Université de Brown. Années 60.

Deux linguistes, Winthrop Nelson Francis et Henry Kučera, influencés par les travaux d'un autre linguiste, Randolf Quirk, se lancent dans la constitution du premier corpus informatisé de l'anglais-américain moderne.

A l'aide d'une petite équipe, ils sélectionnent 500 extraits, de plus de 2000 mots chacun, qui leur paraissent représentatifs d'un anglais américain écrit standard. Ils choisissent parmi des textes variés (éditorial, fictions, écrits scientifiques…) et publiés aux Etats-Unis au cours de l'année 1961.

Ensuite, ils saisissent ce corpus d'environ 1 million de mots sur 100 000 cartes perforées.

En 1964, la première version du *Brown University Standard Corpus of Present-Day American English* (*Brown Corpus* de son petit nom), est publiée. Ses auteurs le rendent immédiatement disponible pour tout chercheur qui en fait la demande.

Sa disponibilité, ainsi que le soin apporté à sa conception, à sa méthodologie d'échantillonnage et sa place de premier corpus informatisé de langue, en font un modèle pour d'autres corpus : le *London-Lund Corpus of Spoken English* et le *Lancaster-Oslo-Bergen*, pour n'en citer que deux.

Le *Brown Corpus* donnera même son nom à une série de corpus basés sur son modèle, surnommée *Brown family*.

Avec les années, d'autres versions du *Brown Corpus* sont publiées. Par exemple, sa version étiquetée est utilisée par Kenneth Ward Church dans le cadre de son programme PARTS (*Program for part of speech tagging*). Cette version étiquetée servira aussi à développer le *Penn Treebank*, corpus de référence pour l'analyse syntaxique.

Aujourd'hui encore, le *Brown corpus* est utilisé dans beaucoup de domaines scientifiques : du traitement automatique du langage naturel à la biomédecine. Il est aussi un des corpus mis à disposition par le *Natural Language Toolkit*, une bibliothèque logicielle open source permettant la création de programmes en Python pour l'analyse de texte.