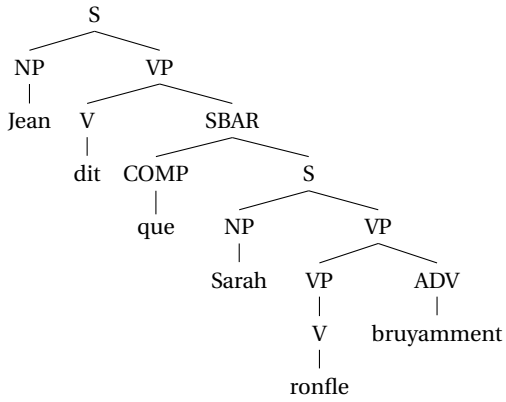
Traitement du langage
Approches linguistiques et empiriques

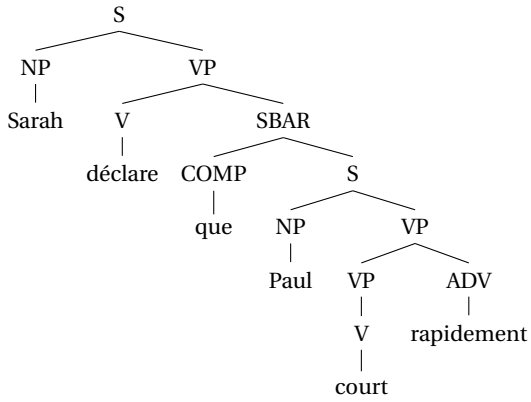
A.A. 2020-2021

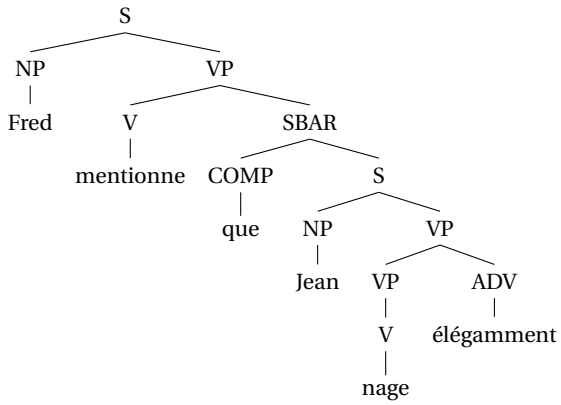
November 5, 2020

1 Extraction de grammaires

Soit un treebank constitué des trois arbres syntaxiques suivants.

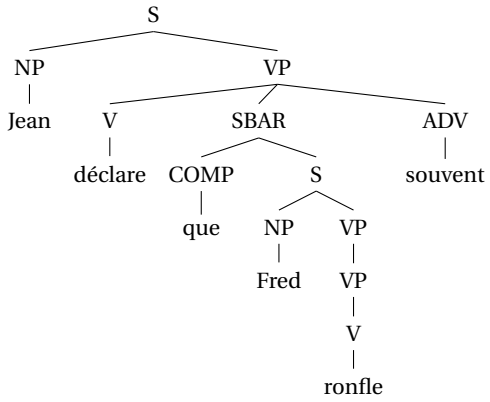






1. Décrivez une grammaire probabiliste de ce corpus, c'est-à-dire notez les règles de grammaire et calculez leurs probabilité.

2. Générez tous les arbres syntaxiques possibles pour la phrase *Jean déclare que Fred ronfle souvent, souvent* est un adverbe (ADV), et calculez leurs probabilités selon la grammaire.



Une des analyses possible pour la phrase *Jean déclare que Fred ronfle souvent* attache l'adverbe *souvent* très haut, au niveau du verbe *déclare*, comme dans l'arbre ci-haut, qui décrit la situation où c'est Jean qui déclare souvent quelque chose.

3 Ce type d'attachement n'a jamais été vu dans le corpus. Afin d'éviter ce genre d'attachements, modifiez les étiquettes des non-terminaux dans le corpus. Votre solution devrait introduire de nouveaux symboles non-terminaux qui permettent à la grammaire de capturer la distinction entre les attachements hauts et bas. La grammaire résultante devrait donner une probabilité de 0 aux arbres avec des attachements hauts.

2 PCFGs

Les PCFG constituent le modèle le plus simple de parsing statistique, mais leur performance est généralement considérée comme insuffisante. Expliquez les raisons de cette inadéquation.