroz, Eslovênia. *Proceedings …* [S.l.]: European Language Resources Association, 2016. p. 1659-1666.

RIBEIRO, C. J. S.; PEREIRA, D. V. A publicação de dados governamentais abertos: proposta de revisão da classe sobre previdência social do

vocabulário controlado do governo eletrônico. *Transinformação*, Campinas, v. 27, n. 1, p. 73-82, 2015.

RINGGAARD, M.; GUPTA, R.; PEREIRA, F. C. N. SLING: a framework for frame semantic parsing. *CoRR*, abs/1710.07032, 2017. Disponível em: <http://arxiv.org/abs/1710.07032>. Acesso em: 6 de dez. de 2018.

SANTOS NETO, A. L. dos et al. Tecnologias de dados abertos para interligar bibliotecas, arquivos e museus: um caso machadiano. *Transinformação*, Campinas, v. 25, n. 1, p. 81-87, 2013.

VICTORINO, M. de C. et al. Uma proposta de ecossistema de big data para a análise de dados abertos governamentais conectados. *Informação & Sociedade*, João Pessoa, v. 27, n. 1, p. 225-242, 2017.