Pontos referidos sobre análise sintática
formação BILLIG, maio de 2020

- cada analisador sintático representa uma teoria sobre a língua
- é difícil avaliar um analisador sintático, porque não há um consenso sobre as categorias que deve ou não distinguir (a questão dos dois tipos de sistemas de Gaizauskas)
- em relação aos formalismos das gramáticas computacionais para sintaxe, existem dois grandes ramos: gramática de constituintes ou sintagmática, e gramática de dependências ou dependencial. Ambas têm vantagens e inconvenientes, ambas têm áreas que tratam melhor. Tradicionalmente, as gramáticas de dependências estão associadas a línguas independentes da ordem enquanto as gramáticas de constituintes estão associadas a línguas de ordem fixa.
- em relação à forma como se chega a uma dado resultado, pode primeiro marcar-se muita coisa e depois tirar os restos, ou pelo contrário ir adicionando casos para aumentar a abrangência

- em relação ao português, não faz sentido começar por POS-tagging (que é um truque de engenharia para o inglês), porque o português é muito diferente do inglês (em particular em relação à ordem muito mais livre e à riqueza morfológica muito maior, etc.)

- a questão da avaliação de algo que tenha a ver com análise sintática
- depende das unidades (o que se considera unidade) e dos níveis de classificação
- é preciso entrar em conta que algumas unidades são fáceis (*de* é sempre uma preposição) mas outras coisas é que são difíceis (mas qual a relação semântica que *de* significa? Posse, origem, etc.)
- vagueza e "teto": comparação de um sistema automático com aquilo com que os humanos concordam…

A questão dos métodos estatísticos
Não são em oposição aos métodos linguísticos!!!
São uma maneira prática de suprir aquilo que não se sabe.

A questão das florestas sintáticas
(as duas mais conhecidas são o Penn treebank (Wall Street Journal) e o SUSANNE do Geoffrey Sampson)
Há duas razões para construir FS: descrição linguística quantitativa; e avaliação: treinar e avaliar analisadores sintáticos sobre ela.
Treinar significa que é o sistema que aprende com base nas soluções (a chamada aprendizagem automática (PT), ou aprendizado de máquina (BR)).

Um problema das florestas sintáticas é que são imutáveis, e são pequenas :-) E ainda por cima podem conter erros humanos… mas são certamente aqueles corpos mais valiosos e mais caros e mais respeitados. "Floresta virgem": um erro chamar assim.