V této práci jsou zkoumány rozdíly mezi metodami tokenizace pro vícejazyčné neuronové modely (multilingual language models) a rovněž jejich vliv na kvalitu jazykového modelu. Je definována sada metrik, které jsou použity pro vyhodnocení kvality tokenizace: pomocí experimentů je demonstrováno, že užité metriky zachycují rozdíly mezi tokenizátory a korelují s výkonem vícejazyčných neuronových modelů.

Některé práce věnované vícejazyčné tokenizaci uvádí, že standardní postup trénování tokenizátorů na vícejazyčném korpusu není vhodný pro vícejazyčné modely. Tato práce hledá důvod uvedených problémů. Jako možné příčiny jsou zkoumány velikost dat, implementace nebo velikost abecedy. V práci docházíme k závěru, že problém je pravděpodobně způsoben nevyvážeností dat mezi jazyky a navrhujeme řešení v podobě rovnoměrného vzorkování trénovacích dat tokenizátoru.

V diplomové práci jsou replikovány tři studie, které se zabývají vylepšením metod vícejazyčné tokenizace a jsou porovnány se standardním trénováním na rovnoměrných datech. Díky porovnání je zjištěno, že princip, který stojí za zlepšením u replikovaných metod, je stejný jako u rovnoměrného vzorkování.

Výsledky diplomové práce poskytují hlubší vhled do problematiky tokenizace pro vícejazyčné modely. Je navržena metodika a doporučení pro trénování vícejazyčných tokenizérů. Nakonec je ukázáno, jak dosáhnout zlepšení tokenizace bez nutnosti použití složitějších tokenizačních metod.