Szkennelt szövegek digitalizálása során keletkező hibák elemzése magyar szövegek esetében

Pataki Máté, Tóth Zoltán

Kivonat

A Meta-Contentum GVOP pályázat keretében az MTA SZTAKI Elosztott Rendszerek osztálya azt is vizsgálta, hogy magyar nyelvű szkennelt szövegekben milyen hibák keletkeznek az eredeti dokumentumhoz képest, és ezek a hibák mire vezethetőek vissza, ennek a vizsgálatnak az eredményét ismertetjük előadásunk során.

A kutatás nagy mennyiségű, valós, digitális szövegként is rendelkezésre álló dokumentumon folyt. A doc, rtf és txt formátumú dokumentumok kinyomtatott és mesterséges hibával is ellátott, majd beszkennelt változatai kerültek összehasonlításra az eredeti szövegekkel, annak érdekében, hogy algoritmikusan is ellenőrizni lehessen a korábban kézi javítás során szerzett tapasztalatokat.

A végleges tesztadatbázis 1 gigabájtnyi szövegből állt, ezen futottak le az összehasonlító algoritmusok, melyek a kétszer 5500 tesztdokumentumból az eredetit és a hozzá tartozó szkennelt változatot összehasonlították, és kigyűjtötték a hibákat, statisztikákat, szavakat. A szógyűjtemények további elemzéseket tettek lehetővé, mint például az ismeretlen szavak hányada az adatbázisban, egy szótőhöz tartozó ragozott alakok átlagos száma, ritkán előforduló szavak, szóalakok.

Linkek

MTA SZTAKI Elosztott Rendszerek Osztály: http://dsd.sztaki.hu