Do português histórico ao kadiwéu: uma plataforma multilingue de corpora anotados

Charlotte Galves, Departamento de Linguística, Unicamp
Em colaboração com Filomena Sandalo, Departamento de Linguística, Unicamp

O Corpus Tycho Brahe do português histórico

http://www.tycho.iel.unicamp.br/~tycho/

Atualmente associado ao projeto temático de pesquisa:

A lingua portuguesa no tempo e no espaço: Contato linguístico, competição de gramáticas e mudança gramatical

(Processo Fapesp 2012/06078-9)

Tycho Brahe Parsed Corpus of Historical Portuguese

presentation| news|catalog|manual|credits|register |Português

Citation »

Galves, Charlotte, and Pablo Faria. 2010. *Tycho Brahe Parsed Corpus of Historical Portuguese*. URL: http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html.

Apresentação »

The *Tycho Brahe Parsed Corpus of Historical Portuguese* is an electronic corpus of texts written in Portuguese by authors born between 1380 and 1881.

At present, **76** texts (**3,298,577** words) are available for research, with a linguistic annotation system in two stages: **part-of-specch tagging** (**45** texts, a total of **2,008,998** words); and **syntactic annotation** (**20** texts, a total of **877,247** words).

The Corpus has been built within the projects:

- Rhythmic Patterns, Parameter Setting and Language Change (1998-2003)
- Rhythmic Patterns, Parameter Setting and Language Change, Phase 2 (2004-2008)
- Portuguese in space and time: linguistic contact, grammars in competition and parametric change. (since 2012)

|: News & Updates :|
(in Portuguese)

|: Registration :|

|: Contact :|

Access to Texts

[Computational tools page] NEW
[Ordered Lists Catalog]
[Query POS files with CorpusSearch]

Download Complete Corpus (compacted .zip files):

[Complete Corpus, syntactic annotation (latest version)]
[Complete Corpus, syntactic annotation (Galves & Faria 2010 version)]
[Complete Corpus, POS tagging]
[Complete Corpus, no annotation]

Quick access - Authors List »

1. Aires, Matias (1705): Reflexões sobre a Vaidade dos Homens
2. Alencar, José de (1829): Iracema, lenda do Ceará
3. Almeida, João Ferreira de (1629): Novo Testamento: João e Atos
4. Almeida, Manuel Pires de (1597): Poesia e Pintura
5. Alorna, Marquesa de (1750): Cartas, Marquesa de Alorna
6. Alorna, Marquês da Fronteira e (1802): Memórias do Marquês da Fronteira e d'Alorna
7. Barros, André (1675): Vida do apostólico padre Antonio Vieira
8. Barros, João de (1497): Gramática, João de Barros
9. Bernardes, Manuel (1644): Nova Floresta
10. Branco, Camilo Castelo (1826): Maria Moisés
11. Branco, Camilo Castelo (1825): A Morgadinha de Val-d'Amores
12. Brandão, Antonio (1584): Monarchia Lusitana
13. Brito, Bernardo de (1569): Da Monarquia Lusitana
14. Brochado, José da Cunha (1651): Cartas, J.C. Brochado
15. Brun, André (1881): Teatro, André Brun
16. Cascais, Joaquim da Costa (1815): A inauguração da estátua equestre
17. Cavaleiro de Oliveira (1702): Cartas, Cavaleiro de Oliveira
18. Chagas, António das (1631): Cartas Espirituais
19. Chagas, Manuel Pinheiro (1842): A Morgadinha de Valflor
20. Chiado, António Ribeiro (1520): Teatro, António Ribeiro Chiado

Corpus Histórico do Português Tycho Brahe

[g_008]

Gandavo, Pero Magalhães de (1502): *História da Província de Santa Cruz*

Versões para Leitura: Transcrição do texto-fonte / Source text transcription — gerar / generate
(Versions for reading)

Versões para Trabalho: Texto simples / Simple text (texto completo, modern — gerar / generate
(Working versions)

Versões anotadas:
(annotated versions)
» Anotação Morfológica (*POS annotation*)
» Anotação Sintática (*syntactic annotation*)

Visualizar anotação sintática

XML Para ver o xml de base, clique em *"exibir" >"código-fonte"* na barra superior do navegador.
To access the XML code please select "View sourcecode" (or something similar) on your web browser menu.

Este texto foi transcrito a partir de fac-simile digital (http://purl.pt/121)
Follow the link above to access the digital fac-simile.

As anotações do Corpus Tycho Brahe

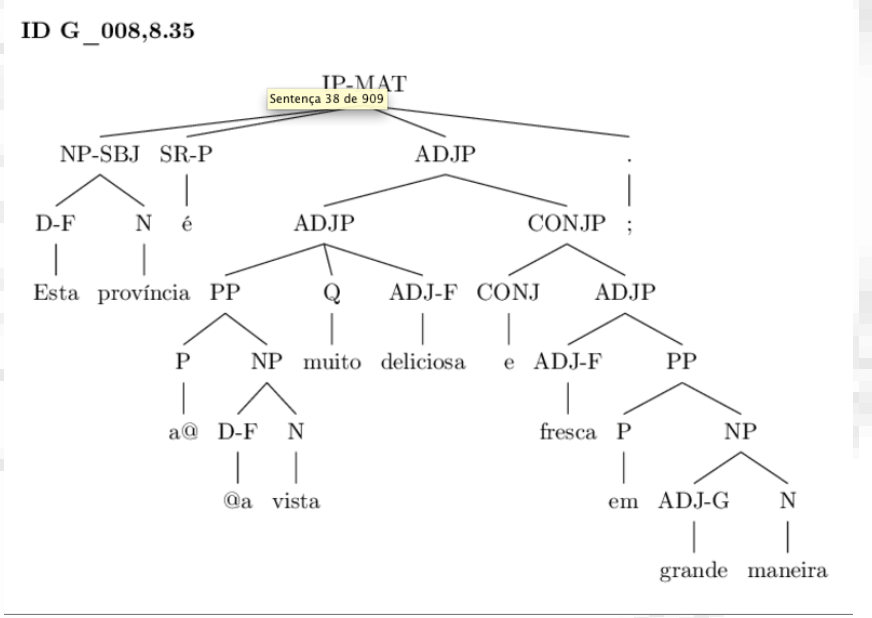
Anotação de classe de palavras

- **Eſta prouincia he á viſta muy delicioſa & freſca em | gram maneira;**

  (Gandavo, Pero Magalhães de (1502): "História da Província de Santa Cruz" (1576)

- Esta/**D-F** província/**N** é/SR-P a@/**P** @a/**D-F** vista/**N** muito/**Q** deliciosa/**ADJ-F** e/**CONJ** fresca/**ADJ-F** em/**P** grande/**ADJ-G** maneira/**N** ;/. G_008,8.35/ID ./PONFP

Anotação sintática



Corpora using the annotation scheme of the Penn-Helsinki Corpora (fonte: http://www.ling.upenn.edu/hist-corpora/other-corpora.html)

- York-Helsinki Parsed Corpus of Old English Poetry
- York-Toronto-Helsinki Parsed Corpus of Old English Prose
- Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English
- Penn-Helsinki Parsed Corpus of Middle English, 2nd edition (PPCME2)
- Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)
- York-Helsinki Parsed Corpus of Early English Correspondence (PCEEC)
- Penn Parsed Corpus of Modern British English, 2nd edition (PPCMBE2)

Corpora using the annotation scheme of the Penn-Helsinki Corpora (fonte: http://www.ling.upenn.edu/hist-corpora/other-corpora.html)

- Tycho Brahe Corpus, a parsed corpus of historical **Portuguese**
  - Charlotte Galves (University of Campinas, Brazil) and collaborators
- Modéliser le changement: les voies du français (Modelling change: the paths of French), a parsed corpus of historical **French**
  - France Martineau (University of Ottawa) and collaborators
- CORDIAL-SIN Corpus, a syntax-oriented corpus of **Portuguese dialects**
  - Ana Maria Martins (Centro de Linguística da Universidade de Lisboa) and collaborators
- **Icelandic** Parsed Historical Corpus (IcePaHC)
  - Eiríkur Rögnvaldsson (University of Iceland) and collaborators
- Word order and word order change in Western European languages (WOChWEL) Corpus, a growing parsed corpus of **Old Portuguese**
  - Ana Maria Martins and Sandra Pereira (Centro de Linguística da Universidade de Lisboa) and collaborators
- Audio-Aligned and Parsed Corpus of **Appalachian English** (AAPCAppE), nearing completion
  - Christina Tortora (City University of New York) and collaborators
- P.S. Post Scriptum - A Digital Archive of Ordinary Writing (**Early Modern Portugal and Spain**), April 2017
  - Rita Marquilhas (Centro de Linguística da Universidade de Lisboa) and collaborators
- NINJAL Parsed Corpus of Modern Japanese (NPCMJ), under construction
  - Prashant Pardeshi (National Institute of **Japanese** Language and Linguistics) and collaborators

Desenvolvimentos recentes

- 2 objetivos:

1) Tornar o corpus inteiramente consultável na web
2) Permitir o desenvolvimento de novos corpora, usando as mesmas ferramentas, com flexibilidade suficiente para incluir línguas e tipos de textos de natureza bem diversa.

A plataforma Tycho Brahe

Ayoreo

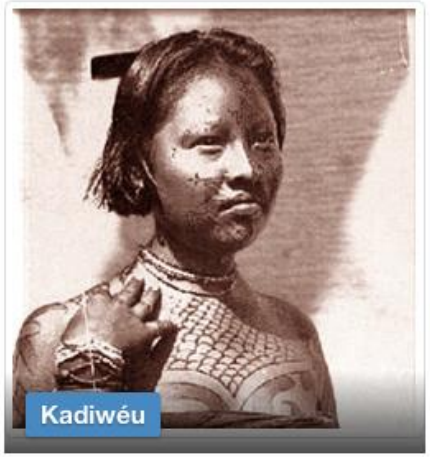
Corpus
Cabinda

Cabinda

Cafundó

Corpus Histórico do Guarani

Corpus Tycho Brahe do Português Histórico



Kadiwéu

Maka

Mocoví

As etiquetas de classe de palavra do Kadiweu

Kadiwéu

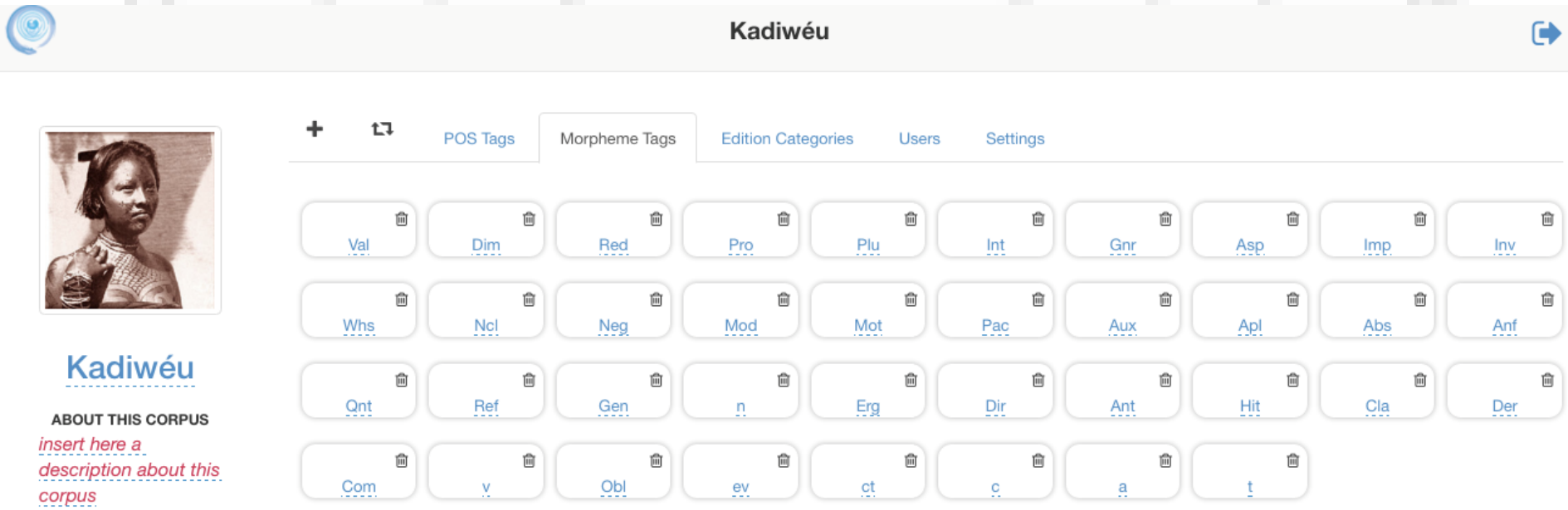
ABOUT THIS CORPUS

insert here a description about this corpus

clone POS tags | POS Tags | Morpheme Tags | Edition Categories | Users | Settings

C+AUX	PRO	CT	ADV	NEG+VB	VB	VB+ADV	C+N	VBAPL	T+VB
NPR	NEG+ADV+D	NEG+ADV	N$	ADJ	AUX	N	C+VB	T+VBT	ADV+VB
T+N	C+D	WD	T	NEG+D	C+VBAPL	MOD+VB	VBTAPL	CT+EV+VBT	CPADV
WPRO	NEG+CT	T+AUX	C	NEG	CONJ	WADV	D	C+D+APL	CT+VBAPL
Q	NUM	NEG+AUX	T+VBAPL	CT+ADV	CONJ+NEG+D	C+NEG	CT+EV+VB	INTJ	CT+NEG+VB
C+ADJ	EV	T+D	CONJ+D	NEG+ADV+NEG+VB	NDIR	CT+VB	T+VBTAPL	VBT	T+ADV
WADV+N									

As etiquetas de morfemas do Kadiweu



As anotações do corpus Kadiweu

Anotações de morfemas e classes de palavras

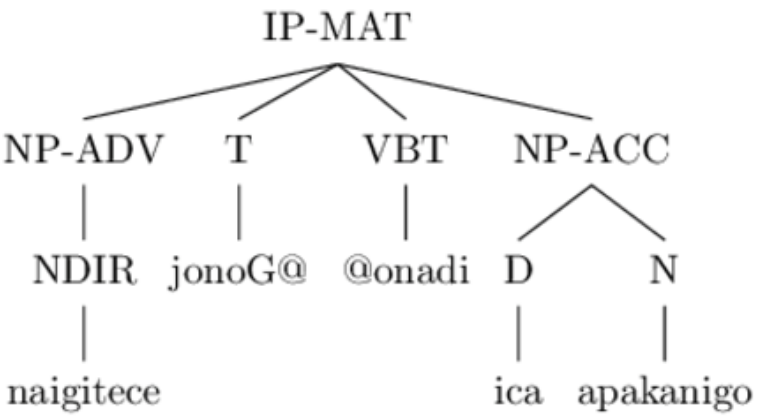
Naigitece jonoGonadi ica apakanigo

(Narrativas Kadiweu -A Mulher onça)

Naigi/**n** tece/**Dir** jonoG/**t** o/**Plu** nadi/**v** i/**Gnr** ca/**Ncl** apakani/**n** go/**Cla**

Naigitece/**NDir** jonoGonadi/**T+VBT** ica/**D** apakanigo/**N**

Anotação sintática



Conclusões

- Os corpora morfossintaticamente anotados são de grande valia para a descrição sincrônica de línguas particulares, o estudo da mudança linguística ou a comparação das línguas – de mesma família ou muito distintas - usando ferramentas computacionais de busca automática.
- O seu tamanho relativamente reduzido – quando comparados com corpora gigantes sem anotação sintática – é compensado pela informação essencial contida na anotação.
- A plataforma Tycho Brahe, em construção, permite usar a mesma metodologia para línguas de tradição escrita e gramatical antiga como o português, e línguas de tradição oral faladas por poucas pessoas, e portanto em perigo de extinção, como o Kadiweu.
- Esse trabalho tem portanto incidência tanto na descrição e análise de línguas ameaçadas, contribuindo a fortalecer sua memória e a favorecer sua preservação, quanto na compreensão das semelhancas e diferenças das línguas de mesmas famílias e de famílias distantes.

Apendix: outras vantagens de corpora do tipo do Corpus Tycho Brahe em comparação com grandes corpora sem anotação sintática

- Perguntas complexas podem ser feitas a respeito da natureza das línguas, e das suas mudanças ao longo do tempo. Isso é impossível se a informação disponível se limita às palavras.
- As buscas automáticas, possibilitadas pela anotação, tornam os resultados mais confiáveis e reprodutíveis.
- Os textos originais são disponíveis, o que permite estudar as sentenças no seu contexto textual-discursivo. Isso é muito relevante, dada a relação entre os processos sintáticos e a organização da informação nos textos.
- A disponibilidade dos textos originais estende o valor do corpus a outras disciplinas (por ex. literatura, história, antropologia)
- Os textos anotados são disponíveis, o que permite que os usuários do corpus avaliem a acurácia da anotação e reportem os erros.

Obrigada!
Gracias!
iniwaagodi!

charlotte.mgc@gmail.com