

Kísérlet cseh nyelvű kivonatok gépi indexelésére

Az Ústředí vědeckých, technických a ekonomických informací /Tudományos, Műszaki és Gazdasági Tájékoztatási Központ/ a pilzeni Skoda-Művek Tájékoztatási Központjával együttműködve 1973 folyamán 779 cseh nyelvű, gőzturbinákról szóló szakirodalmi kivonat gépi indexelésére tett kísérletet. A maga nemében elsőnek tekinthető kísérlet programnyelve a FORTRAN volt, a kísérletben használt számítógép pedig ICL 1905 típusú.

A szóban forgó kivonatokat először intellektuális munkával indexelték, amiből az is következik, hogy a kísérlet elsődleges célja az annak megállapítását tüzték ki, hogy a gépi indexelés mennyire megbízható az intellektuáliséhoz képest. Ennek érdekében az intellektuális indexelést is szuperrevízióknak vetették alá.

A kísérlet befejezése után 200-200 intellektuálisan és gépi uton indexelt kivonatot vetettek össze. Ennek nyomán a következőket lehetett megállapítani:

1. Egy-egy kivonatra átlagosan 4,25 releváns deszkriptor esett. Az intellektuális indexelés egy-egy kivonatot átlagosan 2,73 deszkriptorral jellemzett, a gépi uton történő 5,07-tel. Ebből 3,45 volt a releváns, 1,62 pedig a hibásan megadott deszkriptor. Az intellektuális indexelésnél tehát 1,52, a gépinél 0,80 volt a releváns deszkriptorok átlagos hiánya.

2. A gépi indexelésnél a releváns deszkriptorok hiányát 51,2%-ban a fordító szótár hibái, 31,3%-ban az egyszerű identifikálási szabályok előírásának lehetetlenségei, 17,5%-ban különféle formális vétségek okozták.

3. A hibás deszkriptorok létrejöttének ugyancsak sokféle oka volt, nevezetesen: az identifikátorok szótári készletének hibás kiválasztása /42,5%/, hibás identifikátor-kombinációk létrejötte a mondatok szintjén /27,8%/, hibás identifikátor-kombinációk létrejötte a szöveg szintjén /10,2%/, a szótár átfedései és a program hibái /18,6%/, a lexikális többértelműség hatása /0,9%/.

A kísérlet, illetve az összevetés során tematikusan "kibuktatott" hibák jelentős részét ki lehet küszöbölni. De az is kitűnt, hogy a gépi indexelésben egyelőre nem lehet 100%-os megbízhatóságra törekedni, mint ahogy ez tulajdonképpen a szellemi munkával történő indexelésben sem igen érhető el.

A kísérlet egyébként rámutatott azokra a kutatási szükségletekre is, amelyek nélkül a ma meglévő problémák nem mindegyikét lehet kiküszöbölni a gépi indexelésből /mondattani, a címek és szövegek struktúrájára vonatkozó tanulmányok/ a kivonatok indexeléséről, az ún. extraktumok /az eredeti szövegből kiemelt mondatok/ indexelésére való áttérés lehetőségeinek vizsgálata, a fordító szótárak elméleti kérdéseinek megoldása, stb. /.

/Československá Informatika, 17.k. 1.sz. 1975.
p.4-8./

Futala Tibor