Lórum ipse: magyar vakszöveg-generátor

Nagy Viktor[1], Takács Dávid[1]

[1] Prezi
{viktor.nagy,david.takacs}@prezi.com

Lorem ipsum

A kiadványszerkesztési és weboldaltervezési munkamenetben fontos szerepe van egy olyan előnézeti képnek, amikor az oldal már rendelkezik a szedési terv grafikai jellemzőivel, de a valódi tartalom helyett töltelékszöveg jelenik meg. Ilyenkor a grafikai tervező könnyebben koncentrálhat a grafikai elemekre és a szöveg grafikai jellemzőire (betűcsalád, betűméret, sortávolság stb.). Helykitöltőként hagyományosan a *lorem ipsum*-nak nevezett szöveget és annak változatait használják, amely *Cicero: De finibus bonorum et malorum* című művének töredékeiből származik szavak felcserelésével, hozzátoldásával, halandzsaszavak bevezetésével.

A generált halandzsaszöveg fontos tulajdonsága, hogy nem csak 'ránézésre' hasonlít az imitált nyelvre, hanem statisztikai jellemzői is hasonlóak ahhoz, továbbá tartalmazza ugyanazokat a betűkapcsolatokat, amelyek az adott nyelv helyesírásában vagy tipográfiájában speciálisan kezelendők, például ligatúrákat alkothatnak.

Nem tudunk olyan *lorem ipsum*-generátorról, amely a magyar nyelvre adaptálva a fenti tulajdonságoknak megfelelő szöveget generálna, ugyanakkor a feladat számítógépes nyelvészeti eszközök alkalmazását igényli és nem triviális, ezért döntöttünk megvalósításáról.

Az alkalmazás

A demón bemutatni szánt alkalmazás képes arra, hogy a beadott paramétereknek megfelelően nagy mennyiségű szöveget generáljon elfogadható idő alatt. A nem paraméterezett döntéseket pszeudorandom módon hozza meg, azaz az algoritmus determinisztikus, egy már látott szöveg újragenerálható azonos bemeneti paraméterekkel.

A generálás során lehetőség van a szöveg bizonyos tulajdonságainak megszorítására az alapvető elvárásokon túl, mint pl. a hangzógyakoriságok finomhangolása, a szótövek gyakorisági eloszlásának alakjának meghatározása, a preferált mondatszerkezetek korlátozása.

Az alkalmazott NLP-technikák

A halandzsaszövegtől elvárjuk, hogy hangzásában feleljen meg a mai beszélt magyar köznyelv fonológiájának, ugyanakkor – a funkciószavak és néhány, véges számú egyéb szó kivételével – nem létező szótöveket tartalmazzon, de azokat szabályosan toldalékolva. Továbbá, amennyire lehetséges, grammatikus mondatokból álljon.

Ennek megvalósításához tehát szükség van egy szótőgenerátorra, amely figyelembe veszi mai szókincsünk fonológiai jellegzetességeit és lehetséges, de nem létező alakokat állít elő; egy morfológiai generátorra, amely a képzett szótövekhez, azok szándékolt szófaját ismerve ki tud választani egy megfelelő paradigmát, és generálja annak alakjait; és egy mondatgenerátorra, amely grammatikus magyar mondatokat mintaként felhasználva kvázi-grammatikus halandzsamondatokat állít össze.

Halandzsa szótőtár

A halandzsa szótöveket ngram-modellel generáljuk. A modellt az elemzett Magyar Webkorpusz lexikonján építettük, szófajonként elkülönítve. Az ngram-model véletlen karaktersorozatokat generál, amelyekből kézzel válogattuk ki a jól hangzó szótöveket.

Morfológiai generáló

A szóalak-generáló feladata az elvárt alak generálása a szótőből. A saját fejlesztésű, szabályalapú generálónk fiktív vagy ismeretlen tövekre működik. Megállapítja a tő fonológiai tulajdonságait, és azok alapján választ egy megfelelő paradigmát és meghatározza az esetleges tőváltozásokat. Ha további információra van szükség a szóalak generálásához, azt pszeudorandom módon választja meg.

Szintaktikai generáló

A szintaktikai modul a Magyar Webkorpuszból kinyert mondatsablonokon alapul. A mondatokat szófaji és morfológiai annotációk sorozataként tárolja el, megőrizve a funkciószavak alakjait. A generálási folyamat véletlenszerűen kiválaszt egy sablont, és a tartalmas szavak helyére elhelyez egy-egy véletlenszerűen kiválasztott halandzsa szót a megfelelő morfológiai alakban előállítva.