Corpusul ZiareRom

Corpusul ZiareRom este un corpus de texte culese din variantele on-line ale unor ziare românești din perioada 2004-2007.
Textele articolelor sunt distribuite ierarhic (pe directoare și subdirectoare) pe ani, luni, zile și ziare.
Ziarele inspectate sunt Adevărulonline, BBC-Romanian, Bursa, Capital-RO, Cotidianul, Crainou, Euractiv-ro, Evenimentul Zilei, Jurnalul, Libertatea, Ziarul de Iași, Ziua, 7Plus.
Corpusul însumează peste 86 de milioane de cuvinte (în 2004: 11 mil., în 2005: 20 mil., în 2006: 17 mil., în 2007: 38 mil.).

NOTĂ: Textele au grafii diferite. Majoritatea textelor conțin echivalentele nediacritice ale noii ortografii, altele conțin echivalentele nediacritice ale vechii ortografii (fără â), de exemplu Cotidianul, iar altele conțin diacritice, de exemplu BBC-Romanian.