Tordai-féle szótövező

A tövezőnek négy verziója létezik.

1. A LIGHT1 szótövező csak a leggyakoribb 14 főnévi esetet kezeli. Ennek ellenére már ez is jelentékenyen javíthatja a keresési hatékonyságot. Nézzük a tesztként használt Szeged Korpuszt. Ebben a főneveknek csak 26%-a szerepel toldalékolatlan formában, a maradék 74%-ból további 36%-ot fed le a LIGHT1 által kezelt 14 eset, tehát a legegyszerűbb szótövezővel már a főnevek több mint fele a megfelelő szótőre képződik. A hatékonyságot tovább növeli, hogy a többi névszók közül a melléknevek és számnevek morfológiai paradigmája is hasonló.
2. A LIGHT2 már 21 esetet kezel, valamint a LIGHT1 által figyelmen kívül hagyott egykarakteres toldalékok közül az akkuzativusz (tárgyrag *-t*) és szuperesszívusz (*-n*) toldalékokat is levágja. Mindkét szótövező figyelembe veszi a szótőjelölt hosszát és, hogy tartalmaz-e érvényes mássalhangzó–magánhangzó kombinációt.
3. A MEDIUM változat 12 gyakori főnévi esetet kezel, a birtokos és birtokok, valamint a személyek számát is figyelembe véve. Ezenkívül kezeli a leggyakoribb igealakokat (idő, szám, személy), a melléknevek fokozását, valamint a számneveknél a törtszámnév és sorszámnév toldalékait.
4. A HEAVY szótövező pedig mind a 21 esetet és az összes igealakot figyelembe veszi.

A gyakorlati tesztek azt mutatták, hogy a LIGHT2 tövezőnek vannak a legkedvezőbb tulajdonságai. Nézzük tehát részletesen ennek működését.

A Porter-tövezőhöz hasonlóan a magyar nyelvre is definiálni kell a magánhangzók, illetve mássalhangzók halmazát, ami kiegészül a magyar helyesírás jellegzetességei miatt még a kettős és hosszú mássalhangzók listájával. Tehát a magyar magánhangzók

$$
\begin{aligned}
v &= \{a, á, e, é, i, í, o, ó, ö, ő, u, ú, ü, ű\},\\
c &= \{b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, w, x, z\},\\
d &= \{cs, dz, dzs, gy, ly, ny, ty, zs\},\\
dd &= \{bb, cc, ccs, dd, ff, gg, ggy, jj, kk, ll, lly, mm, nn, nny, pp, rr, ss, ssz, tt,\\
&\qquad tty, vv, zz, zzs\},
\end{aligned}
$$

Definiáljuk az R_1 régiót az alábbiak szerint:

- ha a szó magánhangzóval kezdődik, akkor az első (akár kettős) mássalhangzó utáni szórész;
- ha a szó mássalhangzóval kezdődik, akkor az első magánhangzó utáni szórész;
- ha a szóban nincs mással- és magánhangzó, akkor az R_1 üres.

A szabályok az R_1 régióban való előfordulásra vonatkoznak. Ez gyakorlatilag a Porter-féle *m* érték szerepét játssza. Egy adott végződéslista esetén mindig a leghosszabban illeszkedő végződésre vonatkozó szabály fog tüzelni. Az x* jelölés a megadott minta előtti karakterre szab feltételt. A LIGHT2 tövező az alábbi lépéseket hajtja végre.

1. lépés *Eszközhatározó esetrag törlése*:
Keresett végződésminták: al, el. Ha a végződés R_1-beli és hosszú mássalhangzó előzi meg, akkor a végződés és a hosszú mássalhangzók egyike törlendő. Formálisan: $(R_1 \wedge \text{dd*})$ {al, el} $\rightarrow$ d.

2. lépés *Gyakori esetragok törlése*:
(R_1) {ban, ben, ba, be, ra, re, nak, nek, val, vel, tól, től, ról, ről, ból, ből, hoz, hez, höz, nál, nél, ig, at, et, ot, öt, ért, képp, képpen, kor, ul, ül, vá, vé, onként, enként, anként, ként, en, on, an, ön, n, t} $\rightarrow$; ha az így kapott alak á-ra vagy é-re végződik, akkor az rendre a-ra vagy e-re cserélendő.

3. lépés *Speciális esetragok törlése*:
(R_1) án $\rightarrow$ a; (R_1) ánként $\rightarrow$ a; (R_1) én $\rightarrow$ e.

4. lépés *További esetragok törlése*:
(R_1) {astul, estül, stul, stül} $\rightarrow$; (R_1) ástul $\rightarrow$ a; (R_1) éstül $\rightarrow$ e;

5. lépés *Transzlatívusz esetrag törlése*:
$(R_1 \wedge \text{dd*})$ {á, é} $\rightarrow$ d;

6. lépés *Birtokrag törlése*:
(R_1) {oké, öké, aké, eké, ké, éi, é} $\rightarrow$; (R_1) {áké, áéi, áé} $\rightarrow$ a; (R_1) {éké, ééi, éé} $\rightarrow$ e;

7. lépés *Egyes számú birtokragok törlése*:
(R_1) {ünk, unk, nk, juk, jük, uk, ük, em, om, am, m, od, ed, ad, öd, d, ja, je, a, e} $\rightarrow$; (R_1) {ánk, ájuk, ám, ád, á} $\rightarrow$ a; (R_1) {énk, éjük, ém, éd, é} $\rightarrow$ e;

8. lépés *Többes számú birtokragok törlése*:
(R_1) {jaim, jeim, aim, eim, im, jaid, jeid, aid, eid, id, jai, jei, ai, ei, i, jaink, jeink, eink, aink, ink, jaitok, jeitek, aitok, eitek, itek, jeik, jaik, aik, eik, ik} $\rightarrow$; (R_1) {áim, áid, ái, áink, áitok, áik } $\rightarrow$ a; (R_1) {éim, éid, éi, éink, éitek, éik} $\rightarrow$ e;

9. lépés *Többes szám törlése*:
(R_1) {ök, ok, ek, ak, k} → ; (R_1) ák → a; (R_1) ék → e;

PÉLDA. Tekintsük az alábbi példákat!

$$\textit{fiókáinknak} \xrightarrow{\text{2. lépés}} \textit{fiókáink} \xrightarrow{\text{8. lépés}} \textit{fióka}$$

Ha a szótő k-ra vagy t-re végződik, akkor az algoritmus bizonyos esetekben túltövez.

$$\textit{fiókja} \xrightarrow{\text{7. lépés}} \textit{fiók} \xrightarrow{\text{9. lépés}} \textit{fió*}$$
$$\textit{keret} \xrightarrow{\text{2. lépés}} \textit{ker*}, \quad \text{de} \quad \textit{kerete} \xrightarrow{\text{7. lépés}} \textit{keret}$$

Forrás:
Snowball honlapja.

A. Tordai and M. de Rijke. Hungarian monolingual retrieval at CLEF 2005. In *Working Notes for the CLEF 2005 Workshop*, Vienna, Austria, 2005.