Tato bakalářská práce je zaměřena na základní předzpracování (tokenizaci a segmentaci) českého textu, zejména pro potřeby vytvoření českého internetového korpusu. Texty pro tento korpus budou automaticky získávány z Internetu, a proto samotné segmentaci předchází automatické určení kódování, čištění a rozpoznání jazyka dokumentu. Provádíme experimenty se dvěmi metodami rozpoznání jazyka a předkládáme jejich výsledky. První z metod je založena na porovnávání nejčetnějších n-gramů (podřetězců délky n) získaných z neznámého dokumentu a rozsáhlého českého korpusu. Druhá metoda využívá odhadu podmíněné pravděpodobnosti výskytu znakových trigramů získaných ze stejného korpusu. Pro širší použití je vytvořen modul pro tokenizaci a určování konců vět. Hledání konců vět je řešeno použitím seznamů českých zkratek a analýzou nejbližšího kontextu míst, která by mohla být za konce vět považována. Rozhodovací strom byl trénován na ručně označených datech. Vyhodnocení úspěšnosti bylo založeno na úsudcích nezávislé osoby a výsledky jsou předloženy v práci.