

Carolina: um C rpus Geral do Portugu s Brasileiro com Proveni ncia, Tipologia e Versionamento

Felipe Serras e Mariana Sturzeneker



EQUIPE



L^a



C^a

O **Córp**us **Carolina**

- **Córp**us Geral do Português Brasileiro com Proveniência, Tipologia e Versionamento
 - Tipologia diversa
 - Volume robusto
 - Português brasileiro contemporâneo (pós 1970)
- LaViHD/C4AI-NLP2



Carolina Michaëlis de Vasconcelos

- 1851-1925
- Filóloga
- Primeira Mulher Professora
Universitária em Portugal




Versão Atual



- Versão 1.2 Ada (8 de março de 2023)
 - 823 milhões de tokens
 - 2 milhões de textos
 - 11 GBs







Acesso ao corpus: Hugging Face














Datasets: carolina-c4ai / **corpus-carolina**  3

Tasks:  Fill-Mask  Text Generation Sub-tasks: **masked-language-modeling** **language-modeling** Languages: **Portuguese** Multilinguality: **monolingual**

Size Categories: **1B< n < 10B** Language Creators: **crowdsourced** Annotations Creators: **no-annotation** Source Datasets: **original** License:  **cc-by-nc-sa-4.0**

 Dataset card  **Files and versions**  Community 3

 main  corpus-carolina / corpus  3 contributors  History: 9 commits

 guilhermellemello	Replace white spaces on folder names.	ac16cec	13 days ago
 datasets_and_other_corpora	Replace white spaces on folder names.		13 days ago
 judicial_branch	Replace white spaces on folder names.		13 days ago
 legislative_branch	Replace white spaces on folder names.		13 days ago
 public_domain_works	Replace white spaces on folder names.		13 days ago
 social_media	Replace white spaces on folder names.		13 days ago
 university_domains	Replace white spaces on folder names.		13 days ago



Acesso ao corpus: Portulan Clarin



**PORTULAN
CLARIN****Infraestrutura de Investigação para a Ciência e Tecnologia da Linguagem**

Repositório Bancada Apoio Alcance

pt ▼

[Home](#) / [Repository](#) / [Resource detail](#) *indisponível em Português*

Carolina: General Corpus of Contemporary Brazilian Portuguese with provenance and typology information

[View resource name in all available languages](#)

Carolina: Corpus Geral do Português Brasileiro Contemporâneo com informações de procedência e tipologia

Carolina | *Carolina*

Handle: <https://hdl.handle.net/21.11129/0000-000F-486A-A> (persistent URL to this page)

URL: <https://sites.usp.br/corpuscarolina/> 

Carolina is an open corpus for Linguistics and Artificial Intelligence with a robust volume of texts of varied typology in contemporary Brazilian Portuguese (1970-2021).

[View resource description in all available languages](#)

Carolina é um corpus aberto para Linguística e Inteligência Artificial com um volume robusto de textos de tipologia variada em Português Brasileiro contemporâneo (1970-2021).

[◀ Back](#) [Download](#)



Motivações

- Corpus robusto para Linguística e Inteligência Artificial
- Grandes modelos requerem grande quantidade de dados
 - Controle de licenças está se tornando crítico
- Viéses e proveniência
- Curadoria dos textos e metadados



Pilares de construção

- Proveniência
- Tipologia
- Versionamento
- Integralidade textual



WaC-wiPT

- Investigação prévia de outras metodologias
 - WaC X Córpus tradicionais
- Criação da WaC-wiPT



Levantamentos

- Levantamentos: pesquisas aprofundadas dos domínios disponíveis
- Divididos por tipologia ampla (*broad typology*):
 - Judiciário
 - Legislativo
 - Datasets e Outros Corpora
 - Domínio Público
 - Wikis
 - Portais de Universidades
 - Redes Sociais



Tipologias e Domínio

- Três tags de ordenação dos textos:
 - Tipologia Ampla (*broad typology*)
 - Tipologia da Fonte (*source typology*)
 - Domínio (*domain*)





Cabeçalho

- TEI (Text Encoding Initiative)
 - Link para o texto original
 - Data de download
 - Licença
 - Autoridade



Histórico do Córpus

Versão 0.9 - Embrião

Versão 1.0 - Ada

Versão 1.1 - Ada

Versão 1.2 - Ada

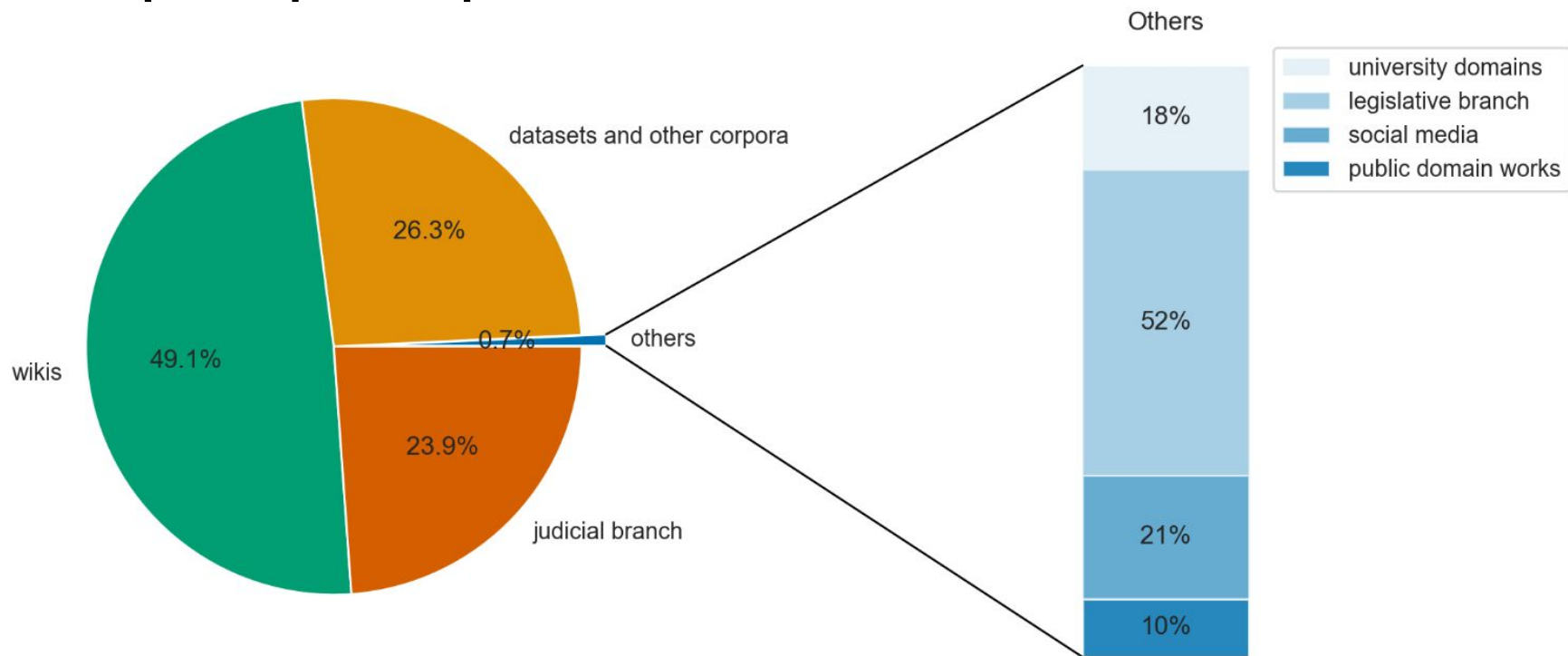


Carolina versão 1.2 Ada

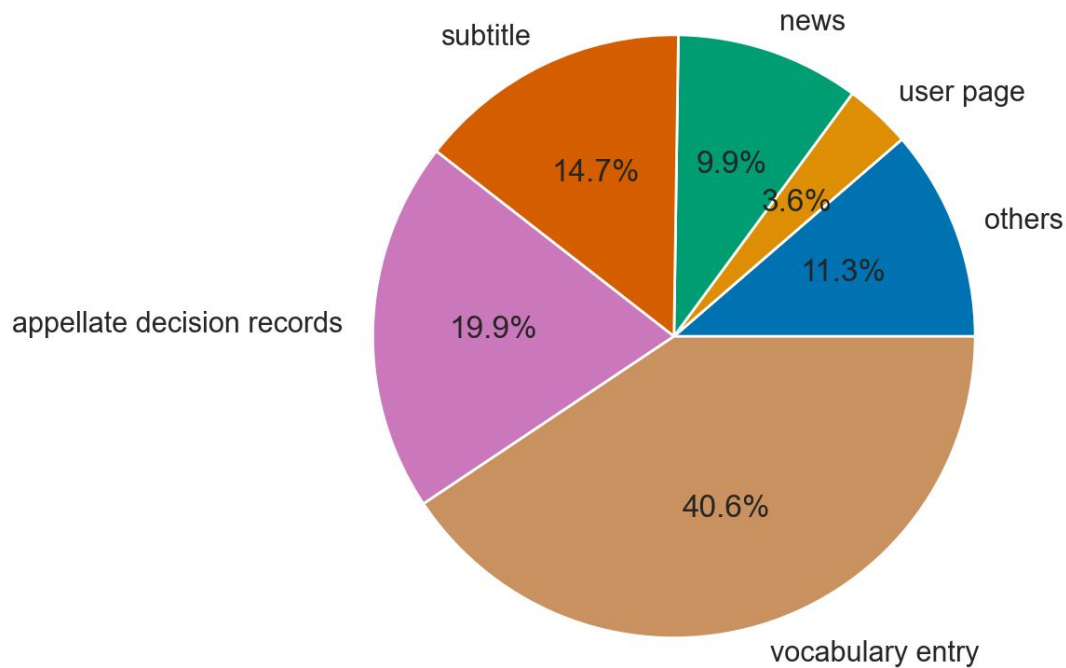
Broad typology	Size (GB)	Number of texts	Number of tokens
Datasets and other corpora	4,4	1.102.049	216.696.066
Judicial branch	1,6	40.464	196.452.059
Legislative branch	0,025	13	3.162.474
Public domain works	0,005	26	601.465
Social media	0,018	3.413	1.280.717
University domains	0,011	941	1.078.967
Wikis	5,3	960.139	403.927.145
Total	11,36	2.107.045	823.198.893



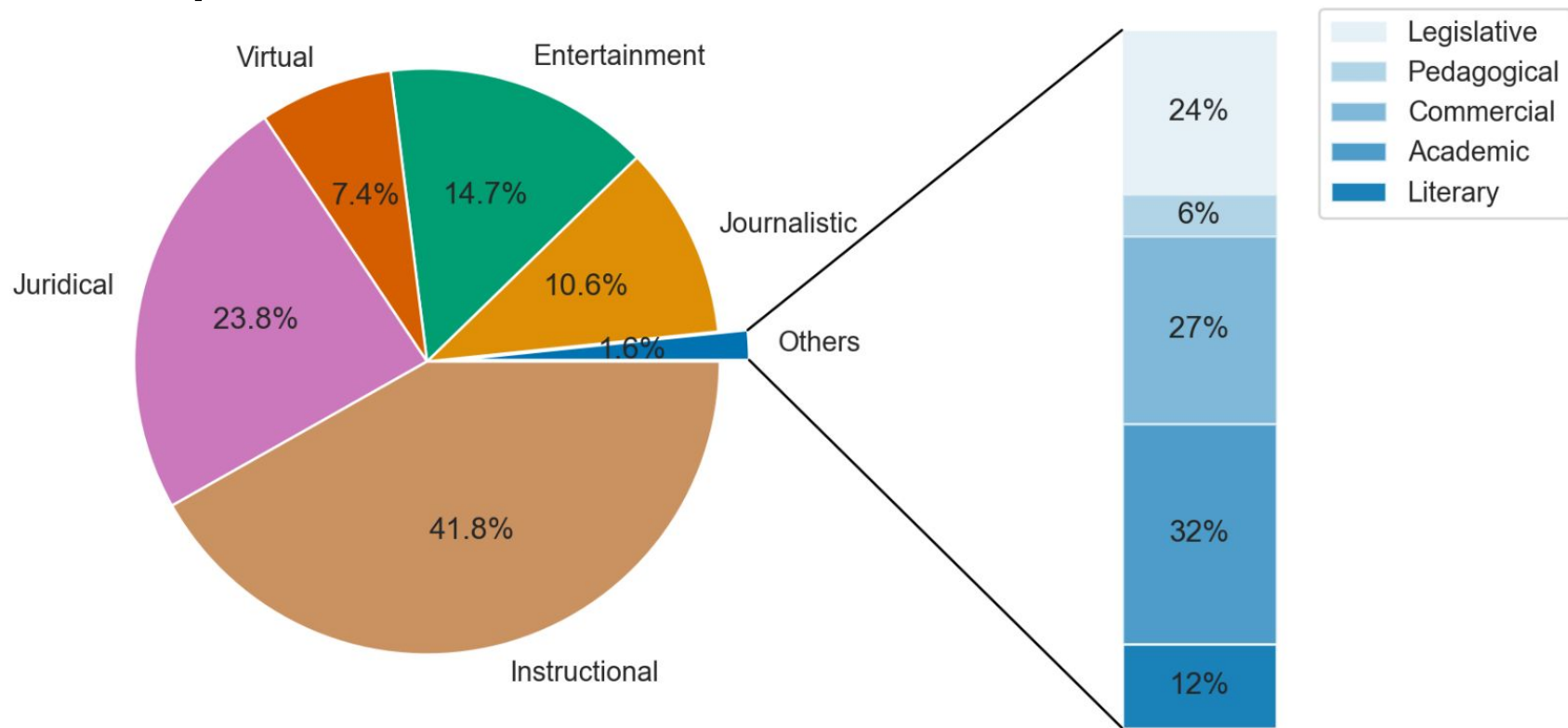
Tokens por tipo amplo



Tokens por tipo da fonte



Tokens por domínio



Conclusão

- Córpus em desenvolvimento contínuo
- Conjunto único de propriedades
- Disponível publicamente



Próximos passos

- Trabalhos derivados
- Criação de uma interface gráfica
- Criação de sub-córpus
- Aumento do volume de textos e balanceamento



Imagens

[Retrato Carolina Michaelis de Vasconcelos](#), disponível em Domínio Público.

[Retrato de Ada Lovelace](#), disponível em Domínio Público.

Template de slides criado por Maria Clara Crespo e Maria Lina Rocha.



Referências

Basseto, B. F. 2015. Breves considerações sobre *Lições de Filologia Portuguesa* de Carolina Michaëlis de Vasconcelos. In **Condé, V. G., Mongelli, L. M., and Vieira, Y. F.**, editors, *Carolina Michaëlis de Vasconcelos: uma homenagem*. FFLCH/USP, São Paulo.

Crespo, M. C., Rocha, M. L., Sturzeneker, M., Serras, F., Mello, G., Costa, A., Palma, M., Mesquita, R., Guets, R., Silva, M., Finger, M., Paixão de Sousa, M. C., Namiuti, C., and Monte, V. 2023. Carolina: a General Corpus of Contemporary Brazilian Portuguese with Provenance, Typology and Versioning Information. In *arXiv*. DOI: <https://doi.org/10.48550/arXiv.2303.16098>

Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., Zasina, A. J., and Benko, V. 2020. Comparing web-crawled and traditional corpora. *Language Resources and Evaluation*, 54(3):713–45.

Sturzeneker, M., Crespo, M. C., Rocha, M. L., Finger, M., Paixão de Sousa, M. C., do Monte, V. M., and Namiuti, C. 2022. Carolina's methodology: building a large corpus with provenance and typology information. In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing*, pp. 53–8.

TEI Consortium 2021. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.2.2 Last updated on 9th April 2021. Retrieved May 20, 2021 from <https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>.



Obrigada!



<https://sites.usp.br/corpuscarolina/>

lavihd@usp.br