

## 4. Caso de uso 04 - Criação, configuração e edição dos corpora

A ferramenta de criação de corpora é um componente fundamental para a construção e gestão eficiente de coleções de textos destinados a análises linguísticas. Este processo pode ser realizado por Administradores e usuários devidamente cadastrados e com as credenciais corretas (de edição).

Para os administradores, a ferramenta oferece recursos avançados para a criação e configuração dos corpora, atribuição de permissões e gerenciamento etc.

Há três formas [REVISAR] para executar a criação de corpora na ferramenta Tycho Brahe:

1. *Translation mode* (Modo traduções) ou *Translation edictor*
2. [REVISAR: INSERIR LISTA DE MANEIRAS DE DEV DE CORPUS] Antes de ver em detalhe cada uma das maneiras de criação de corpora, é necessário apresentar quais são as opções disponíveis para configuração dos corpora.

A seguir, é apresentada uma descrição dos botões de configuração dos corpora (acesso pela ferramenta "Corpora Management"). O acesso à área de configuração dos corpora pode ser realizada a partir da página inicial da plataforma em <https://www.tycho.iel.unicamp.br/home>, acessando-se a ferramenta "Corpora Management", como apresentado na Figura x, a seguir:

The screenshot shows the Plataforma Tycho Brahe interface. At the top, there's a navigation bar with links like 'Corpus Histórico', 'Inovação', 'Ferramentas', 'Projetos', 'Manuais', 'Instituições', and 'Contato'. A dropdown menu indicates 'PT-BR'.

**Ferramentas (Tools) section:**

- Visualizador**: Utilizado para visualização dos corpora disponíveis.
- eDitor**: Um editor de textos (eDitor) que inclui edição e revisão de textos e sentenças com diferentes interfaces dependendo das parametrizações do corpus: transcrição de imagens, áudios, traduções, etc.
- Syntrees**: Ferramenta gráfica para revisão de estruturas sintáticas com uma interface intuitiva e integrada ao parser.
- Syntviewer**: Ferramenta para converter expressões de coletivas sintáticas em árvores gráficas. É integrado ao analisador e exporta árvores para imagens.
- Parser**: Mecanismos para parser sintático e etiquetagem POS.
- Search**: Ferramenta filológica e morfossintática para corpora anotados.
- IO**: Importação e exportação de dados, funcionando como módulo de interoperabilidade entre diferentes formatos de dados utilizados por outras ferramentas.
- Corpora Management**: Acesso e configuração geral de parâmetros. This module is highlighted with a red dashed border.

**Corpora Management configuration page:**

1 - Role a página até a área de "Ferramentas" e selecione "Corpora Management".  
 2 - Crie um novo corpus em "Create new corpus" ou selecione um previamente criado.  
 3 - A seleção de "Parâmetros" abre a caixa de botões de alternância para configuração básica dos corpora.

Type	Status	No. of documents	No. of words	No. of users
PUBLIC	ACTIVE	4	78232	2
PRIVATE	ACTIVE	7	16533	23
PUBLIC	INACTIVE	16	5356	1
PUBLIC	ACTIVE	1	84	3
PUBLIC	ACTIVE	16	2807	7
PRIVATE	INACTIVE	1	4518	1
PUBLIC	INACTIVE	3	0	4
PRIVATE	INACTIVE	2	0	0
PRIVATE	INACTIVE	1	7565	0
PUBLIC	INACTIVE	3	1883	2

Figura x: Acessando área de configuração de corpora

- Public corpus: seleciona se o corpus é disponível ao público em geral ou se é privado;
- Active: configura se o corpus está ativo;
- Featured corpus: atualmente não está sendo utilizado [REVISAR QUAL A FUNCIONALIDADE]
- Use split: habilita o uso campos referentes morfemas, incluindo glossa, nos corpora.

The screenshot shows the eDictor interface with a red arrow pointing to the 'use split' button in the 'Áudio' (Audio) section. The 'Áudio' section contains a text area with Portuguese and English translations, and a 'Syntactic Rules' section. The main part of the interface is a grid-based morphological analysis table.

original	ica	ejiwajegi	inoa	icoa	ane	doitalo	me	yeligo
etiqueta POS	D	N	Q	D	WPRO	VBAPL	C	VB
gloss-br	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio
gloss	o	ejiwajegi	inoa	icoa	ane	doitalo	me	yeligo
morfemas	i Gnr	ca Ncl	ejawa n	jegei Der	na Ncl	wa Plu	ca Ncl	wa Plu
etiqueta	masc	ausente	vazio	nom	masc	vindo	masc	ausente
gloss-br	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio
gloss	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio

Figura X: Configuração de corpora: botão "use split"

- Use Sound: a habilita o uso de áudio no corpus.
  - Como podemos observar na Figura x abaixo, a habilitação do botão "Use Sound" habilita uma seção de Áudio, que permite ao analista incluir o áudio a ser transcritado.

The screenshot shows the eDictor interface with a red arrow pointing to the 'Use Sound' button in the 'Áudio' (Audio) section. The 'Áudio' section contains three buttons: play, stop, and pause. The main part of the interface is a grid-based morphological analysis table.

original	ica	ejiwajegi	inoa	icoa	ane	doitalo	me	yeligo
etiqueta POS	D	N	Q	D	WPRO	VBAPL	C	VB
gloss-br	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio
gloss	o	ejiwajegi	inoa	icoa	ane	doitalo	me	yeligo
morfemas	i Gnr	ca Ncl	ejawa n	jegei Der	na Ncl	wa Plu	ca Ncl	wa Plu
etiqueta	masc	ausente	vazio	nom	masc	vindo	masc	ausente
gloss-br	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio
gloss	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio

Figura X: Configuração de corpora: botão "Use Sound"

- O "Use Sound" também habilita, na ferramenta "edictor", um botão para dar play nos áudios associados às sentenças do documento selecionado, como observamos na Figura x abaixo:

Documento:	Ejijwajegi dinibolodi	
1	ica ejijwajegi inoa ane doitalo me yeligo	Olam kadiéwu tem isso que é um medo de comer (algumas coisas)
2	ejijwajegi aona yeligo tabidatGa	Diz que o kadiéwu não come raspa queimada.
3	atone yoe lotidi ica iwaalo	pois se diz/acredita (surpreendentemente) que a mulher não dá leite (se comer)
4	codaa me iloneeGa ayeligo	Dal que pessoas jovens não comem isso.
5	niale ela novadi eleci ayeligo	Outra coisa que não sei
6	dotigí daGia owadi nige dirigajae	O temor é de parir gêmeos
7	odaa aagnaaGa iloneeGa idaGee ayeligo	Então os homens jovens
8	dotigí deGetacdeteca aca lodawa	Eles tem medo de judar da esposa.
9	niGina lati waca nigaraungi iloneGaa awikje aGoikateke moyeligo	As crianças, moças e moços jovens, não podiam comer pâncreas de vaca.
10	atoneo me jomolociaGati niGina me iweniti ica oko	Porque não deixa afundar quando você mergulha pessoas.
11	tibige leeGodí ica me idinaGataGatinig ika ninyoGodí	Talvez porque escondiamos na água.
12	eleci aonaGa oyeligo niGina beGee nigaanjipswaanigi	Outra coisa que também não come enquanto é criança:
13	niGina latikilo libitagi	o tutano do osso.
14	one doita ica ejijwajegi daGia dilaike	Diz que o kadiéwu também tem medo de ter cabelos grisalhos.
15	niGina waca lotidi iwkakapadi eleci odolenaGatidi ica beGee noji	Outra coisa que eles colocavam medo enquanto novo é de tomar a nata do leite da vaca.
16	leeGodí one iwkakapadi latobi	Porque diz que enrugava o rosto.
17	nige yeligo	Quando come.

Figura X: Configuração de corpora: botão "Use Sound"(no edictor)

- Use translations: o botão "use translations" habilita a seção de Traduções no edictor, como apresentado na Figura x:

Plataforma Tycho Brahe: eDitor

Corpus Kaduwé

Documento: Ejiwajegi dinibolodi

UD Parse + Adicionar nova Remover

ica ejiwajegi inoa icoa ane doitalo me yeligo

original	ica	ejiwajegi	inoa	icoa	ane	doitalo	me	yeligo
etiqueta POS	D	N	Q	D	WPRO	VBAPL	C	VB
gloss-br	vazio	vazio	yazio	vazio	vazio	vazio	vazio	vazio
gloss	ø	kaduwéu	um	isso	que	tem medo de	que	comer
morefemas	!	ca	ejawa jegi	na wa	ca wa	d qí talo	x	eligo
etiqueta	Gnr	Ncl	n Der	Gnr Ncl Plu	Gnr Ncl Plu	Inv y Apl	Erg y	y
gloss-br	masc	ausente	vazio nom	masc yendo	masc ausente	vazio vazio	vazio	vazio
gloss	vazio	vazio	vazio vazio	vazio	vazio	vazio vazio	vazio	vazio

Áudio

O botão "use translations" habilita a seção de traduções na ferramenta editor (sentença selecionada no documento)

Traduções

Português

Olum kaduwéu tem isso que é um medo de comer (algumas coisas)

English

Syntactic Rules

estrutura Sintática

Figura x: Configuração de corpora: botão "use translations"

- Use Lexicon: é utilizado quando há um parser disponível, mas não há um etiquetador automático (que só funciona se houver uma quantidade mínima de palavras para treinamento do etiquetador). No caso de esta condição não estar satisfeita, o analista deve utilizar o léxico para realizar a etiquetagem automática.
  - Use Grid: ao acessar o catálogo, o usuário tem a possibilidade de apresentação de documentos de duas maneiras: lista, ou grid. Ao habilitar o botão "Use Grid", o usuário configura a apresentação em grid como default, como apresentado na Figura x abaixo:



Figura x: Configuração de corpora: botão "Use Grid"

- Use Parser: [REVISAR: O Luiz mencionou que iria revisar se este botão está sendo utilizado..a mesma funcionalidade está presente na aba à esquerda]. Este botão habilita a utilização de parser.
- Use Category: o botão "Use Category" habilita a possibilidade de categorização e subcategorização dos corpora (por exemplo, categorização com base em parâmetros demográficos etc), como apresentado na Figura x, com o corpus CE-DOHS selecionado no catálogo. Note-se que a criação dos rótulos para as categorias pode ser realizada por um usuário com permissões de administrador

Categoria	Descrição	No. Documentos
Cartas para vários destinatários (a partir de 1724)	Edição fac-similar e semidiplomática composta por 208 cartas e 114 remetentes, extraídas de Carneiro (2005). Essas cartas, dirigidas a diversos destinatários, estão depositadas no Instituto Geográfico...	100
Cartas para Cícero Dantas Martins, Barão de Jeremoabo (a partir de 1850)	Edição fac-similar e semidiplomática composta por 190 cartas e 43 remetentes. Essas cartas, dirigidas a Cícero Dantas Martins, Barão de Jeremoabo, estão depositadas no Centro de Documentação da Fundação...	190
Cartas para Severino Vieira, Governador da Bahia (a partir de 1850)	Edição fac-similar e semidiplomática composta por 172 cartas e 40 remetentes, extraídas de Santiago (2019). Essas cartas, dirigidas a Severino Vieira, Governador da Bahia, estão depositadas no Centro de Documentação da Fundação...	102
Cartas particulares do Recôncavo da Bahia (a partir de 1770)	Trata-se de 158 cartas, emitidas entre 1770 e 1886, uma amostra primorosa para a aplicação de estudos dentro de uma perspectiva sociodiscursiva...	42
Cartas do Acervo Dantas Jr (a partir de 1880)	Edição fac-similar e semidiplomática composta por 242 cartas e 113 remetentes. Essas cartas, dirigidas a João da Costa Pinto Dantas Jr., neto do Barão de Jeremoabo, estão depositadas no Centro de Docu...	242
Cartas Baianas: Acervo do Dr. João da Costa Pinto Victória (a partir de 1860)	Edição fac-similar e semidiplomática composta por 102 cartas e 05 remetentes. Essas cartas, dirigidas a integrantes da família Costa Pinto, estão depositadas no arquivo Particular do Dr. João da Costa...	101
Correspondências Amigas: o Acervo de Valente, Bahia (a partir de 1960)	Edição fac-similar e semidiplomática composta por 94 correspondências, sendo 79 cartas, 15 cartões com escrita do remetente, além de mais 9 cartões apenas com ilustrações impressas e 38 remetentes. Es...	94
Cartas em Sisal/Mãos Cândidas/Cartas de Inábels (a partir de 1870)	Edição fac-similar e semidiplomática composta por 131 cartas, escritas por 53 remetentes, extraídas de Santiago (2019). Essas cartas foram escritas por sertanejos oriundos de comunidades rurais dos mu...	131
Cartas do Acervo Particular da Família Soledade (a partir de 1900)	Conjunto de 100 cartas manuscritas (apenas 1 é datilografada), trocadas entre baianos cultos, Otto Soledade Júnior e René da Silva Barros Soledade ,	29

Figura x: Configuração de corpora: botão "Use Category"

- Use Edition Tiers: o botão "Use Edition Tiers", contrapondo-se ao "use split", habilita a disponibilização de diversos campos para edição, incluindo POS, spelling etc. Por exemplo, acessando-se o catálogo, selecione o corpus Tycho Brahe do Português Histórico e selecione o documento "Atas dos Brasileiros - Tomo 02", como apresentado na Figura X:

Plataforma Tycho Brahe: Catálogo

+ Adicionar novo documento

Ref.	Nome	Autor	Páginas	Status	Data de inclusão	Status da revisão
b_010.xml	A Morgadinha de Val-d'Amores	Camilo Castelo Branco	130	FINALIZADO	2023-11-20	<div style="width: 100%;">total: 1784 completado: 0.00 % 0 951 833 9</div>
c_009.xml	A Morgadinha de Valfior	Manuel Pinheiro Chagas	174	FINALIZADO	2023-11-20	<div style="width: 100%;">total: 3667 completado: 0.00 % 0 159 1910 9</div>
c_006.xml	A arte de furtar	Manuel da Costa	157	FINALIZADO	2023-11-11	<div style="width: 100%;">total: 1274 completado: 0.00 % 0 9 1274</div>
c_008.xml	A inauguração da estátua equestre	Cascais, Joaquim da Costa	154	FINALIZADO	2023-11-20	<div style="width: 100%;">total: 4899 completado: 0.00 % 0 119 1083 2625</div>
f_003.xml	A partilha	Miguel Falabella	34	FINALIZADO	2023-11-17	<div style="width: 100%;">total: 2700 completado: 0.00 % 0 2 2700</div>
s_001.xml	A vida de Frei Bertolameu dos Mártires	Luis de Sousa	15	FINALIZADO	2023-11-15	<div style="width: 100%;">total: 736 completado: 0.00 % 0 2 736</div>
va_003.xml	Atas dos Africanos	Various	127	FINALIZADO	2023-11-15	<div style="width: 100%;">total: 756 completado: 0.00 % 0 3 756</div>
va_002.xml	Atas dos Brasileiros	Various	18	FINALIZADO	2023-11-18	<div style="width: 100%;">total: 3556 completado: 0.00 % 0 0 3556</div>
Atas dos Brasileiros - Tomo 02	Atas dos Brasileiros - Tomo 02	Various	101	FINALIZADO	2023-11-14	<div style="width: 100%;">total: 1696 completado: 5.17 % 24 1036 624</div>
o_002.xml	Aventuras de Diófanes	Teresa Margarida da Silva e Orta	304	FINALIZADO	2023-11-14	<div style="width: 100%;">total: 3617 completado: 0.00 % 0 0 3617</div>

Exibido 10 resultados por página Total: 93

Figura x: Configuração de corpora: botão "Use Edition Tiers"

Ações

- Este documento possui inconsistências. Clique aqui para verificar.
- Continuar a edição do documento**
- Continuar anotação sintática. Continue a partir da última sentença revisada.
- Revisar etiquetas POS
- Revisar estruturas sintáticas
- Imprimir. Impressão de todas as sentenças com seus detalhes (uma sentença por página).
- Remover documento. Atenção! Este processo é irreversível.

Corpus Tycho Brahe do Português Histórico

Ref.	Nome	Autor	Páginas	Status	Data de inclusão	Status da revisão
Branco	A Morgadinha de Val-d'Amores	Camilo Castelo Branco	130	FINALIZADO	2023-11-20	<div style="width: 100%;">total: 1784 completado: 0.00 % 0 951 833 9</div>
Chagas	A Morgadinha de Valfior	Manuel Pinheiro Chagas	174	FINALIZADO	2023-11-20	<div style="width: 100%;">total: 3667 completado: 0.00 % 0 159 1910 9</div>
Costa	A arte de furtar	Manuel da Costa	157	FINALIZADO	2023-11-11	<div style="width: 100%;">total: 1274 completado: 0.00 % 0 9 1274</div>
Falabella	A partilha	Miguel Falabella	34	FINALIZADO	2023-11-17	<div style="width: 100%;">total: 2700 completado: 0.00 % 0 2 2700</div>
Sousa	A vida de Frei Bertolameu dos Mártires	Luis de Sousa	15	FINALIZADO	2023-11-15	<div style="width: 100%;">total: 736 completado: 0.00 % 0 2 736</div>
Africanos	Atas dos Africanos	Various	127	FINALIZADO	2023-11-15	<div style="width: 100%;">total: 756 completado: 0.00 % 0 3 756</div>
Brasileiros	Atas dos Brasileiros	Various	18	FINALIZADO	2023-11-18	<div style="width: 100%;">total: 3556 completado: 0.00 % 0 0 3556</div>
Atas dos Brasileiros - Tomo 02	Atas dos Brasileiros - Tomo 02	Various	101	FINALIZADO	2023-11-14	<div style="width: 100%;">total: 1696 completado: 5.17 % 24 1036 624</div>
Diófanes	Aventuras de Diófanes	Teresa Margarida da Silva e Orta	304	FINALIZADO	2023-11-14	<div style="width: 100%;">total: 3617 completado: 0.00 % 0 0 3617</div>

Figura x: Configuração de corpora: botão "Use Edition Tiers (abrindo a)."'

Com o cursor acima da oração alvo, um clique com o botão direito do mouse abre este pequeno painel. Clicar em "abrir" abre um segundo painel que apresenta a sentença e as anotações salvas.

O cursor destaca a sentença no texto.

Imagem não disponível

Aos dezoito dias do mês de junho de mil e oitocentos e trinta e sete estando o provedor e mais mesários da nossa devoção leu-se o termo do que ficou aguiado

original | Aos | dezoito | dias | do | mês | de | Junho | de | mil | e | oitocentos | e | trinta | e | sete | estando | o | provedor

POS tag | P+D-P | NUM | N-P | P+D | N | P | N | P | NUM | CONJ | NUM | CONJ | NUM | CONJ | NUM | ET-G | D | N | CONJ

- Selecionar este botão para abrir o quadro completo para edição das camadas.

Figura x: Configuração de corpora botão "Use Edition Tiers:"

Esta ação abre um painel com uma matriz para edição das camadas de edição, como mostra a Figura x. Note-se: na mesma tela que seria análoga à sentença com o "use split" selecionado, são apresentadas as camadas de edição.

	*exp*	Aos	osdezoito	dimas	do	mes	de	Junho	de	mil	e	oitocentos	e	trinta	e	sete	estando	o
original	vazio	P+D-P	NUM	N-P	P+D	N	P	N	P	NUM	CONJ	NUM	CONJ	NUM	CONJ	NUM	ET-G	D
etiqueta POS	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	
grafia	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	
ilegível	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	
expansão	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	
modernização	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	
correção	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	
padronização	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	
pontuação	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	
flexão	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	vazio	

Áudio não disponível

Idiomas não disponíveis para tradução.

Estrutura Sintática

#### Figura x: Matriz para edição em camadas

- Use Translations/Use Edictor/Use Designer/Use Transcriber/Use Syntrees/: Estes cinco botões não são excludentes, i.e., podem ser habilitados todos ao mesmo tempo em um determinado corpus. Esta ação disponibiliza todas as ferramentas para utilização no corpus. (Haverá seções específicas para cada um dos botões)

Nos tópicos a seguir, serão abordados os principais fluxos de trabalho para a criação de novos corpora, além de tutoriais detalhados que guiam administradores e usuários no uso eficiente da ferramenta, garantindo uma experiência fluida e colaborativa na gestão dos catálogos.