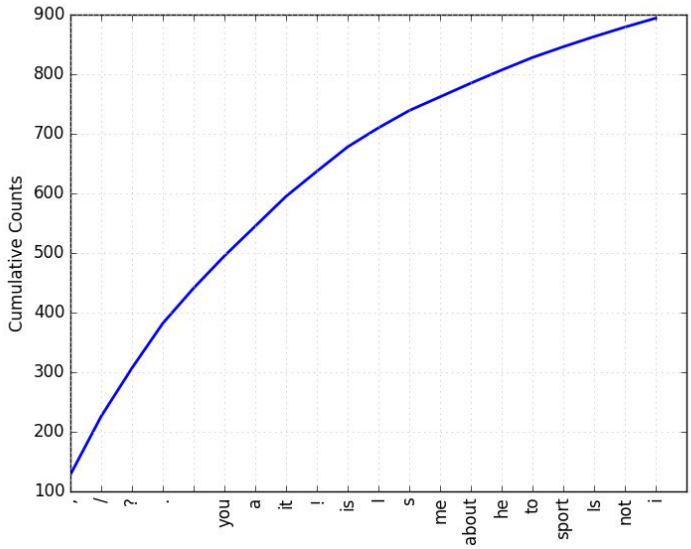
Analyse du corpus (projet TAL)

Nous avons pu récolter un corpus d'une centaine de lignes, où chaque ligne représente une interaction d'un utilisateur avec un bot et la réponse donné par le bot, donc environ 200 phrases. Nous avons demandé à des personnes de notre entourage de nous donner ces exemples d'interactions attendues. Nous leur avons expliqué le contexte de notre jeu, le tout sans leur montrer d'exemple concret pour éviter de les influencer.
Nous avons ensuite analysé ce corpus grâce à NLTK. Voici quelques statistiques:

- Le corpus contient 1865 mots (ici les ponctuations sont incluses)
- Il y a 438 éléments linguistiques (environ 430 sans les ponctuations)
- Les 20 plus fréquents sont : [(',', 128), ('/', 98), ('?', 81), ('.', 75), (' ‘ ', 59), ('you', 54), ('a', 50), ('it', 50), ('!', 42), ('is', 41), ('I', 32), ('s', 29), ('me', 23), ('about', 23), ('he', 22), ('to', 21), ('sport', 18), ('Is', 17), ('not', 16), ('i', 15)]
- Ces 20 éléments constituent environ 48% du corpus



- les mots les plus long sont : ['Cheerleading', 'basketball', 'individual', 'easygoing', 'portuguese', 'gymnastics', 'congratulations', 'tournament']

Nous avons aussi tagger notre corpus et obtenu le résultat suivant:

- types les plus fréquents: [('PRP', 221), ('NN', 216), ('.', 198), ('NNP', 135), (',', 128), ('JJ', 126), ('VBZ', 101), ('DT', 96), ('RB', 96), ('IN', 94)]
  PRP = personal pronoun
  NN = Noun, singular or mass
  . = ponctuation
  NNP = proper noun, singular