

# **ETIQUETAS DE SEGMENTAÇÃO:** UMA PROPOSTA PARA ANÁLISE DA SEGMENTAÇÃO EM LEGENDAS INTRALINGUÍSTICAS DE FILMES BRASILEIROS

Élida Gama Chaves  
Vera Lúcia Santiago Araújo



VI Escola Brasileira de Linguística Computacional (EBRALC 2012)  
XI Encontro de Linguística de Corpus (ELC 2012)



# Introduction

---

- ▶ This work is in the field of Corpus Based Translation Studies (Baker: 1996 and Audiovisual Translation (AVT), more specifically in the studies of subtitling.
- ▶ It aims at presenting a proposal of personalized tags to be used in the analysis of segmentation in intralingual subtitles.



# Segmentation in subtitling

---

- ▶ Segmentation is a subtitling parameter related to the subdivision of the subtitles and the distribution of the text within subtitles (**line break**) and across subtitles. It can occur
  - ▶ visually (on the basis of shot cuts)
  - ▶ rhetorically (on the basis of speech rhythms)
  - ▶ linguistically (on the basis of semantic units)



- 
- ▶ Previous studies suggest that any segmentation problem forces the viewers to decode the subtitle text, and thus they may get tired more quickly and lose the pleasure afforded by audiovisual product.



# Rhetorical segmentation problems

---

- ▶ The information is anticipated or delayed in the subtitles, or do not follow the speech, including hesitations, pauses and features of oral speech.
- ▶ “The way subtitles are segmented and distributed must reflect some of the dialogues dynamics. Good rhetorical segmentation helps convey surprise, suspense, irony, hesitation, etc.” (Cintas and Remael (2007, p. 179))



# Linguistic segmentation problems

---

The constituents (phrases and clauses) are broken inside.

Karamitroglou (1998) states that subtitled text should appear segmented at the **highest syntactic nodes** possible. This means that each subtitle flash should ideally contain one complete sentence.



- ▶ From the Karamitroglou proposal (1998), Perego (2008) analyses cases of poor segmentation in a corpus of varied subtitles, then defines the following categories according to each kind of problem.
- 

- ▶ Breaking of Noun Phrase
- ▶ Breaking of Prepositional Phrase
- ▶ Breaking of Verb Phrase
- ▶ Breaking of Coordinated and Subordinated clauses.



▶ Based on the categories proposed by Perego (2008), it was possible to define personalized tags to the investigation of segmentation for Brazilian Portuguese subtitles.

- ▶ 4 tags indicating segmentation problems.
- ▶ 19 tags to analyse segmentation problems



# Defining Tags

---

- ▶ In order to find out segmentation problems in the subtitling an investigation was carried out to determine the nature of these problems.
- ▶ Each segmentation problem was annotated manually, then defined as a segmentation tag.



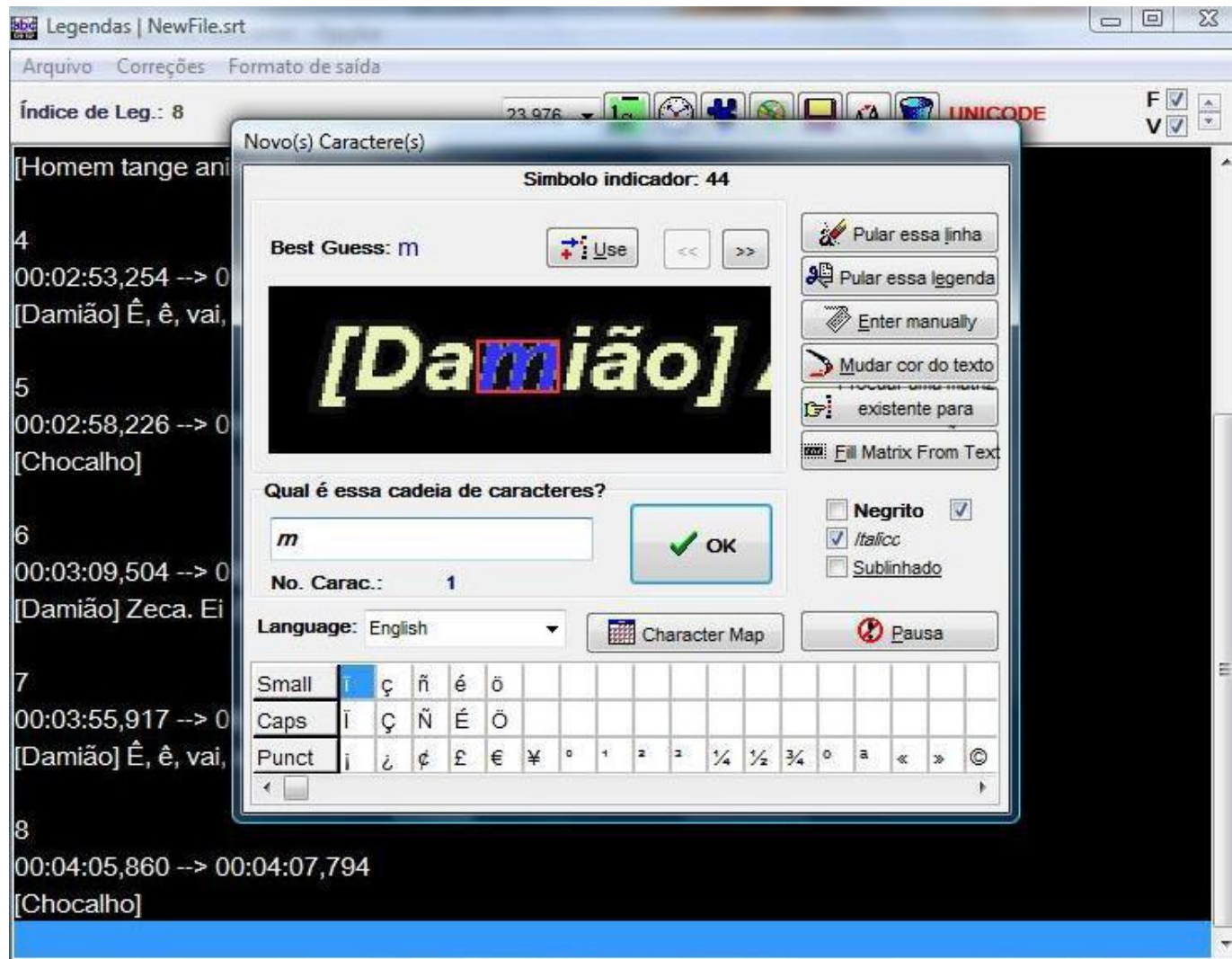
# Materials and Methods

---

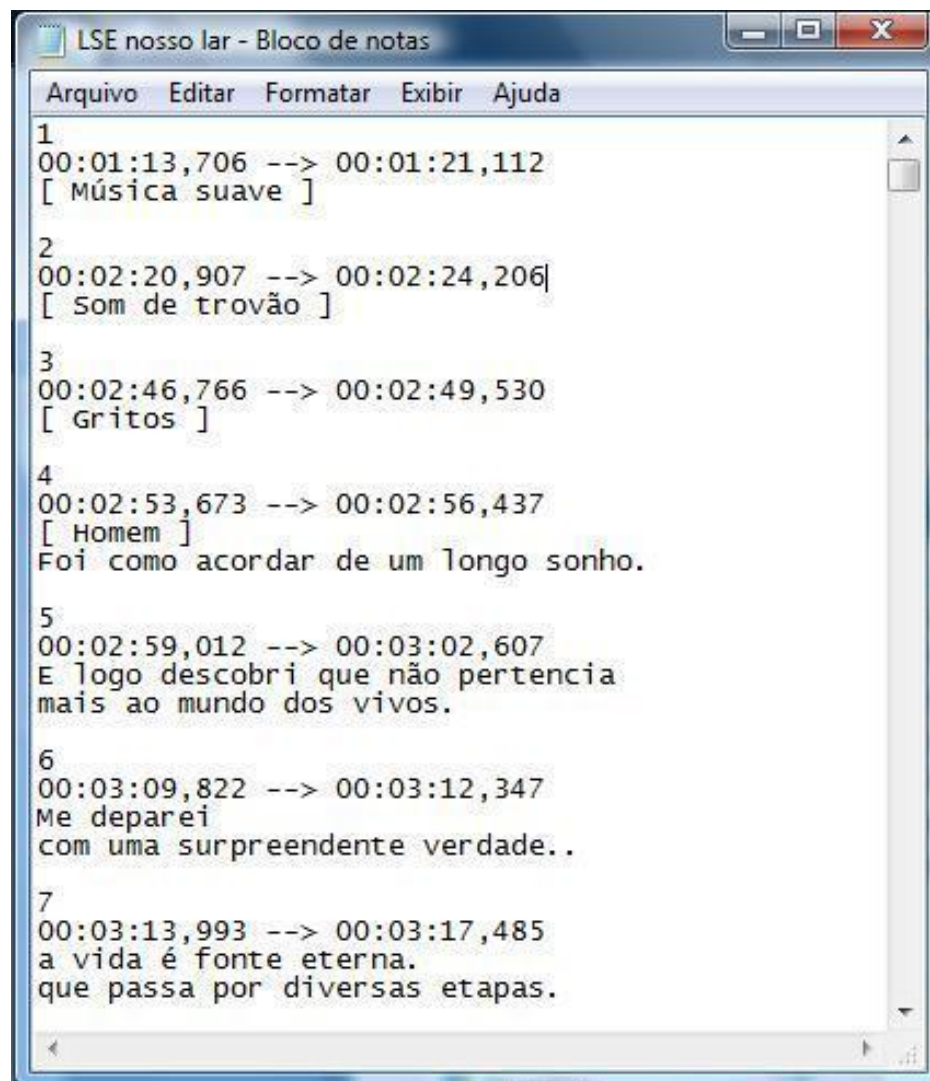
- ▶ Corpus
  - ▶ 1132 subtitles of the Brazilian film *Nosso Lar* (2010)
- ▶ Subtitle extractor
  - ▶ SubRip (.srt)
- ▶ Manual Tagger
  - ▶ *Windows Notepad* (.txt)
- ▶ Analysis
  - ▶ WordSmith Tools 5.0
    - ▶ Concord



# SubRip Interface



# Subtitle file (.srt)



```
LSE nosso lar - Bloco de notas
Arquivo  Editar  Formatar  Exibir  Ajuda

1
00:01:13,706 --> 00:01:21,112
[ Música suave ]

2
00:02:20,907 --> 00:02:24,206
[ Som de trovão ]

3
00:02:46,766 --> 00:02:49,530
[ Gritos ]

4
00:02:53,673 --> 00:02:56,437
[ Homem ]
Foi como acordar de um longo sonho.

5
00:02:59,012 --> 00:03:02,607
E logo descobri que não pertencia
mais ao mundo dos vivos.

6
00:03:09,822 --> 00:03:12,347
Me deparei
com uma surpreendente verdade..

7
00:03:13,993 --> 00:03:17,485
a vida é fonte eterna.
que passa por diversas etapas.
```

# Unannotated X Annotated subtitles

UNANNOTATED SUBTITLES	ANNOTATED SUBTITLES
<p>179 00:26:01,560 --&gt; 00:26:05,052 Logo eu! Logo eu que nem acreditava em vida após a morte!</p>	<p>&lt;sub179&gt;&lt;L&gt;179 &lt;t&gt;00:26:01,560 --&gt; 00:26:05,052&lt;/t&gt; &lt;cpl24&gt;Logo eu! Logo eu que <u>nem</u>&lt;PROSEGG&gt;&lt;SV_negação+verbo&gt; &lt;cpl32&gt;<u>acreditava</u> em vida após a morte! &lt;velocidade da legenda_alta 56c/3,5&gt;&lt;/sub179&gt;</p>
<p>196 00:26:49,441 --&gt; 00:26:53,309 Todo o ceticismo termina quando se acorda no mundo espiritual.</p>	<p>&lt;sub196&gt;&lt;L&gt;196 &lt;t&gt;00:26:49,441 --&gt; 00:26:53,309&lt;/t&gt; &lt;cpl31&gt;Todo o ceticismo termina <u>quando</u>&lt;PROSEGG&gt;&lt;SUBORD_conj+oração&gt; &lt;cpl30&gt;<u>se acorda no mundo espiritual</u>. &lt;velocidade da legenda_alta 61c/3,8&gt;&lt;/sub196&gt;</p>
<p>197 00:26:54,646 --&gt; 00:26:57,740 O amigo parece ter compreendido o sentido da água,</p>	<p>&lt;sub197&gt;&lt;L&gt;197 &lt;t&gt;00:26:54,646 --&gt; 00:26:57,740&lt;/t&gt; &lt;cpl18&gt;O amigo <u>parece ter</u>&lt;PROSEGG&gt;&lt;SV_verbo+verbo&gt; &lt;cpl31&gt;<u>compreendido</u> o sentido da água, &lt;velocidade da legenda_alta 49c/3,1s&gt;&lt;/sub197&gt;</p>
<p>201 00:27:05,190 --&gt; 00:27:09,126 <u>Lísias</u>, eu não vou parar enquanto não souber o que está acontecendo.</p>	<p>&lt;sub201&gt;&lt;L&gt;201 &lt;t&gt;00:27:05,190 --&gt; 00:27:09,126&lt;/t&gt; &lt;cpl33&gt;<u>Lísias</u>, eu não vou parar <u>enquanto</u>&lt;PROSEGG&gt;&lt;SUBORD_conj+oração&gt; &lt;cpl34&gt;<u>não souber o que está acontecendo</u>. &lt;velocidade da legenda_alta 67c/4s&gt;&lt;/sub201&gt;</p>
<p>523 00:47:19,870 --&gt; 00:47:22,065 O que sabe sobre a medicina espiritual?</p>	<p>&lt;sub523&gt;&lt;L&gt;523 &lt;t&gt;00:47:19,870 --&gt; 00:47:22,065&lt;/t&gt; &lt;cpl16&gt;O que sabe <u>sobre</u>&lt;PROSEGG&gt;&lt;SP_prep+subst&gt; &lt;cpl22&gt;<u>a medicina espiritual?</u> &lt;velocidade da legenda_alta 38c/2,2s&gt;&lt;/sub523&gt;</p>
<p>805 01:09:49,285 --&gt; 01:09:51,845 Não vou esperar <u>anos</u> e <u>anos</u> para retornar.</p>	<p>&lt;sub805&gt;&lt;L&gt;805 &lt;t&gt;01:09:49,285 --&gt; 01:09:51,845&lt;/t&gt; &lt;cpl20&gt;Não vou esperar <u>anos</u>&lt;PROSEGG&gt;&lt;SAdv&gt; &lt;cpl21&gt;<u>e anos</u> para retornar. &lt;velocidade da legenda_alta 41c/2,5s&gt;&lt;/sub805&gt;</p>

# Segmentation Tags

ETIQUETA INDICATIVA DE PROBLEMA DE SEGMENTAÇÃO LINGÜÍSTICA (Gramatical)
<PROSEGG>
ETIQUETAS INDICATIVA DE PROBLEMA DE SEGMENTAÇÃO RETÓRICA
<PROSEGR_antecipouinformação>
<PROSEGR_atrasouinformação>
ETIQUETA INDICATIVA DE PROBLEMA DE SEGMENTAÇÃO VISUAL
<PROSEGV_vazou>
ETIQUETAS DE ANÁLISE DE SINTAGMA NOMINAL (SN)
<SN_pre-nucleares+subst>
<SN_nominal+modif/modif+nominal>
<SN_superlativo+adj>
<SN_relativo+oração incompleta>
<SN_nome próprio>
<SN_título+nome próprio>
<SN_colocações/idiom/conv>
ETIQUETAS DE ANÁLISE DE SINTAGMA PREPOSICIONADO (SP)
<SP_prep+subst>
ETIQUETAS DE ANÁLISE DE SINTAGMA VERBAL (SV)
<SV_verbo+verbo>
<SV_verbo+adv>
<SV_colocações>
<SV_negação+verbo>
<SV_(verbo)+obliquo+verbo>
ETIQUETAS DE ANÁLISE DE SINTAGMA ADVERBIAL (SAdv)
<SAdv>
ETIQUETAS DE ANÁLISE DE SINTAGMA ADJETIVO (SAdj)
<SAdj_subst+adj>
ETIQUETAS DE ANÁLISE DE ORAÇÃO COORDENADA (COORD)
<COORD_coordenador+oração>
<COORD_negativa>
ETIQUETAS DE ANÁLISE DE ORAÇÃO SUBORDINADA (SUBORD)
<SUBORD_conj+oração>
<SUBORD_se>

# WordSmith Tools – Concordance lines

The screenshot displays the WordSmith Tools interface. The main window shows a concordance search for the phrase "if nouns are bricks, are verbs mortar?". A smaller window titled "Concord" is open, displaying a list of concordance lines. The table below represents the data shown in this window.

N	Concordance	Set	Tag	Word #	t.	#	os.	#	os.	#	os.	File
1	01.06.34.013</t> <cpl11>E se eu não<PROSEGG><SV_negação+verbo>			10,836	264	7%	1	6%	0	9%	tiquetadoTXT.txt	
2	<cpl18>E se eu não quiser<PROSEGG><SV_verbo+verbo>			10,852	265	8%	1	6%	0	9%	tiquetadoTXT.txt	
3	a lidar com as separações<PROSEGG><SAdj_subst+adj>			10,773	264	1%	1	5%	0	9%	tiquetadoTXT.txt	
4	<cpl14>O ministro vai<PROSEGG><SV_(verbo)+>			10,163	252	7%	1	8%	0	5%	tiquetadoTXT.txt	
5	--> 01.03.21,120</t> <cpl8>Há casos<PROSEGG><SAdj_subst+adj>			10,257	254	9%	1	9%	0	6%	tiquetadoTXT.txt	
6	<cpl22>e ditar para os nossos<PROSEGG><SN_pre-nucleares+subst>			11,356	271	8%	2	4%	0	3%	tiquetadoTXT.txt	
7	<cpl26>que vocês vão permitir que<PROSEGG><SUBORD_conj+oração>			11,389	271	9%	2	5%	0	3%	tiquetadoTXT.txt	
8	<cpl27>[ Emmanuel ] Este livro que<PROSEGG><SN_relativo+oração>			11,339	271	3%	2	4%	0	2%	tiquetadoTXT.txt	
9	<cpl20>Não vou esperar anos<PROSEGG><SAdv> <cpl21>e anos			11,150	269	4%	1	0%	0	1%	tiquetadoTXT.txt	
10	<cpl22>Agora sou eu que estou<PROSEGG><SV_verbo+verbo>			11,167	269	6%	2	0%	0	1%	tiquetadoTXT.txt	
11	necessário um planejamento<PROSEGG><SAdj_subst+adj>			10,114	250	7%	1	8%	0	5%	tiquetadoTXT.txt	
12	<cpl14>Arrependimento<PROSEGG><SN_nominal+>			8,416	220	6%	1	8%	0	4%	tiquetadoTXT.txt	
13	<cpl27>Tudo perde o sentido quando<PROSEGG><SUBORD_conj+oração>			8,653	226	4%	1	1%	0	5%	tiquetadoTXT.txt	
14	<cpl14>Senhor, eu vim<PROSEGG><SV_verbo+verbo>			8,251	219	8%	1	6%	0	3%	tiquetadoTXT.txt	
15	<cpl27>os seres precisam de alguma<PROSEGG><SN_pre-nucleares+subst>			8,066	217	3%	1	4%	0	2%	tiquetadoTXT.txt	
16	<cpl19>vamos entrar e você<PROSEGG><COORD_coord>			8,135	218	8%	1	4%	0	2%	tiquetadoTXT.txt	
17	--> 01.01:39,260</t> <cpl9>Não quero<PROSEGG><SV_verbo+verbo>			9,807	244	8%	1	4%	0	3%	tiquetadoTXT.txt	
18	<cpl19>Há espaço para mais<PROSEGG><SN_pre-nucleares+subst>			9,822	244	5%	1	4%	0	3%	tiquetadoTXT.txt	
19	<cpl22>meu pai demorou apenas<PROSEGG><SAdv> <cpl33>18 anos			9,741	244	7%	1	3%	0	2%	tiquetadoTXT.txt	
20	<cpl24>Você é médico, mas antes<PROSEGG><SN_expressão>			8,783	229	7%	1	2%	0	6%	tiquetadoTXT.txt	

At the bottom of the Concord window, there are tabs for "concordance", "collocates", "plot", "patterns", "clusters", "filenames", "follow up", "source text", and "notes". The "concordance" tab is currently selected.

<SV\_verbo+verbo>



O amigo parece ter  
compreendido o sentido da água,

O amigo parece ter/compreendido o sentido da água,



## <SV\_negação+verbo>



Logo eu! Logo eu que nem/acreditava em vida após a morte

<SP\_prep+subst>



O que sabe sobre/a medicina espiritual?

<SAdv>

---



Não vou esperar anos/e anos para retornar.

---



## <SUBORD\_conj+oração>



Todo o ceticismo termina quando/se acorda no mundo espiritual.

## <SUBORD\_conj+oração>



Lísias, eu não vou parar enquanto/não souber o que está acontecendo.

# Findings

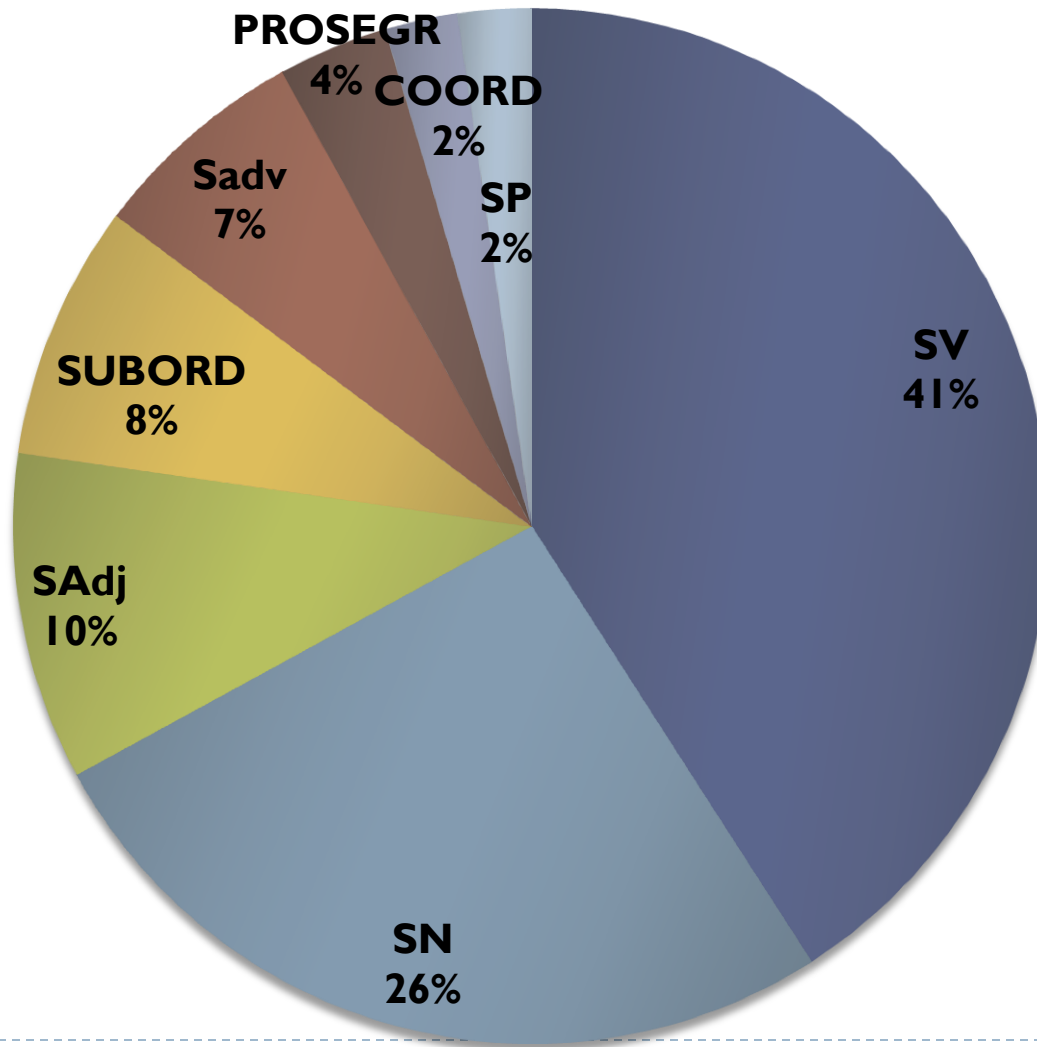
---

- ▶ Linguistic segmentation problems occurred in the levels of noun, verb, prepositional, adjective and adverbial phrases, and in the level of coordinated and subordinated clauses whenever there is a break in the internal structure of this linguistic constituents



# Findings

---



# Our emails and blogs

---

[verainnerlight@uol.com.br](mailto:verainnerlight@uol.com.br)

[elidagama@hotmail.com](mailto:elidagama@hotmail.com)

[www.leaduece.blogspot.com](http://www.leaduece.blogspot.com)

[www.atavbrasil.blogspot.com](http://www.atavbrasil.blogspot.com)

