Revista brasileira de iniciação científica
ISSN: 2359-232X

OS LIMITES DA PALAVRA E DA SENTENÇA NO PROCESSAMENTO AUTOMÁTICO DE TEXTOS

WORD AND SENTENCE BOUNDARIES IN AUTOMATIC TEXT PROCESSING

LOS LÍMITES DE LA PALABRA Y DE LA FRASE EN EL PROCESAMIENTO AUTOMÁTICO DE TEXTOS

Tatiana Cavalcanti[1]
Aline Silveira[2]
Elvis de Souza[3]
Cláudia Freitas[4]

Resumo: Este trabalho tem como objetivo apresentar os principais desafios linguísticos envolvidos na etapa de pré-processamento de um corpus composto por teses e dissertações da área de petróleo e gás. Como resultado, além do levantamento de questões específicas do domínio e de textos técnico-científicos, medimos o quanto o tratamento destas mesmas questões dificulta o processamento automático, e disponibilizamos para a comunidade de Processamento de Linguagem Natural de língua portuguesa um corpus padrão-ouro no que se refere apenas a tokenização e sentenciação, repleto de casos complexos, que serve para avaliação de métodos automáticos de segmentação, contribuindo também para a qualidade das etapas posteriores de processamento.
Palavras-chave: Processamento de Linguagem Natural. Linguística Computacional. Pré-processamento. Tokenização. Sentenciação.

Abstract: This paper aims to explore the major linguistic challenges involved in the preprocessing of a corpus composed of theses and dissertations from the Oil and Gas domain. Besides posing specific questions related to this domain and to scientific texts, we measured to which extent dealing with these matters hinders the automatic processing. We built a gold standard corpus of tokenization and sentence segmentation comprising several difficult cases, which are now available to the Portuguese Natural Language Processing community. This corpus can be used to evaluate automatic tokenization methods, as well as to improve the quality of subsequent steps in processing.
Keywords: Natural Language Processing. Computational linguistics. Preprocessing. Tokenization. Text segmentation.

[1] Graduanda em Licenciatura em Letras. PUC-Rio. E-mail: tatiana.shc@hotmail.com. Orcid: https://orcid.org/0000-0002-4378-5851
[2] Graduanda em Licenciatura em Letras. PUC-Rio. E-mail: silveira26aline@gmail.com. Orcid: https://orcid.org/0000-0002-4742-3014
[3] Graduando em Bacharelado em Letras. PUC-Rio. E-mail: elvis.desouza99@gmail.com. Orcid: https://orcid.org/0000-0001-9373-7412
[4] Doutora em Letras e professora na PUC-Rio. PUC-Rio. E-mail: claudiafreitas@puc-rio.br. Orcid: http://orcid.org/0000-0001-6807-8558

Revista brasileira de iniciação científica
ISSN: 2359-232X

Resumen: Este trabajo tiene como objetivo presentar los principales desafíos lingüísticos involucrados en la etapa de preprocesamiento de un corpus compuesto por tesis y disertaciones en el área de petróleo y gas. Como resultado, además de plantear cuestiones específicas relacionadas con este dominio y con textos científicos, determinamos en qué medida el tratamiento de estas mismas cuestiones dificulta el procesamiento automático. Creamos un corpus estándar de oro de tokenización y segmentación que comprende varios casos difíciles, que ahora están disponibles para la comunidad de Procesamiento del Lenguaje Natural de lengua portuguesa. Este corpus se puede utilizar para evaluar los métodos de tokenización automática, así como para mejorar la calidad de los pasos posteriores del procesamiento.

Palabras llave: Procesamiento del Lenguaje Natural. Lingüística computacional. Preprocesamiento. Tokenización. Segmentación.

Submetido 22/02/2021 **Aceito 30/05/2021** **Publicado 15/10/2021**

Revista brasileira de iniciação científica
ISSN: 2359-232X

Introdução

Processamento de Linguagem Natural (PLN) é o nome que se dá ao campo que estuda o processamento automático de línguas humanas por computadores. O PLN pode ser feito, de modo geral, com base em regras linguísticas, aprendizado de máquina, ou ainda utilizando métodos híbridos. O aprendizado de máquina requer uma quantidade considerável de dados com indicações consistentes do que o modelo deve aprender para que possa desempenhar suas tarefas adequadamente. No caso do aprendizado voltado para o PLN, portanto, devemos dispor de corpora *anotados* (ou *datasets*), isto é, conjuntos de textos anotados com informação linguística.

Embora para línguas como o inglês e para alguns gêneros textuais os resultados de tarefas básicas de PLN sejam em grande parte satisfatórios, quando há diferenças entre o tipo de texto que alimenta o treinamento de algoritmos e o tipo de texto onde eles são aplicados os resultados pioram consideravelmente – enquanto um anotador de classes gramaticais (Kazama, 2001) obtém desempenho de 96,84% na anotação de um corpus de textos jornalísticos, a mesma tarefa vê seu desempenho cair para 83,5% em um corpus de resumos de artigos científicos da área da biomedicina (THOMPSON et al., 2017). Nesse sentido, uma solução para o processamento de textos em um domínio específico envolve preparar, com quantidade e qualidade, um corpus desse mesmo domínio para que a máquina seja capaz de generalizar o aprendizado a partir de textos semelhantes.

Este trabalho relata os desafios e as soluções encontradas durante o processo de desenvolvimento de um corpus padrão-ouro no que se refere às etapas de tokenização e sentenciação. O corpus, chamado PetroTok, foi produzido no âmbito do projeto BIG Oil[5] – Ciência de Dados para Óleo e Gás, que inclui a construção de um corpus de teses e dissertações pertencentes à área de petróleo e gás em língua portuguesa – o corpus Petrolês. Além de relatar os desafios e soluções linguísticas, avaliamos o grau de dificuldade deste tipo de tarefa, comparando os resultados da análise humana com os da análise automática efetuada pela ferramenta UDPipe (Straka, 2016), anotador morfossintático baseado em aprendizado de

[5] O projeto é uma colaboração entre o grupo ComCorHd (Linguística Computacional, Corpus e Humanidades Digitais), do Departamento de Letras, e o ICA (Laboratório de Inteligência Computacional Aplicada), do Departamento de Engenharia Elétrica, ambos da PUC-Rio.

Revista brasileira de iniciação científica
ISSN: 2359-232X

máquina do projeto Universal Dependencies (Nivre, 2016), que prevê um formato de anotação gramatical unificado para diversas línguas. Assim, disponibilizamos um material de qualidade para a avaliação de tokenização e sentenciação.

O Laboratório de Inteligência Computacional Aplicada (ICA) da Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) desenvolveu para o projeto, em 2019, uma ferramenta de pré-processamento de textos no formato PDF que realiza a conversão de artigos acadêmicos para o formato TXT. Essa conversão considera todas as alterações necessárias para que o formato original do documento, como formatação e imagens, não impacte negativamente a legibilidade do texto quando convertido para o formato de texto plano, questões essas que já foram exploradas em Silveira et al. (2019).

Tendo como horizonte a construção de um sistema de buscas semanticamente orientado para o domínio de petróleo e gás, o Petrolês deverá contar com diversas camadas de anotação linguística, como classes gramaticais, dependências sintáticas e entidades mencionadas específicas do domínio, como tipos de rochas e localizações geográficas. No entanto, para lidar com essas informações linguísticas, é necessário que etapas anteriores tenham sido concluídas satisfatoriamente.

A tokenização – isto é, a identificação de unidades linguísticas (*tokens*) – e a sentenciação – a delimitação das sentenças – são tarefas com bons resultados no PLN, tendo obtido 99,54% e 89,24%, respectivamente, para a língua portuguesa utilizando o sistema UDPipe (Straka, 2016) tomando por base um corpus de textos jornalísticos. No entanto, verificamos uma qualidade inferior de segmentação automática em um corpus como o nosso, de textos acadêmicos e de um domínio técnico específico.

O "gabarito" que construímos se constitui de casos difíceis já solucionados por nós e que, por isso, serve para avaliação de métodos de tokenização e sentenciação automáticos, contribuindo consequentemente para a qualidade das etapas posteriores do projeto. Durante a confecção desse material, organizamos diretivas com critérios de tokenização e sentenciação cuidadosamente debatidas pela equipe de anotação e que serão exibidas ao longo deste trabalho.

É interessante notar que, além de se apresentarem como desafio para o processamento automático, os tópicos aqui discutidos também suscitam reflexões linguisticamente ricas, como a própria noção do que podemos considerar uma palavra (BIDERMAN, 2001). Assim, a

construção do Petrolês, além de possibilitar um avanço no PLN para língua portuguesa em termos de material para avaliação de métodos, também fomenta a discussão de diferentes perspectivas para se estudar a língua de forma empírica, como sugere Sampson (2001).

O artigo está estruturado da seguinte maneira: na próxima seção, apresentamos os conceitos-chave relativos à etapa de pré-processamento linguístico de um corpus: sentenciação e tokenização. Em seguida, trazemos trabalhos relacionados ao nosso, que ilustram algumas formas de lidar com as questões levantadas na seção anterior. Na seção Metodologia, que se divide em duas partes, descrevemos as etapas envolvidas na criação do corpus PetroTok: em primeiro lugar, detalhamos os procedimentos para seleção das frases que constituem o corpus; em segundo lugar, descrevemos os principais problemas relativos à segmentação e à tokenização, bem como as soluções tomadas para cada caso. Na seção Resultados e Avaliação, fazemos uma apresentação quantitativa do corpus criado, e comparamos os resultados do material antes e depois da nossa intervenção. Por fim, na última seção, delineamos nossas considerações finais.

Pré-processamento, Sentenciação e Tokenização

Quando se trata de um trabalho de anotação de corpora, um fato essencial a ser considerado é a existência de etapas anteriores à anotação e à revisão em si. Em primeiro lugar, deve-se lembrar que o material deve ser pré-processado, momento em que se descartam elementos indesejáveis para a análise morfossintática que se seguirá, como imagens, tabelas, cabeçalhos e links. No caso específico de textos do gênero acadêmico, como teses, dissertações, e monografias, a presença de listas itemizadas, figuras e tabelas dificultam a leitura da ferramenta computacional e rompem com a linearidade do texto – este ponto já foi abordado em Silveira et al. (2019). Em seguida, assumindo que se tenha um texto limpo e pronto para ser segmentado e analisado em níveis mais complexos, deve haver a segmentação dos elementos em sentenças e em tokens (HEARST, 1999).

Uma vez com o texto pré-processado, fazemos a sua sentenciação, que é a delimitação do texto corrido em sentenças. A tokenização, por sua vez, é o processo de segmentar uma sentença já delimitada em *tokens*, as unidades linguísticas mínimas do processamento automático. Muitas vezes, o *token* corresponde ao que o senso comum entende como palavra –

Revista brasileira de iniciação científica
ISSN: 2359-232X

à exceção de alguns poucos casos, como as pontuações, símbolos como o de porcentagem e contrações[6].

No entanto, a despeito de um senso comum sobre o que sejam palavras, delimitá-las não é tarefa fácil. Biderman (2001), por exemplo, aponta para as fragilidades dos três critérios comumente utilizados por linguistas: o critério fonológico, o morfossintático (gramatical) e o semântico. Para a autora, sozinhos, nenhum dos três critérios consegue delimitar todas as palavras. No PLN, o critério usado na identificação de palavras costuma ser o critério gráfico, observando a presença de espaços vazios entre cadeias de caracteres, uma vez que oferece uma pista formal e confiável no que se refere à delimitação das unidades. Contudo, como será mostrado, mesmo esse critério pode não ser suficiente.

Um texto mal tokenizado traz, de fato, consequências negativas para a anotação linguística, já que tokenização, sentenciação e anotação são etapas intimamente relacionadas: uma vez decidido o que contará como uma palavra (ou frase), é a esta palavra que será atribuída uma classe gramatical, como pode ser visto em 1.

Figura 1: Tokenização e classificação gramatical da frase “A menina viajou.”

id	token	pos
1	A	DET
2	menina	NOUN
3	viajou	VERB
4	.	PUNCT

Fonte: Elaboração própria, 2020.

No exemplo da Figura 1, é possível delimitar claramente quais são as unidades mínimas e atribuir-lhes a devida anotação gramatical. Já em exemplos retirados do nosso corpus, encontramos sentenças como a da Figura 2.

[6] No modelo com o qual trabalhamos, a palavra “da”, por exemplo, deve ser tokenizada como duas unidades: a preposição “de” e o artigo “a”

Revista brasileira de iniciação científica
ISSN: 2359-232X

Figura 2: Exemplo de frase com compostos químicos

```
# sent_id = 107-20121015-TESEMSC_0_resumo-13
# text = As espécies de Hidrocarbonetos Policíclicos Aromáticos
encontrados foram: naftaleno, 1 Metilnaftaleno, Bifenil, Acenafteno,
Fluoreno, Dibenzothiofeno, Fenantreno, Antraceno, Fluoranteno, Pireno,
Benzo(a)antraceno, Criseno, Benzo(b)fluoranteno, Benzo(k)fluoranteno,
Benzo(a)pireno, Perileno, Benzo(e)pireno, Indeno(1,2,3-cd)pireno,
Dibenzo(a,h)antraceno e Benzo(g.h.i)perileno.
```

Fonte: PetroTok. A frase pode ser encontrada no corpus pelo identificador *sent_id*.

Nesse caso, ao olharmos para os compostos químicos, a definição das fronteiras linguísticas se mostra menos óbvia. Na Figura 3, observa-se uma entre muitas formas de se tokenizar “Indeno(1,2,3-cd)pireno”.

Figura 3: Possível tokenização do composto químico Indeno(1,2,3-cd)pireno

id	token
1	Indeno(
2	1,2,
3	3-cd
4	)
5	pireno

Fonte: Elaboração própria, 2020.

Apesar de possível, a tokenização ilustrada traz complicações, pois, pensando na anotação linguística, surgem dúvidas sobre a classe gramatical de, por exemplo, “1,2,” ou “3-cd”, assim como sobre a sua futura anotação sintática.

Dessa forma, as dificuldades impostas pelo gênero acadêmico e pelo domínio específico de petróleo motivaram as decisões acerca do pré-processamento, da sentenciação e da tokenização deste corpus. Com isso, o material aqui apresentado dispõe de uma série de padrões que podem vir a ser aplicados também em trabalhos de natureza semelhante.

Na seção seguinte, serão brevemente apresentados alguns trabalhos de pesquisa inseridos na área de PLN que também abrangem os desafios e estratégias correspondentes relacionados ao tratamento dos textos na construção de um corpus.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Trabalhos relacionados

Grefenstette & Tapanainen (1994) discutem a tokenização como um problema para a lexicografia computacional. Eles apresentam objetivos e métodos de tokenização, além de padrões linguísticos para reconhecer acrônimos, abreviações, números e datas a partir de expressões regulares. Os autores apresentam os problemas encontrados no processo e discutem os efeitos de escolhas aparentemente inocentes. A presença de hífen na quebra de linha – situação com a qual também nos deparamos – é um dos problemas enfrentados por eles e resolvidos com uma única regra, que juntava todas as palavras separadas por hífen, eliminando-o juntamente com o espaço que a ele às vezes se sucede. Para os autores, os casos em que as regras não se aplicavam corretamente eram deixados para serem resolvidos posteriormente ou simplesmente ignorados, sendo aceitos como ruído no sistema.

O trabalho de Lopes & Vieira (2013), também voltado para a constituição de um corpus de documentos técnico-científicos, enfrentou igualmente problemas relativos à tokenização. As resoluções relatadas no trabalho seguem um direcionamento similar ao que foi feito no PetroTok: em grande parte, regras feitas com expressões regulares para tratamento da segmentação. No entanto, o trabalho se diferencia do nosso em relação a algumas escolhas feitas. As autoras relatam, por exemplo, modificações na formatação de números (“1.927” se torna “1927”; “3,1415” se torna “31415”), o que não precisou ser feito por nós, porque este não se mostrou um problema. Além disso, as referências no meio do texto – tão presentes no gênero acadêmico e desafiadoras para nós – não foram tratadas mas, antes, retiradas do texto, o que não fizemos, já que a ideia é manter ao máximo a integridade dos documentos.

Ainda no âmbito da construção de corpora, Freitas & Afonso (2007) documentaram o resultado das discussões conjuntas entre os membros do projeto de construção da Floresta Sintá(c)tica (SANTOS et al., 2007). Nesse documento, as autoras registram todas as decisões linguísticas tomadas, desde a definição de palavra até categorias morfossintáticas. O formato utilizado serviu como exemplo para a produção das nossas diretivas acerca da tokenização e da sentenciação.

Sanchez (2019) trata da tarefa de detecção de sentenças em um domínio específico: o jurídico. Esse domínio exige olhar próprio especialmente por causa da variedade de pontuação, estrutura e sintaxe do texto legal. O trabalho descreve essas peculiaridades e avalia três

abordagens para a tarefa de sentenciação. Isso mostra que, mesmo sendo uma tarefa antiga na área de PLN, a detecção de fronteiras de sentenças continua, em 2019, sendo uma questão relevante. Além disso, cada domínio precisará identificar seus próprios problemas, bem como desenvolver suas soluções.

No contexto de diálogo entre PLN e Humanidades Digitais – área que une ciências humanas e recursos digitais com o objetivo de elaborar novos pontos de vista e métodos para o primeiro campo –, Rocha et al. (2019) descrevem correções feitas no corpus Obras Brasileiras (OBras) (Santos et al., 2018), composto de obras literárias brasileiras em domínio público, a fim de obter buscas com resultados mais precisos, sendo algumas dessas correções relacionadas à má segmentação dos textos. De fato, o projeto conta com uma série de instruções (muitas delas relacionadas ao pré-processamento) que devem ser seguidas pelos colaboradores do projeto para que uma determinada obra seja incluída[7]. Isso mostra que a segmentação não é um desafio apenas para o processamento de textos técnico-científicos, mas também de obras literárias.

Manning & Schütze (1999), ao falarem sobre o uso de métodos heurísticos – isto é, que visam solucionar um problema específico dividindo-o em partes menores – para a detecção de fronteiras de sentenças, afirmam que esse tipo de método requer conhecimento do tipo de texto em questão. Por isso, é compreensível que cada gênero textual, ou tipo de texto, tenha seus próprios desafios.

Dessa forma, a apresentação dos trabalhos selecionados visa tornar mais fácil que se compreenda a dimensão do desafio relacionado ao tratamento de um texto que integrará um corpus. Considerando as soluções propostas pelos autores aqui discutidas, a seção seguinte tratará da metodologia da pesquisa.

Metodologia

Esta é uma pesquisa aplicada, que tem como objetivo final a construção de um recurso para o PLN de língua portuguesa. Nesta seção, descrevemos o processo de criação do corpus e,

[7]As instruções de segmentação encontram-se em: https://www.linguateca.pt/OBRAS/instrucoes_obras.html. Acesso em: 7 jan. 2021.

em seguida, relatamos os principais desafios trazidos por este material no que se refere à sentenciação e à tokenização, bem como as soluções dadas para cada caso.

O presente trabalho foi feito em quatro etapas: (i) familiarização, análise de problemas e documentação das soluções; (ii) anotação automática de uma grande amostra do corpus com a ferramenta UDPipe; (iii) busca por casos complexos; (iv) correção da análise automática com auxílio da ferramenta Interrogatório, construindo um corpus padrão-ouro apenas com casos complexos de tokenização e sentenciação.

A primeira etapa foi a de familiarização com as peculiaridades dos textos, tanto por serem do gênero acadêmico quanto pelo domínio específico. Analisamos os textos, discutimos quais seriam as melhores soluções para possíveis desafios e construímos uma documentação de como delimitar sentenças e tokenizar as palavras. Na documentação, a preocupação central é a consistência: mesmos fenômenos devem receber o mesmo tratamento.

Uma vez tendo sido apresentados às especificidades do texto, experimentamos realizar a sentenciação e tokenização de uma amostra de 400 textos pela ferramenta de anotação automática UDPipe (Straka, 2016) para guiar as nossas correções a partir dos casos que o próprio sistema não conseguiu segmentar corretamente. Então, utilizando a bibliografia apresentada (Grefenstette & Tapanainen, 1994; Lopes & Vieira, 2013; Freitas & Afonso, 2007; Sanchez, 2019; Rocha et al., 2019), buscamos os casos clássicos de segmentação complexa, como a ocorrência de parênteses, ponto e vírgula e títulos de seções. Realizamos a busca utilizando a ferramenta Interrogatório (de Souza & Freitas, 2019) – sistema de buscas linguisticamente orientadas em corpora no formato CoNLL-U[8].

Por fim, tendo encontrado um número considerável de frases complexas segmentadas incorretamente, lançamos mão de funcionalidades do Interrogatório que permitem realizar a tokenização e a sentenciação manualmente, agilizando a correção e garantindo uniformidade em ações que envolviam a inclusão ou eliminação de tokens e a junção ou separação de sentenças no material que viria a se tornar o nosso padrão-ouro PetroTok.

A seguir, serão discutidos os problemas centrais relacionados à segmentação do texto na construção do corpus e as devidas estratégias desenvolvidas pela equipe para lidar com eles.

[8] CoNLL-U é o formato utilizado para codificar os metadados e anotações linguísticas dentro do projeto Universal Dependencies, descrito em <https://universaldependencies.org/format.html>. Acesso em 2 de jun. 2021.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Discussão dos principais problemas

- **Hifenização**

Um dos primeiros problemas com os quais nos deparamos, também mencionado em Grefenstette & Tapanainen (1994), foi relacionado a palavras com hífen nas quebras de linha, o que chamamos de *hifenização*. O hífen, como ficou decidido pela equipe, não se configura como critério de separação de palavras, de modo que uma palavra como "auto-sustentável" conta como uma única unidade. Contudo, percebemos que, na maior parte dos casos, as palavras hifenizadas não eram palavras compostas, mas palavras que receberam hífen devido à quebra de linha no documento original, isto é, se deviam à separação de sílabas. Isso acontecia de tal maneira que a palavra era dividida em duas, como em "subloca- ram", "per- dem", "po- voados", entre outros. Nesse sentido, foram desenvolvidas estratégias para lidar com a hifenização, ilustradas na Figura 4.

Figura 4: Exemplos de mudança na hifenização

Item	Como estava no corpus	Outras ocorrências no corpus	Como ficou
a)	po- voados	povoados	povoados
b)	mato- grossense	mato-grossense	mato-grossense
c)	infra- estrutura	infra-estrutura infraestrutura	infraestrutura
d)	metil- acetato	nenhuma	metil-acetato

Fonte: Elaboração própria, 2020.

Em primeiro lugar, fizemos uma busca em todas as palavras do corpus e as comparamos com as palavras que apresentavam hífen seguido de espaço. Caso fosse encontrada uma ocorrência no corpus em que a palavra tivesse sido escrita sem o hífen e sem o espaço, o sistema automaticamente retirava esses dois caracteres e, assim, a palavra se tornava uma só (item *a* da figura). Se o sistema encontrasse um exemplo da palavra apenas com o hífen, sem o espaço,

Revista brasileira de iniciação científica
ISSN: 2359-232X

esse último era removido da palavra (item *b* da figura). Na situação de encontrar um exemplo da mesma palavra escrita junta e um exemplo separada com hífen, prevalecia a escrita junta (item *c* da figura). Se, ao final, não houvesse uma palavra de comparação, retirava-se apenas o espaço da palavra e o hífen era mantido (item *d* da figura).

- **Compostos químicos**

Outra questão foi o caso dos compostos químicos, ocorrências típicas do domínio de petróleo. O problema reside no tratamento das pistas formais de tokenização nesses compostos e na diferença entre casos como *poli(bisfenol A-co-epicloridrina)* e *Dibenzo(a,h)antraceno*. Como regra geral da documentação para o Petrolês, sinais como o parênteses, a vírgula e o ponto final são tokens por si só, ou seja, não estão atrelados a nenhuma palavra. Assim, a divisão em tokens de *poli(bisfenol A-co-epicloridrina)* seria realizada do seguinte modo:

Figura 5: Tokenização de “poli(bisfenol A-co-epicloridrina)”

id	token
1	poli
2	(
3	bisfenol
4	A-co-epicloridrina
5	)

Fonte: Elaboração própria, 2020.

Entretanto, casos como *Dibenzo(a,h)antraceno*, *Indeno(1,2,3-cd)pireno* e *Benzo(g.h.i)perileno* não poderiam ser segmentados da mesma maneira, pois a falta de espaço funciona como uma pista gráfica para que cada um deles seja tratado como um token só, diferentemente do caso de *poli(bisfenol A-co-epicloridrina)*, no qual há um espaço entre *bisfenol* e *A-co-epicloridrina,* permitindo a separação em tokens distintos.

- **Citações no corpo do texto**

Além disso, outro fenômeno precisou de atenção especial: a ocorrência de citações no corpo do texto, especificamente a expressão “et al.”, como em “ALLINGER et al., 1978”. Nesse caso, o ponto que marca a abreviação faz parte da expressão. Desse modo, decidiu-se por realizar a divisão em dois tokens “et” e “al.” Em outros casos, como na frase “De acordo com Kennish (1997), ovos, larvas e estágios juvenis são mais sensíveis ao óleo”, ao nos depararmos com a citação no corpo do texto, foi preciso decidir se estávamos diante de uma unidade apenas, ou de quatro unidades. Seguindo o critério de deixar o pré-processamento o mais transparente possível, optamos pela separação em tokens distintos, como na Figura 6.

Figura 6: Tokenização de “Kennish (1997)”

id	token
1	Kennish
2	(
3	1997
4	)

Fonte: Elaboração própria, 2020.

Na Figura 6, encontramos a tokenização de “Kennish (1997)”, que ilustra a nossa escolha do parênteses como um token próprio. Se optássemos por tratar “(1997)” como uma unidade, por exemplo, estaríamos reforçando uma ideia do parênteses como algo colado à informação trazida dentro dele, e não como um símbolo de pontuação.

- **Unidades de medida e símbolos**

O tratamento de unidades de medida e símbolos também exigiu atenção especial. Quantas unidades temos em casos como “60 km/h”, “5 V” ou “50%”? A solução adotada foi

Revista brasileira de iniciação científica
ISSN: 2359-232X

baseada na presença ou não do espaço entre o número e a unidade ou símbolo. Por exemplo, 50% contaria como um token apenas, enquanto 60 km/h seria dividido em dois tokens, “60” e “km/h”.

- **Pontos finais, de exclamação e de interrogação como únicos delimitadores de sentença**

Quanto à sentença, os pontos finais, como o ponto propriamente, o ponto de interrogação e o de exclamação, foram considerados os únicos indicadores de fim de sentença. No entanto, após a anotação automática feita pelo UDPipe (Straka, 2016), foram encontrados casos em que pontuação diferente de ponto final, como vírgula, travessão, parênteses, dois pontos e ponto e vírgula apareciam no fim de sentença. Por exemplo, listas e enumerações (Figura 7) se revelaram potencialmente problemáticas, na medida em que o ponto e vírgula e os dois pontos eram tidos constantemente como delimitadores de sentença. Nesses casos, a decisão – devidamente documentada – é de que esses itens todos formem uma única sentença, como apresentado na coluna PetroTok (Figura 7).

Figura 7: A ocorrência do ponto e vírgula no arquivo em PDF

PDF	UDPipe	PetroTok
De acordo com CETESB (2002) as características mais relevantes em um derrame são: 1. Tipo e quantidade de petróleo, sendo os mais tóxicos os óleos leves devido à presença de uma quantidade maior de compostos aromáticos; 2. Amplitude de maré, podendo esta agravar o efeito do derrame ou mesmo contribuir para processo de limpeza; 3. Época do ano, por causar consideráveis variações na estrutura e composição das comunidades biológicas costeiras; (...) 9. Formas de limpeza aplicadas ao derrame.	**Frase 1:** De acordo com CETESB (2002) as características mais relevantes em um derrame são: **Frase 2:** 1. Tipo e quantidade de petróleo, sendo os mais tóxicos os óleos leves devido à presença de uma quantidade maior de compostos aromáticos; **Frase 3:** 2. Amplitude de maré, podendo esta agravar o efeito do derrame ou mesmo contribuir para processo de limpeza; **Frase 4:** 3. Época do ano, por causar consideráveis variações na estrutura e composição das comunidades biológicas costeiras; (...) **Frase 10:** 9. Formas de limpeza aplicadas ao derrame.	**Frase 1:** De acordo com CETESB (2002) as características mais relevantes em um derrame são: 1. Tipo e quantidade de petróleo, sendo os mais tóxicos os óleos leves devido à presença de uma quantidade maior de compostos aromáticos; 2. Amplitude de maré, podendo esta agravar o efeito do derrame ou mesmo contribuir para processo de limpeza; 3. Época do ano, por causar consideráveis variações na estrutura e composição das comunidades biológicas costeiras; (...) 9. Formas de limpeza aplicadas ao derrame.

Fonte: Elaboração própria, 2019.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Como ilustra a Figura 7, cada item da lista termina com ponto e vírgula. Contudo, o que ocorre é que as listas itemizadas – ou seja, com número ou qualquer outro marcador como círculos, traços, quadrados ou setas – possuem grande variação em relação à pontuação do fim dos itens da lista. Por isso, definimos qual seria a diretriz do pré-processamento em cada um dos casos. Nos eventos em que os itens são separados por ponto final, cada item forma uma sentença própria, visto que o ponto final é sempre delimitador de sentença. Nos eventos em que os itens são separados por ponto e vírgula, tem-se o caso da figura 7, ou seja, tornam-se todos uma só sentença.

- **Títulos e subtítulos de seções**

Outra situação frequente está relacionada à presença de títulos e subtítulos de seções. Visto que, em geral, estes não apresentam ponto final, essas sequências são encadeadas no processamento automático (UDPipe) como se fossem uma só, o que se observa na Figura 8. Esses casos são difíceis, pois não têm ponto final – a pista formal de delimitação de sentenças – mas, ainda assim, precisam estar separados. Esse tipo de ocorrência, muito comum no gênero acadêmico, é análogo ao caso das manchetes de jornal, presentes no gênero de notícias.

Figura 8: A ocorrência de títulos e subtítulos no PDF, o retorno automático do UDPipe e a versão corrigida no PetroTok

PDF	UDPipe	PetroTok
Essa retenção poderá viabilizar a utilização da membrana condutora protônica a maiores temperaturas, o que aumenta a eficiência das CCs. 3. Parte experimental 3.1. Materiais Em zeólitas onde os cátions de compensação de carga são prótons, aparecem grupos hidroxilas ponte em cada sítio AlO4-, isto é, sítios ácidos de Bronsted.	**Frase 1:** Essa retenção poderá viabilizar a utilização da membrana condutora protônica a maiores temperaturas, o que aumenta a eficiência das CCs.3. Parte experimental3.1. Materiais Em zeólitas onde os cátions de compensação de carga são prótons, aparecem grupos hidroxilas ponte em cada sítio AlO4-, isto é, sítios ácidos de Bronsted.	**Frase 1:** Essa retenção poderá viabilizar a utilização da membrana condutora protônica a maiores temperaturas, o que aumenta a eficiência das CCs. **Frase 2:** 3. Parte experimental. **Frase 3:** 3.1. Materiais. **Frase 4:** Em zeólitas onde os cátions de compensação de carga são prótons, aparecem grupos hidroxilas ponte em cada sítio AlO4-, isto é, sítios ácidos de Bronsted.

Fonte: Elaboração própria, 2019.

A Figura 8 apresenta uma sequência de parágrafos que inclui título e subtítulo retirada do corpus PetroTok. Na primeira coluna, vemos a organização original do texto em PDF. Na segunda, temos a saída automática feita pela ferramenta UDPipe. Por fim, na terceira se encontra a solução de segmentação do texto adotada no PetroTok.

- **Quando o ponto final não é delimitador de sentença**

Além disso, uma questão que exigiu cuidado foi a de pontos que não são o ponto final, mas que são reconhecidos automaticamente como tal. Por exemplo, o ponto encontrado nas expressões "etc." e "et al." não equivale a uma marcação de fim de sentença, e sim a uma abreviação. O mesmo ocorre na Figura 9, em que o sentenciador automático, reconhecendo o marcador de abreviação como ponto final, divide uma única sentença em várias, erroneamente.

Figura 9: A ocorrência do ponto marcador de abreviação no PDF, o retorno automático do UDPipe e a versão corrigida no PetroTok

PDF	UDPipe	PetroTok
Seus principais autores são, entre outros, Douglass North, James G. March, Johann P. Olsen, Paul J. DiMaggio e Walter W. Powell.	**Frase 1:** Seus principais autores são, entre outros, Douglass North, James G. **Frase 2:** March, Johann P. **Frase 3:** Olsen, Paul J. **Frase 4:** DiMaggio e Walter W. **Frase 5:** Powell.	**Frase 1:** Seus principais autores são, entre outros, Douglass North, James G. March, Johann P. Olsen, Paul J. DiMaggio e Walter W. Powell.

Fonte: Elaboração própria, 2020.

Por fim, ainda sobre os erros na identificação do ponto final, existem os pontos que acompanham números ou letras nas listas, como em "1." ou "ii.". Nesses casos, as sentenças também eram erroneamente delimitadas no ponto, de modo que o conteúdo listado passava a compor uma nova frase – o que pode ser visto na Figura 10.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Figura 10: A ocorrência de pontos que acompanham letras nas listas, o retorno automático do UDPipe e a versão corrigida no PetroTok

PDF	UDPipe	PetroTok
Para a produção de gás de síntese a partir do gás natural são utilizados os seguintes processos: i. Reforma a vapor; ii. Oxidação parcial; iii. Reforma a seco.	**Frase 1:** Para a produção de gás de síntese a partir do gás natural são utilizados os seguintes processos: i. **Frase 2:** Reforma a vapor; ii. **Frase 3:** Oxidação parcial; iii. **Frase 4:** Reforma a seco.	Para a produção de gás de síntese a partir do gás natural são utilizados os seguintes processos: i. Reforma a vapor; ii. Oxidação parcial; iii. Reforma a seco.

Fonte: Elaboração própria, 2020.

A partir desses conjuntos de instruções, pudemos uniformizar a tokenização e a sentenciação no corpus. A seguir, encontram-se os resultados e a avaliação das medidas tomadas em relação às questões descritas nesta seção.

Resultados e Avaliação

Como resultado do estudo das situações problema, criamos um corpus cuidadosamente revisto quanto a *sentenciação* e *tokenização,* composto por 1.140 sentenças e 115.140 tokens. O corpus não contém frases na sequência em que aparecem nos textos originais, mas uma seleção de frases que podem ser especialmente difíceis para o processamento automático, como ilustramos nas seções anteriores. O corpus está disponível na página do projeto[9] e pode ser utilizado para a avaliação de métodos automáticos de sentenciação e tokenização.

Por fim, como uma maneira de quantificar os desafios que relatamos aqui, comparamos o corpus padrão-ouro com o trabalho realizado pelo tokenizador automático embutido na ferramenta UDPipe (STRAKA, 2016). Os resultados da comparação podem ser encontrados na tabela 1.

[9] <http://petroles.ica.ele.puc-rio.br/>

Revista brasileira de iniciação científica
ISSN: 2359-232X

Como podemos observar, há alguma diferença com relação ao número de sentenças. os números são maiores no resultado do sistema, indicando a tendência de separar sentenças onde não é devido.

Tabela 1: Comparação entre a segmentação dos textos que realizamos e os resultados automáticos do UDPipe

	PetroTok	**UDPipe**
Sentenças	1.140	1.314
Sentenças sem verbo	137	241

Fonte: Elaboração própria, 2021.

Na tabela 2, medimos a precisão, a abrangência e a medida F (média harmônica) do desempenho do UDPipe na tarefa de sentenciação automática do texto original que deu base ao PetroTok[10]. Os cálculos foram realizados da seguinte forma:

- São verdadeiros positivos (VP) os casos de sentenças que o sistema identificou corretamente;
- São falsos negativos (FN) os casos de sentenças que não foram identificadas pelo sistema;
- São falsos positivos (FP) os casos de sentenças que foram identificadas pelo sistema mas que não deveriam ter sido;

Assim, precisão (P), abrangência (A) e medida F (F) seguem as equações (1)-(3):

$$P = \frac{VP}{VP + FP} \qquad (1)$$

[10] Não calculamos a tokenização porque alguns dos problemas relatados, especialmente a hifenização, já haviam sido resolvidos de forma automática.

Revista brasileira de iniciação científica
ISSN: 2359-232X

$$A = \frac{VP}{VP + FN} \qquad (2)$$

$$F = 2 \cdot \frac{P \cdot A}{P + A} \qquad (3)$$

Tabela 2: Avaliação da sentenciação automática desempenhada pelo UDPipe

	Precisão	**Abrangência**	**Medida F**
Sentenciação	71,8%	82,5%	76,8%

Fonte: Elaboração própria, 2021.

A análise da Tabela 2 nos permite interpretar que o sistema peca mais pelo excesso do que pela falta: ele tende a segmentar demais, o que lhe permite encontrar por volta de 80% das sentenças corretas (abrangência), mas também errar outros 30% de sentenças (precisão), muitas das quais não deveriam existir.

Considerações finais

Para que seja viável a construção de um corpus anotado de qualidade deve-se primeiro considerar uma série de cuidados iniciais que preparem os textos para serem processados. O pré-processamento, apesar de ser comumente relegado a uma categoria de pouca importância, merece atenção, e uma preocupação com essa etapa garante um melhor andamento na sequência de etapas de qualquer projeto de PLN. Assim, a conjugação de uma boa tokenização e uma boa sentenciação se destaca como algo essencial e não trivial.

Como foi descrito ao longo do texto, preparamos – e disponibilizamos publicamente – um material revisado e corrigido em termos de tokenização e sentenciação de textos do gênero acadêmico, em português, do domínio de petróleo. Comparamos nosso padrão-ouro com os

resultados de um sistema automático e, assim, asseguramos que o material que preparamos apresenta, de fato, melhorias consideráveis em relação a uma segmentação automática.

As decisões tomadas relativas a esse processo de construção do PetroTok foram amplamente discutidas e documentadas pela equipe, que buscou embasamento em um referencial teórico linguístico e em outros trabalhos em PLN. Desse modo, a partir dos exemplos práticos apresentados, retirados de um material pertencente ao domínio de petróleo e gás, também foi possível verificar a existência de questões de segmentação específicas de domínio e de gênero textual.

Referências bibliográficas

BIDERMAN, Maria Tereza Camargo. **Teoria lingüística: teoria lexical e lingüística computacional**. Martins Fontes, 2001.

DE SOUZA, Elvis; FREITAS, Cláudia. ET: uma Estação de Trabalho para revisão, edição e avaliação de corpora anotados morfossintaticamente. In: **VI Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana (TILic 2019)**. TILic 2019, Salvador, BA, Brazil, Outubro, 15-18, 2019.

FREITAS, Cláudia; AFONSO, Susana. **Bíblia Florestal: Um manual lingüístico da Floresta Sintá(c)tica**. 2007. Disponível em: <http://www.linguateca.pt/Floresta/BibliaFlorestal/>. Acesso em: 14 jul. 2020.

GREFENSTETTE, Gregory; TAPANAINEN, Pasi. **What is a Word, What is Sentence? Problems of Tokenization**, Grenoble: Rank Xerox Research Centre. 1994.

HEARST, Marti. Untangling text data mining. in: **Proceedings of the 37th Annual meeting of the Association for Computational Linguistics**. 1999. p. 3-10.

KAZAMA, Jun'ichi; MIYAO, Yusuke; TSUJII, Jun'ichi. A maximum entropy tagger with unsupervised hidden markov models. In: **Proc. of the 6th NLPRS**. 2001. p. 333-340.

LOPES, Lucelene; VIEIRA, Renata. Building domain specific parsed corpora in portuguese language. in: **Proceedings of ENIAC 2013**, 2013, Brasil., 2013.

MANNING, Christopher.; SCHÜTZE, Hinrich. **Foundations of statistical natural language processing**. MIT press, 1999.

NIVRE, Joakim et al. Universal dependencies v1: A multilingual treebank collection. In: **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**. 2016. p. 1659-1666.

ROCHA, Luísa; FREITAS, Cláudia; SANTOS, Diana. Preparação para Leitura Distante em português: diálogos entre PLN e Humanidades Digitais. In: **VI Workshop de Iniciação Científica em**

Tecnologia da Informação e da Linguagem Humana (TILic 2019). TILic 2019, Salvador, BA, Brazil, Outubro, 15-18, 2019.

SAMPSON, Geoffrey. **Empirical Linguistics**. London: Continuum, 2001.

SANCHEZ, George. Sentence boundary detection in legal text. In: **Proceedings of the Natural Legal Language Processing Workshop 2019**. 2019. p. 31-38.

SANTOS, Diana; BICK, Eckhard; AFONSO, Susana. **Floresta Sintá(c)tica: apresentação e história do projecto**. 2007. Disponível em https://www.linguateca.pt/Diana/download/SantosBickAfonsoFlorestaSet2007.pdf. Acesso em: 12 ago. 2020

SANTOS, Diana; FREITAS, Cláudia; BICK, Eckhard. OBras: a fully annotated and partially human-revised corpus of Brazilian literary works in public domain. In: **OperCor Forum.** Canela, RS, 24 de setembro de 2018.

SILVEIRA, Aline; DE SOUZA, Elvis; CAVALCANTI, Tatiana; FREITAS, Cláudia. Do PDF ao TXT: Desafios na extração de informação em textos técnico-científicos. In: **VI Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana (TILic 2019)**. TILic 2019, Salvador, BA, Brazil, Outubro, 15-18, 2019.

STRAKA, Milan; HAJIC, Jan; STRAKOVÁ, Jana. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In: **LREC**. 2016.

THOMPSON, Paul; ANANIADOU, Sophia; TSUJII, Jun'ichi. The GENIA Corpus: Annotation Levels and Applications. **In: Handbook of Linguistic Annotation**. Springer, Dordrecht, 2017. p. 1395-1432.