Recursos Filológicos I: *Treebanking*

2020/2 Seminário em Filosofia Antiga - Humanidades digitais
PPG em Filosofia da UFMG
19/jan/ 2021

Anise D'Orange Ferreira*
Unesp, FCL-Ar

Pesq.Resp.CNPq PQ2 - n.307431/2019-3

Plano da aula

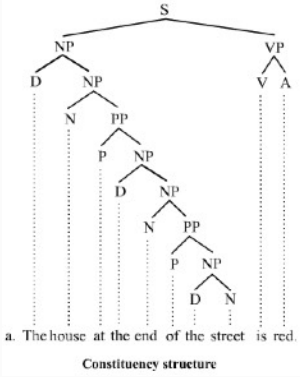
- O que são treebanks e treebanks de dependência sintática
- Treebanks no contexto dos estudos linguísticos e aplicados
- Treebanks do grego antigo
- AGDT - Ancient Greek Dependency Treebank
- Treebanking - Procedimento de anotação em treebank usando o editor ou parser Arethusa, na plataforma Perseids
- Exemplo sobre passagem do diálogo *Parmênides* do Platão, 137d - ss.

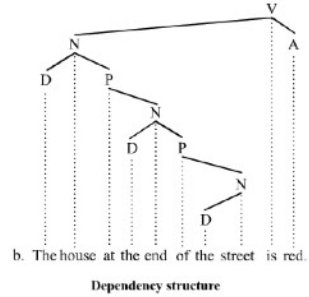
O que são treebanks?

- Chamada de **florestas sintáticas pelos portugueses;** bancos de **árvores sintáticas**

- Definição de Celano, G. (2019)"*treebank* de dependência é um corpus que contém uma representação simbólica da sintaxe de um ou mais textos. Pode ser definido como um conjunto de sentenças separadas (parsed) de acordo com o formalismo linguístico da **gramática de dependência**."

- Mambrini, F. (2011) "Na lingüística contemporânea, o termo 'treebank' é usado para definir um corpus anotado no qual os textos, subdivididos em unidades estruturais (fichas) cada vez menores, são enriquecidos por um conjunto de meta-informações descrevendo a morfologia de palavras únicas e suas relações sintáticas dentro da frase. A natureza das anotações adicionadas também pode incluir outros níveis de análise lingüística e variar de acordo com os propósitos aos quais o corpus se destina."





Para que servem?

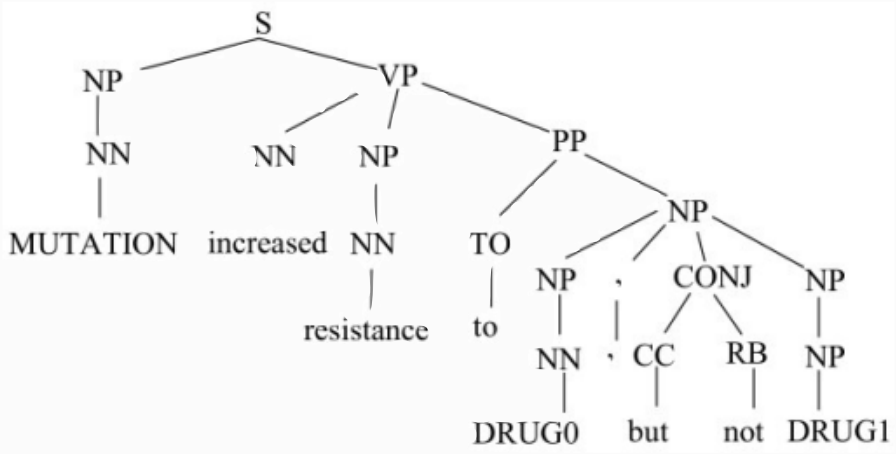
- Ao registrar o(s) percurso(s) sintáticos dos leitores serve ao **ensino da lingua**, à pesquisa **de interpretação textual**, à **tradução** e à **pesquisa linguística** propriamente dita:
- "Treebanks are a powerful resource for **data-driven linguistic research** which are likely to have a great impact on the way the grammar of the ancient languages is studied." (Mambrini, 2016, p. 84).
- "Although some experiments on (semi-)automatic parsing of Latin and Greek have already been carried out, so far, **all the information, including part of speech, morphological features (tense, mood, person etc.), and the syntactic relations between each word in the texts, have been entered manually by human annotators.** This process of word-by-word enrichment **can be facilitated with the help of graphical interfaces and online tools, such as Arethusa**. (Mambrini, 2016, p. 85)

Sintaxe e árvore de dependência

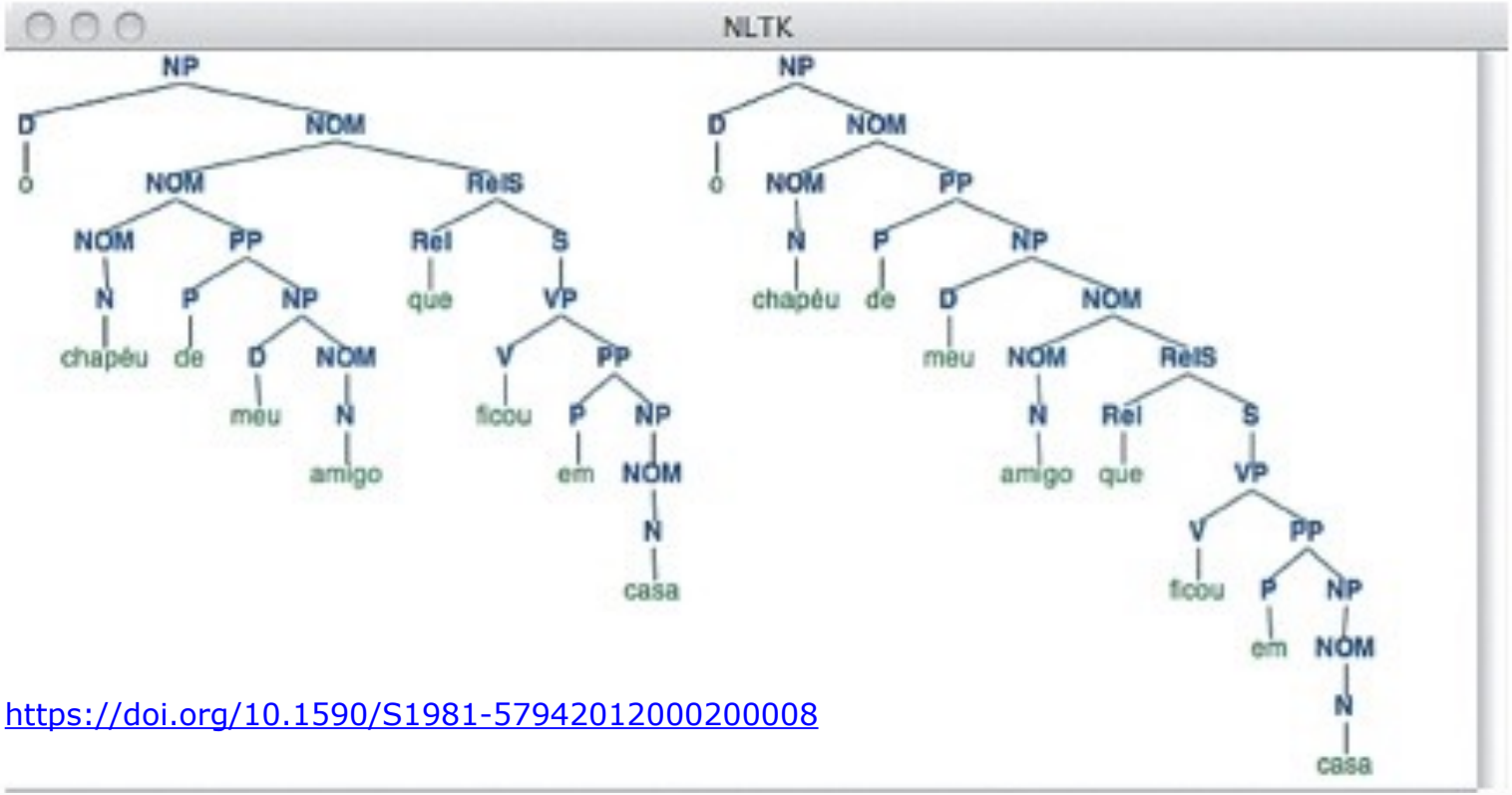
- A sintaxe de dependência tem sua base na ideia de **função gramatical** : sujeito, objeto, etc. e na **relação de dependência**
- A árvore de dependência se presta a línguas com "**descontinuidade**" sintática, i.e., com dependências distantes entre uma e outra palavra da sentença.
- A árvore de dependência tem **estrutura hierárquica** oposta à linear
- Lucien Tesnière (1959) é o **pioneiro** em utilizar o conceito de dependência junto a uma forma de representação sintática hierárquica em árvore.
- Oposta à linearidade da árvore de constituintes da gramática Gerativa de Chomsky.
- Os principais treebanks das línguas histórias são de **dependência**

Árvores de constituintes

- Tem relação de constituintes
- Tem estrutura frasal, contínua e linear
- Tem origem na gramática gerativista do Chomsky

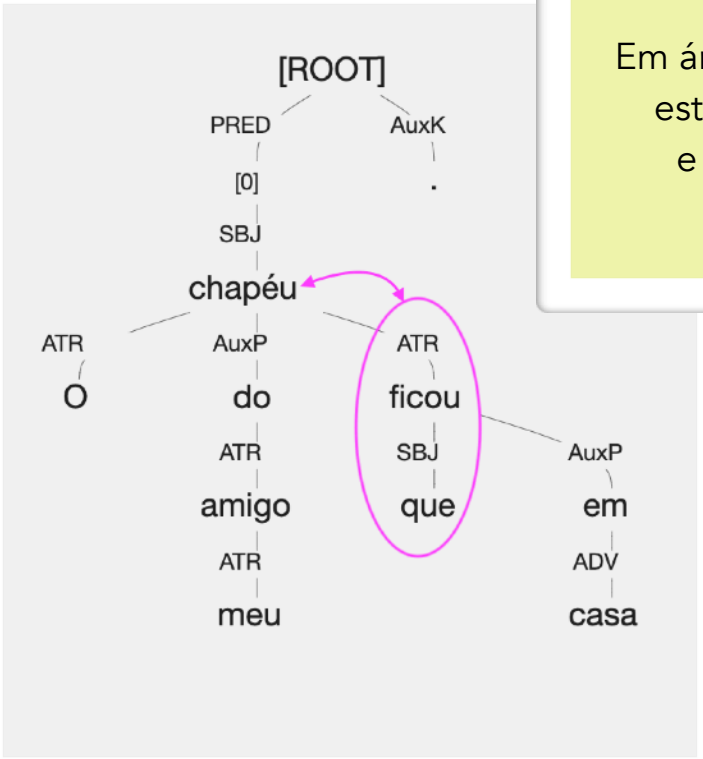


Ex. O chapéu de meu amigo que ficou em casa...?

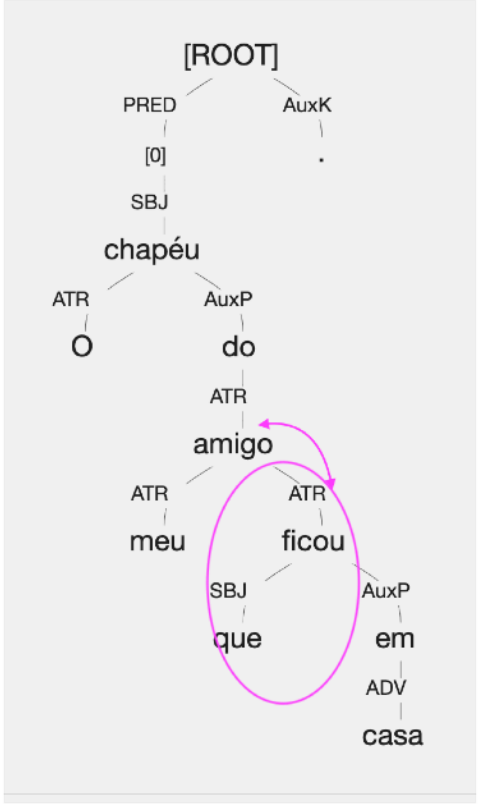


https://doi.org/10.1590/S1981-57942012000200008

Ex. o chapéu de meu amigo que ficou em casa...



Em árvore de dependência, esta frase é incompleta e requer um aT para PRED



Treebanks no contexto dos estudos linguísticos e aplicados

- A sintaxe **não é o único fator** para permitir a leitura e interpretação de um texto. Necessário mas não suficiente
  - situação de produção, receptores, destinatários, contexto sócio-histórico uma visão social de linguagem >> impacta na definição de gênero e nos aspectos pragmáticos do ato comunicativo
  - textos antigos -> produção inclui mss, variantes e transmissão
- Aspectos semânticos > co-texto e dependências sintáticas
- Anotação em *treebanking* depende da elaboração de um TAGSET
  - etiquetas > **POS** part of speech (classe gramatical)
  - etiquetas > REL (sintaxe)
  - etiquetas > SG (semânticas)
- Vários conjuntos de anotação - tagsets- para árvores de dependência.

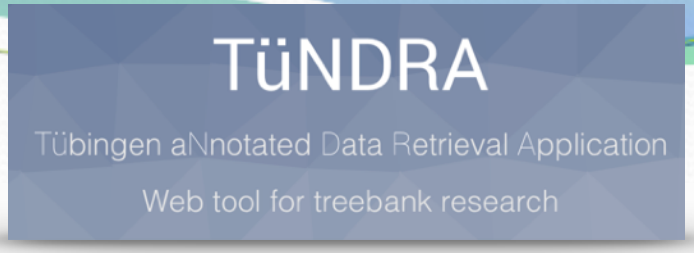
Treebanks do grego antigo

- **Ancient Greek and Latin Dependency Treebank (AGLDT)**,
- The Index Thomisticus Treebank (IT-TB),
- PROIEL Treebank (universal dependency)
- SEMATIA (segue AGDT no seu corpus)
- UD Perseus (universal dependency)

AGDT Anotação em árvore sintática de dependência

- **Ancient Greek and Latin Dependency Treebank (AGLDT: AGDT e ALDT)**
  - https://perseusdl.github.io/treebank_data/
- Ferramenta Arethusa na plataforma Perseids (Tufts U.) http://perseids.org (login c/ google etc)
- Manual: guidelines em inglês, original:
  - AGDT 1. https://github.com/PerseusDL/treebank_data/blob/master/v1/greek/docs/guidelines.pdf
  - AGDT 2. https://github.com/PerseusDL/treebank_data/blob/master/AGDT2/guidelines/Greek_guidelines.md

Consulta no Tündra

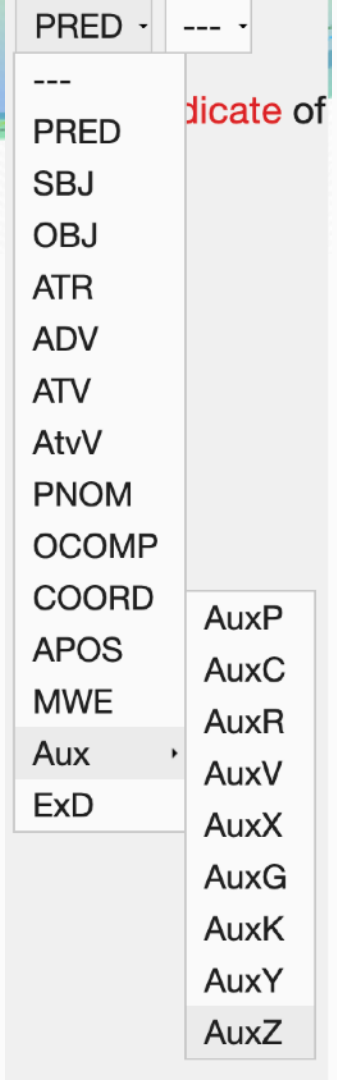


- **Tundra Perseus AGDT**
  - https://perseusdl.github.io/treebank_data/
  - https://weblicht.sfs.uni-tuebingen.de/Tundra/PerseusGreek
- Linguagem usada para consulta no Tundra https://drive.google.com/file/d/1xmUwcixj-nynI98nrPe0_0ysZU-zHsyJ/view?usp=sharing

Tagset AGDT

- **PRED - verbo finito, pessoal, da oração principal (*suj. SEMPRE nom.*)**
- **ATR** - adjetivos, adj poss. adj. dem., artigos na função de artigos e **nó/v. de or. adjetiva**
- SBJ - sujeito (de qq or.)
- **OBJ - compl. v. obrigatório, v. or. completivas**
- **ADV - compl. adv. opc. / v. or. adverbiais**
- PNOM - predicativo do suj; v. or. predicativa
- OCOMP - compl./pred. do OBJ
- COORD - conj. coordenativas

- **COORD - conj. coordenativas dentro da sent. e iniciais no nível do texto**
- AuxC - conj. subordinativas
- AuxP - preposições (adp)
- ExD - vocativo (depend. externa)
- APOS - aposto
- MWE - multiple word expression
- ATV e AtvV - atributos verbais = predic. verbo-nominal
- **AuxZ** - partículas de neg. e enfáticas de palavras específicas
- **AuxY**
  - partículas adv. oracionais
  - conectivos repetidos anteriores ao COORD



Regras básicas de configuração da árvore

- **PRED fica na raiz**, exceto quando há um coord inicial que fica na raiz e o PRED_CO depende do COORD.
- Elementos coordenados com COORD recebem o **sufixo _CO**
- **Incluir aT (artificial token) para elipse verbal do PRED**
- **SBJ e OBJ** dependem do seu verbo/PRED
- **Mais de um PRED, só por coordenação** (outros verbos finitos e pessoais serão dependentes ou como OBJ, ADV c/ conj. ou partic. ou infinit.
- Um nó pode ter vários ramos ∧∧ // mas **cada ramo só pode ter um nó**.
- Preposições AuxP e Conjunções AuxC **são sempre nó de um ramo**, nunca folha solta; estão acima dos termos que dependem delas.
- Os **nós verbais** de oração **ATR dependem do termo nominal que** é expandido pela oração ATR
- O editor pode dividir partículas com dupla função, por exemplo, **οὔ-τε**

AGDT 2. Smyth Grammar - SG

- Tagset para aspectos semânticos de nomes: substantivos, pronomes, adjetivos, verbos e advérbios
- Segue a classificação gramatical da Smyth Grammar
- Especificidade do uso do caso (qual dativo?)
- Especificidade do verbo/oração: independente, dependente

Passagem Platão, Parmênides, 137c-142a

- Probl. específicos
  - sentenças em discurso indireto só c/ infinitivo, sem um verbo finito.
  - sequências dialogadas s/ v. princ.
- Comum
  - constr. infinitivo sujeito
  - aT p/ elipse verbal
  - interrogativas s/ verbo
  - discurso indireto

Referências

- Celano, G. 2019.The Dependency Treebanks for Ancient Greek and Latin. Leipzig, U. https://doi.org/10.1515/9783110599572-016
- Mambrini, F. 2011. L'Ancient Greek Dependency Treebank'. Un nuovo strumento per lo studio della lingua greca, ***Lexis,*** 29.
- Mambrini, F. 2016. The Ancient Greek Dependency Treebank: Linguistic Annotation in a Teaching Environment. In: Bodard, G & Romanello, M (eds.) ***Digital Classics Outside the Echo-Chamber: Teaching, Knowledge Exchange & Public Engagement***, Pp. 83–99. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bat.f.
- Tesnière, L. 2015[1966] ***Elements of Structural Syntax***. Amsterdam: John Benjamins
- Tesnière, L. 1959. ***Elements de syntaxe structurale.***Paris: Klincksieck