

Documentação relativa à tokenização e à sentencição do corpus Petrolês/PetroTok

Versão 1.1 – 07/08/2020

Aline Silveira

Elvis de Souza

Tatiana Cavalcanti

Wograine Evelyn

Cláudia Freitas

1. Dentro da frase:

- **hífen**: palavras separadas por hífen sempre contam como um único token, ou seja, ele não é critério de separação de palavras (Ver Critérios de separação de frases), e o espaço que houver entre o hífen e um dos termos da palavra composta deverá ser eliminado
Exemplo: “**auto- sustentável**” é um token só, e vira “**auto-sustentável**”
- **travessão**: se houver espaço entre as palavras e o travessão, separá-los como tokens distintos. Se não houver espaço, tudo forma um único token.
Exemplo:
Faixa de temperatura: 160oC-220oC → tokens: 160oC; –; 220oC
Faixa de temperatura: 80oC-120oC → token: 80oC-120oC
- **barra**: palavras separadas por barra são um token só
Exemplo: “propriedades *físico/químicas*”; “fator *homem/hora*”; “*km/h*”; “*S/cm*”
→ **OBS.: UNIDADES e SÍMBOLOS** (60 km/h; 10-5 S/cm; 5V; 200°C; 10ml; 50%)
 - a) Se *houver* espaço entre o número e a unidade, temos *dois* tokens distintos.
Exemplo: “condutividade de 10-5 S/cm e janela de estabilidade eletroquímica maior que 5 V.” → tokens: **10-5** e **S/cm**
 - b) Se *não houver* espaço (100oC), tudo se configura como *um* token.
Exemplo: perdem água a temperaturas próximas a 100oC” → token: **100oC**

- **parênteses:** o parênteses é um token por si só, como outros sinais de pontuação (ponto final, dois pontos, vírgula etc.), e portanto não deve estar atrelado a nenhuma palavra. Uma exceção a essa regra se encontra em compostos químicos como “Indeno(1,2,3-cd)pireno”, “Benzo(g,h,i)perileno” ou “Dibenzo(a,h)antraceno”, cada um deles contando como um token. Não faria sentido segmentá-los ao meio, pois constituem unidades de sentido.

Exemplo: O **poli(bisfenol A-co-epicloridrina)** (PBE) é uma resina epóxi contendo grupos éter que podem coordenar cátions (Figura 4).

poli	nsubj(resina)
(flat:name(poli)
bisfenol	flat:name(poli)
A-co-epicloridrina	flat:name(poli)
)	flat:name(poli)

- **et al.:** separado em dois tokens, “et” e “al.”, que terão como pos NOUN

→ **OBS.2:** Números referentes às notas de rodapé, assim como equações, não devem estar no txt.

2. Delimitação de frases; sentencição

- **Títulos e subtítulos:** se for possível, tirar a numeração. Colocar ponto final para separá-los como uma sentença (Ver SEPARADORES DE SENTENÇA – “.” marca fim de frase)

PDF x TXT:

3-Introdução:

3.1-Caracterização Geral:

Desde 1887, quando se teve o início da “era da propulsão mecânica” e posteriormente com o surgimento da indústria petroquímica em 1930, o petróleo tem tido importante função na sociedade, como fonte combustível e fornecendo matéria sintética para diversos produtos (CETESB, 2002).

```
# sent_id = 6-20140908-MONOGRAFIA_0-1
# text = Introdução.
Introdução
.
-
-
-
-

# sent_id = 6-20140908-MONOGRAFIA_0-2
# text = Caracterização Geral.
Caracterização
Geral
.
-
-
-
-
-
```

– **Listas itemizadas:**

Número + ponto final OU qualquer outro marcador (bolinha, tracinho) são eliminados.

- Quando separados por **ponto e vírgula (ou vírgula)**: os itens formam uma única sentença (exemplos 1, 2 e 3).
- Quando **não há pontuação** sucedendo o item: adição de ponto e vírgula, colocando ponto final apenas no último item da lista (exemplo 3).
- Quando separados por **ponto final**: cada item forma uma sentença própria (ponto final é delimitador de sentença SEMPRE).

Exemplo 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 á 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;

Como proceder:
vem após

O que número

+ ponto final é deslocado para logo após o dois pontos ou ponto e vírgula, seguido de espaço:

De acordo com CETESB (2002) as características mais relevantes em um derrame são: XXXX

(ponto final é separador de sentença, mas ; e : **não** são – para mais informações, ver “Critérios de separação de frases”)

Resultado:

De acordo com CETESB (2002) as características mais relevantes em um derrame são: Tipo e quantidade de petróleo, sendo os mais tóxicos os óleos leves devido á presença de uma quantidade maior de compostos aromáticos; Amplitude de maré, podendo esta agravar o efeito do derrame ou mesmo contribuir para processo de limpeza; Época do ano, por causar consideráveis variações na estrutura e composição das comunidades biológicas costeiras;

Exemplo 2:

PEMFC (proton exchange membran fuel cell – célula a combustível de membrana de troca de próton):

Eletrólito: membrana polimérica de condução protônica;

Faixa de temperatura: 80°C-120°C;

Vantagens: alta densidade de potência, operação flexível, mobilidade;

Desvantagens: custo da membrana e catalisador, contaminação do catalisador com monóxido de carbono;

Aplicações: veículos automotores, espaçonaves, unidades estacionárias.

Como proceder: PEMFC é subtítulo, portanto colocamos ponto final.

#text = PEMFC (proton exchange membran fuel cell – célula a combustível de membrana de troca de próton).

#text = Eletrólito: membrana polimérica de condução protônica; Faixa de temperatura: 80oC-120oC; Vantagens: alta densidade de potência, operação flexível, mobilidade; Desvantagens: custo da membrana e catalisador, contaminação do catalisador com monóxido de carbono; Aplicações: veículos automotores, espaçonaves, unidades estacionárias.

Exemplo 3:

Os dados relativos aos projetos foram agrupados em tabelas contendo:

- a localização geográfica, obtidas através da padronização das latitudes e longitudes em graus e décimos de graus,
- salinidades na superfície e no fundo das estações;
- temperaturas na superfície e no fundo das estações
- estação do ano;
- abundância de ovos e larvas, identificadas ao menor nível taxonômico possível e padronizados em número de indivíduos/100 m³ (N/100 m³).

text = Os dados relativos aos projetos foram agrupados em tabelas contendo: a localização geográfica, obtidas através da padronização das latitudes e longitudes em graus e décimos de graus, salinidades na superfície e no fundo das estações; temperaturas na superfície e no fundo das estações; estação do ano; abundância de ovos e larvas, identificadas ao menor nível taxonômico possível e padronizados em número de indivíduos/100 m3 (N/100 m3).

Critérios de separação de frases

1. SEPARADORES DE SENTENÇA

a. ponto final

text = Nesta reação de adição, um mol de ligações duplas conjugadas sempre consumirá dois mols de iodo.
sent_id = 0-20150121-TEEMSC_O-11

OBS.: O ponto final se diferencia do ponto marcador de abreviações, como aquele encontrado na expressão “et al.” – tokenizada como “et” e “al.”

2. NÃO SEPARADORES DE SENTENÇA

a. vírgula (,)

text = Estes compostos diminuem a qualidade dos produtos petrolíferos devido à sua fácil polimerização, já que as suas ligações duplas conjugadas apresentam alta reatividade.

sent_id = 0-20150121-TEEMSC_O-3

b. ponto e vírgula (;)

text = Adicionalmente, tem limitações relacionadas com a concentração e a natureza do fenol (Smith, 1987; Spiker, 1992; Aitken, 1993 e Wada, 1994).

sent_id = 2-20150126-TEEDSC_O-9

OBS.: Os itens de uma lista itemizada, quando terminados em ponto e vírgula, formam uma única sentença (ver tópico “Listas itemizadas” mais acima).

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 á 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;

No TXT:

text = De acordo com CETESB (2002) as características mais relevantes em um derrame são: Tipo e quantidade de petróleo, sendo os mais tóxicos os óleos leves devido à presença de uma quantidade maior de compostos aromáticos; Amplitude de maré, podendo esta agravar o efeito do derrame ou mesmo contribuir para processo de limpeza; Época do ano, por causar consideráveis variações na estrutura e composição das comunidades biológicas costeiras; Grau de hidrodinamismo, determinado pela quantidade, intensidade e força das ondas e correntes locais; Ciclo construtivo/destrutivo do ambiente: determinado pelo grau de erosão e deposição das praias; Tipo de substrato; Tipo de comunidade; Exposição prévia a outros impactos; Formas de limpeza aplicadas ao derrame.

sent_id = 6-20140908-MONOGRAFIA_0-45

c. dois pontos (:)

text = Especificamente em organismos planctônicos, a contaminação pode vir de diferentes formas: através da fração solúvel, do contato direto com a mancha ou mesmo pela ingestão de alimentos contaminados por petróleo.

sent_id = 6-20140908-MONOGRAFIA_0-50

OBS.: Mesmo quando precede uma lista itemizada separada do texto, dois pontos não separam sentenças (ver tópico de “Listas itemizadas” mais acima).

No PDF:

Os dados relativos aos projetos foram agrupados em tabelas contendo:

- a localização geográfica, obtidas através da padronização das latitudes e longitudes em graus e décimos de graus,
- salinidades na superfície e no fundo das estações;
- temperaturas na superfície e no fundo das estações
- estação do ano;
- abundância de ovos e larvas, identificadas ao menor nível taxonômico possível e padronizados em número de indivíduos/100 m³ (N/100 m³).

No TXT:

text = Os dados relativos aos projetos foram agrupados em tabelas contendo: a localização geográfica, obtidas através da padronização das latitudes e longitudes em graus e décimos de graus, salinidades na superfície e no fundo das estações; temperaturas na superfície e no fundo das estações; estação do ano; abundância de ovos e larvas, identificadas ao menor nível taxonômico possível e padronizados em número de indivíduos/100 m³ (N/100 m³).

d. travessão (–)

#text = Neste trabalho estudou-se a degradação oxidativa usando enzima cloroperoxidase (CPO) de *Caldariomyces fumago* dos compostos fenólicos [...] e de

compostos fenólicos presentes nas águas residuais de refinaria em efluente bruto - água ácida (EB) e água de fundos de tanque de armazenamento de óleo cru (TA).

#sent_id: 2-20150126-TESEDSC_O_resumo-1

e. parênteses ()

text = Além dos dienos conjugados, o estireno e seus derivados (devido à conjugação da ligação dupla com o sistema aromático) também apresentam forte tendência à polimerização (POLÁK et al., 1986).

sent_id = 0-20150121-TESEMSC_O-4

PROBLEMAS RELACIONADOS ÀS QUESTÕES DE TOKENIZAÇÃO/SENTENCIAÇÃO

1) Corte de frases

- Ponto final *não separou* a sentença
- Houve corte de frases entre “padrões” e “encontra-se” (a parte destacada em amarelo foi eliminada)

NO PDF:

Todas as amostras foram analisadas por monitoramento seletivo de íons (MSI), para identificação e quantificação dos compostos em estudo. O modo de análise por varredura linear (SCAN) de 35–450 daltons foi utilizado para a obtenção dos espectros de massas e confirmação do tempo de retenção dos compostos em questão. A identificação dos compostos foi realizada utilizando a biblioteca eletrônica de espectros de massas Willey275 e espectros de massas de compostos padrões.

II.4.3 – Avaliação da Recuperação do Método

Foi realizado um estudo de recuperação empregando-se o método CG/EM, em seis amostras de nafta dopadas com diferentes padrões (amostras de nafta com valor de dieno conjugado conhecido). Estas amostras foram preparadas mediante a adição de padrões de dienos conjugados e não conjugados, em amostras de nafta. O preparo das amostras de nafta dopadas encontra-se no item II.4.1 (páginas 46 e 47).

2)

```
# sent_id = 10-20150122-MONOGRAFIA_0-108
# text = 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, -, isto é, sítios ácidos de aparecem grupos hidroxilas ponte em cada sítio AlO4 Bronsted.
```


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.

```
# sent_id = 16-20150122-MONOGRAFIA_0-46
# text = Para a produção de gás de síntese a partir do gás natural são utilizados os seguintes processos: i.
1      Para      para      ADP      _      _      3      case      _      start_char=0|end_char=4
2      a          o          DET      _      _      _      _      _      3      det      _
3      produção  produção  NOUN      _      _      Gender=Fem|Number=Sing 15      obl      _      start_ch
4      de          de          ADP      _      _      5      case      _      start_char=16|end_char=18
5      gás        gás        NOUN      _      _      Gender=Masc|Number=Sing 3      nmod      _      start_char=19|end_char=2
6      de          de          ADP      _      _      7      case      _      start_char=23|end_char=25
7      síntese    síntese    NOUN      _      _      Gender=Fem|Number=Sing 5      nmod      _      start_char=26|end_char=3
8      a          a          ADP      _      _      12     case      _      start_char=34|end_char=35
9      partir     partir     NOUN      _      _      8      fixed     _      start_char=36|end_char=42
10-11 do          _          _          _      _      _      _      _      _      start_char=43|end_char=45
10     de          de          ADP      _      _      8      fixed     _      _
11     o          o          DET      _      _      Definite=Def|Gender=Masc|Number=Sing|PronType=Art 12     det      _
12     gás        gás        NOUN      _      _      Gender=Masc|Number=Sing 3      nmod      _      start_char=46|end_char=4
13     natural    natural    ADJ      _      _      Gender=Masc|Number=Sing 12     amod      _      start_char=50|end_char=5
14     são        ser        AUX      _      _      Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin 15     aux:pass
15     utilizados utilizar    VERB      _      _      Gender=Masc|Number=Plur|VerbForm=Part|Voice=Pass
16     os          o          DET      _      _      Definite=Def|Gender=Masc|Number=Plur|PronType=Art 18     det      _
17     seguintes  seguinte  ADJ      _      _      Gender=Masc|Number=Plur 18     amod      _      start_ch
18     processos  processo  NOUN      _      _      Gender=Masc|Number=Plur 15     nsubj:pass
19     ;          ;          PUNCT      _      _      15     punct      _      start_char=95|end_char=96
20     i          i          PROPON      _      _      Gender=Masc|Number=Sing 18     appos      _      start_char=97|end_char=9
21     .          .          PUNCT      _      _      15     punct      _      start_char=98|end_char=99
```

```
# sent_id = 16-20150122-MONOGRAFIA_0-47
# text = Reforma a vapor; ii.
1      Reforma  reforma  NOUN      _      _      Gender=Fem|Number=Sing 0      root      _      start_char=0|end_char=7
2      a          a          ADP      _      _      3      case      _      start_char=8|end_char=9
3      vapor     vapor     NOUN      _      _      Gender=Masc|Number=Sing 1      nmod      _      start_char=10|end_char=15
4      ;          ;          PUNCT      _      _      1      punct      _      start_char=15|end_char=16
5      ii        ii        ADJ      _      _      Gender=Masc|NumType=Ord|Number=Sing 1      amod      _      start_char=17|end_char=19
6      .          .          PUNCT      _      _      1      punct      _      start_char=19|end_char=20
```

```
# sent_id = 16-20150122-MONOGRAFIA_0-48
# text = Oxidação parcial; iii.
1      Oxidação  oxidação  NOUN      _      _      Gender=Fem|Number=Sing 0      root      _      start_char=0|end_char=8
2      parcial   parcial   ADJ      _      _      Gender=Fem|Number=Sing 1      amod      _      start_char=9|end_char=16
3      ;          ;          PUNCT      _      _      1      punct      _      start_char=16|end_char=17
4      iii        iii        ADJ      _      _      Gender=Masc|NumType=Ord|Number=Sing 1      amod      _      start_char=18|end_char=21
5      .          .          PUNCT      _      _      1      punct      _      start_char=21|end_char=22
```

```
# sent_id = 16-20150122-MONOGRAFIA_0-49
# text = Reforma a seco.
1      Reforma  reforma  NOUN      _      _      Gender=Fem|Number=Sing 0      root      _      start_char=0|end_char=7
2      a          a          ADP      _      _      3      case      _      start_char=8|end_char=9
3      seco      seco      NOUN      _      _      Gender=Masc|Number=Sing 1      nmod      _      start_char=10|end_char=14
4      .          .          PUNCT      _      _      1      punct      _      start_char=14|end_char=15
```

☐ Abreviação de nomes: o ponto foi visto como ponto final

```
# sent_id = 90-20141119-TESEDSC_1-61
# text = Seus principais autores são, entre outros, Douglass North, James G.
1      Seus      seu      DET      _      _      Gender=Masc|Number=Plur|PronType=Prs 3      det      _      start_char=0|end_char=4
2      principais principal ADJ      _      _      Gender=Masc|Number=Plur 3      amod      _      start_char=5|end_char=15
3      autores    autor  NOUN      _      _      Gender=Masc|Number=Plur 9      nsubj      _      start_char=16|end_char=23
4      são        ser      AUX      _      _      Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin 9      cop      _      start_char=24|end_char=27
5      ,          ,          PUNCT      _      _      7      punct      _      start_char=27|end_char=28
6      entre      entre  ADF      _      _      7      case      _      start_char=29|end_char=34
7      outros     outro  PRON      _      _      Gender=Masc|Number=Plur|PronType=Ind 9      nmod      _      start_char=35|end_char=41
8      ,          ,          PUNCT      _      _      7      punct      _      start_char=41|end_char=42
9      Douglass   Douglass PROPON      _      _      Gender=Masc|Number=Sing 0      root      _      start_char=43|end_char=51
10     North      North  PROPON      _      _      Gender=Masc|Number=Sing 0      root      _      start_char=52|end_char=62
```

3) Troca da ordem de palavras

- Troca de pedaços do texto: “inserção de polímeros nos mesmos, possibilitando a formação de nanocompósitos.” se tornou “nanocompósitos. inserção de polímeros nos mesmos, possibilitando a formação de.”
- Zeólita (subtítulo) não recebeu ponto final
- A sentença logo após o subtítulo foi puxada para sentença anterior (por causa da falta do ponto final depois do subtítulo)

NO PDF:

para a H-ZSM-5, H-mordenita e HY, respectivamente. Neste trabalho, optou-se pela Zeólita Y, que apesar de apresentar maior E_a , possui canais mais largos para a inserção de polímeros nos mesmos, possibilitando a formação de nanocompósitos¹⁷.

Zeólita

A zeólita Y, cuja estrutura encontra-se representada na Figura 5, foi doada pela PETROBRAS e possui área de $728 \text{ m}^2/\text{g}$, com a seguinte composição: $\text{SiO}_2 = 65,9\%$; $\text{Al}_2\text{O}_3: 20\%$, $\text{Na}_2\text{O}: 13,1\%$ (razão Si/Al molar: 2,8).

NO TXT:

```
# sent_id = 10-20150122-MONOGRAFIA_0-110
# text = Neste trabalho, optou-se pela Zeólita Y, que apesar de apresentar maior Ea, possui canais mais largos para a nanocompósitos. inserção de polímeros nos mesmos, possibilitando a formação de Zeólita A zeólita Y, cuja estrutura encontra-se representada na Figura 5, foi doada pela PETROBRAS e possui área de 728 m2/g, com a seguinte composição: SiO2 =65,9%; Al2O3:20%, Na2O: 13,1% (razão Si/Al molar: 2,8).
```

4) Bibliografia: A bibliografia não foi retirada na tese 0-20150121-TESEMSC_0. A imagem abaixo mostra a última sentença da tese antes das referências. Todas as sentenças que se seguem no txt fazem parte da bibliografia.

```
# sent_id = 0-20150121-TESEMSC_0-414
# text = Recomenda-se então a aplicação do método voltamétrico na determinação do valor de dienos conjugados pela indústria petroquímica em substituição ao método atual, o UOP-326, visto que o método proposto necessita de um pequeno volume de amostra (100 µL), apresenta uma boa repetibilidade (desvio padrão relativo inferior a 5%), além de apresentar um tempo de análise relativamente curto (1 hora).3. BARMAN, B.
```

5) Sem espaço entre os itens da lista (tese 90)

text = Para atender a este objetivo, foram definidos os seguintes objetivos específicos:1.2.2 Específicos1) Analisar o papel das principais instituições executoras do PNPB.2) Caracterizar o perfil da agricultura familiar brasileira.3) Analisar as características produtivas de cada uma das principais culturas oleaginosas de biodiesel, produzidas pela agricultura familiar nas diferentes regiões brasileiras, e sua relação com as medidas de desigualdades regionais.4) Criar índices estaduais referentes à produção das oleaginosas, a partir das possíveis concentrações dos fatores (econômico, tecnológico, sociopolítico e socioambiental), a partir da análise dos resultados da agricultura familiar.1.3 Estrutura do trabalho O trabalho está dividido em cinco capítulos, sendo o primeiro esta introdução, em que se procura expor a ideia central e a importância da pesquisa.

6) n° descontraído: n° virou “no” □ em + o

48	embora	embora	SCONJ	-	-	59	mark	-	start_char=238 end_char=244	
49-50	na	-	-	-	-	51	case	-	start_char=245 end_char=247	
49	em	em	ADP	-	-	51	case	-		
50	a	o	DET	-	-		Definite=Def Gender=Fem Number=Sing PronType=Art	51	det	-
51	Lei	Lei	PROPN	-	-		Gender=Fem Number=Sing	59	nsubj	start_char=248 end_char=251
52-53	n°	-	-	-	-				start_char=252 end_char=254	
52	em	em	ADP	-	-	54	case	-		
53	o	o	DET	-	-		Definite=Def Gender=Masc Number=Sing PronType=Art	54	det	-
54	4.504	4.504	PROPN	-	-		Number=Sing	51	flat:name	start_char=255 end_char=260

Importante trocar as **aspas duplas** pelos símbolos do Bosque (<< >>), caso contrário o parser não identifica que são **PUNCT** e faz uma sintaxe estranhíssima
<https://meet.google.com/linkredirect?authuser=0&dest=http%3A%2F%2Finterrogatorio.ica.ele.puc-rio.br%2Fcgi-bin%2Finterrogar.cgi%3Fcorpus%3Dbosque2.5-workbench.conllu%26params%3D1%2520ccomp%253Aparataxis>

47	"	"	PROPN	Gender=Masc Number=Sing	45	appos	start_char=227 end_char=228
48	International	International	PROPN	Number=Sing	45	flat:name	start_char=228 end_char=241
49	Organization	Organization	PROPN	Number=Sing	47	flat:name	start_char=242 end_char=254
50	for	for	PROPN	Number=Sing	47	flat:name	start_char=255 end_char=258
51	Standardization	Standardization	PROPN	Number=Sing	47	flat:name	start_char=259 end_char=274
52	"	"	PROPN	Number=Sing	47	flat:name	start_char=274 end_char=275

8) R\$ - foi tokenizado separado (deve ser junto com pos SYM)

<https://meet.google.com/linkredirect?authuser=0&dest=http%3A%2F%2Finterrogatorio.ica.ele.puc-rio.br%2Fcgi-bin%2Finterrogar.cgi%3Fcorpus%3Dbosque2.5-workbench.conllu%26params%3D1%2520R%255C%2524>

77	ficou	ficar	VERB	Mood=Ind Number=Sing Person=3 Tense=Past VerbForm=Fin	47	advcl	start_char=380 end_char=385
78	em	em	ADP	79	case	start_char=386 end_char=388	
79	R	R	PROPN	Gender=Masc Number=Sing	77	obl	start_char=389 end_char=390
80	\$	\$	PROPN	Number=Sing	79	flat:name	start_char=390 end_char=391
81	431,82	431,82	NUM	NumType=Card	82	nummod	start_char=392 end_char=398
82	real	real	NOUN	Gender=Masc Number=Plur	77	obj	start_char=399 end_char=404
83	por	por	ADP	84	case	start_char=405 end_char=408	
84	hectares	hectare	NOUN	Gender=Masc Number=Plur	82	nmod	start_char=409 end_char=417

9) Falta de espaço no # text depois de “segunda,” □ fez a tokenização ficar estranha:

# sent_id = 90-20141119-TESEDSC_1-664									
#	text	# Com relação ao percentual destinado às lavouras permanentes, a primeira utilizou 25,27%, e a segunda,23,54%%.							
1	Com	com	ADP	2	case	start_char=0 end_char=3			
2	relação	relação	NOUN	Gender=Fem Number=Sing	14	obl	start_char=4 end_char=11		
3-4	ao	a	ADP	5	case	start_char=12 end_char=14			
4	o	o	DET	Definite=Def Gender=Masc Number=Sing PronType=Art	5	det	start_char=15 end_char=25		
5	percentual	percentual	NOUN	Gender=Masc Number=Sing	2	nmod	start_char=26 end_char=35		
6	destinado	destinar	VERB	Gender=Masc Number=Sing VerbForm=Part	5	acl	start_char=36 end_char=38		
7-8	às	a	ADP	9	case	start_char=39 end_char=47			
8	as	o	DET	Definite=Def Gender=Fem Number=Plur PronType=Art	9	det	start_char=48 end_char=59		
9	lavouras	lavouras	NOUN	Gender=Fem Number=Plur	6	obj	start_char=60 end_char=62		
10	permanentes	permanente	ADJ	Gender=Fem Number=Sing PronType=Art	13	det	start_char=63 end_char=71		
11	,	,	PUNCT	6	punct	start_char=72 end_char=80			
12	a	o	DET	Definite=Def Gender=Fem Number=Sing PronType=Art	13	det	start_char=81 end_char=87		
13	primeira	primeiro	ADJ	Gender=Fem NumType=Ord Number=Sing	14	nsubj	start_char=88 end_char=90		
14	utilizou	utilizar	VERB	Mood=Ind Number=Sing Person=3 Tense=Past VerbForm=Fin	0	root	start_char=91 end_char=92		
15	25,27%	25,27%	NUM	NumType=Card	14	obj	start_char=93 end_char=97		
16	,	,	PUNCT	20	punct	start_char=97 end_char=98			
17	e	e	CCONJ	20	cc	start_char=99 end_char=100			
18	a	a	ADP	20	case	start_char=101 end_char=102			
19	segunda,23,54%	segunda,23,54%	NUM	NumType=Card	20	nummod	start_char=103 end_char=107		
20	%	%	SYM	15	conj	start_char=107 end_char=108			
21	.	.	PUNCT	14	punct	start_char=108 end_char=109			

10) Equação no corpo do texto

# sent_id = 16-20150122-MONOGRAFIA_0-15									
# text = CH3X CH3OCH3 + HX X = Cl, Br Reação 1.3 - Formação de DME a partir de halometano Este trabalho apresenta resultados da conversão do clorometano									
1	CH3X	CH3X	PROPN	-	Gender=Masc Number=Sing	22	nsubj	-	start_char=0 end_char=4
2	CH3OCH3	CH3OCH3	PROPN	-	Number=Sing	1	flat:name	-	start_char=5 end_char=12
3	+	+	PUNCT	-	4	punct	-	-	start_char=13 end_char=14
4	HX	HX	PROPN	-	Gender=Fem Number=Sing	1	conj	-	start_char=15 end_char=17
5	X	X	PROPN	-	Number=Sing	4	flat:name	-	start_char=18 end_char=19
6	=	=	PUNCT	-	7	punct	-	-	start_char=20 end_char=21
7	Cl	Cl	PROPN	-	Gender=Fem Number=Sing	1	conj	-	start_char=22 end_char=24
8	,	,	PUNCT	-	9	punct	-	-	start_char=24 end_char=25
9	Br	Br	PROPN	-	Gender=Fem Number=Sing	1	conj	-	start_char=26 end_char=28
10	Reação	Reação	PROPN	-	Number=Sing	9	flat:name	-	start_char=29 end_char=35
11	1.3	1.3	PROPN	-	Number=Sing	9	flat:name	-	start_char=36 end_char=39
12	-	-	PUNCT	-	13	punct	-	-	start_char=40 end_char=41
13	Formação	Formação	PROPN	-	Gender=Fem Number=Sing	1	conj	-	start_char=42 end_char=50
14	de	de	ADP	-	15	case	-	-	start_char=51 end_char=53
15	DME	DME	PROPN	-	Number=Sing	13	nmod	-	start_char=54 end_char=57
16	a	a	ADP	-	19	case	-	-	start_char=58 end_char=59
17	partir	partir	NOUN	-	16	fixed	-	-	start_char=60 end_char=66
18	de	de	ADP	-	16	fixed	-	-	start_char=67 end_char=69



Este trabalho apresenta resultados da conversão do clorometano e água a DME sobre a zeólita ZSM-5 trocada com alguns cátions.

12) Números entre colchetes (referência à bibliografia) tokenizados como "[10" e "]"

#text = [10]

[10

]