A mondatokra bontó algoritmus működése

A potenciális mondathatároló jelekre megvizsgáljuk, hogy milyen szabályok illeszkednek a jel adott környezetére, és ennek alapján döntünk. Ha több szabály illeszkedik, akkor a szabályok súlyának aggregálásával határozzuk meg a végső konfidenciaértéket (ez lehet additív vagy multiplikatív).

A szegmentálást támogató nyelvfüggő leírófájlnak a következő elemeket kell tartalmaznia:

- mondathatároló jelek,
- mondathatárolást leíró szabályok,
- rövidítéslista.

Mondathatároló jelek: .!?
Attribútuma:

- típus: kijelentő, felszólító, kérdő.

Mondathatároló szabályok: A szabályok szabályelemekből állnak.
Attribútuma:

- típus: és | vagy | nem | kizáró vagy — Ezek azt írják le, hogy a szabályelemek között milyen logikai kapcsolatnak kell teljesülnie.
- súly (valós) — Annak értéke, hogy az adott szabály teljesülése esetén mennyivel változik a mondathatárjelölt konfidenciaszintje. Additív algoritmus esetén $+/-$ értékek jelzik a szabály orientációját, multiplikatív esetben: gyengítő súly $(0, 1)$; erősítő súly $(1, \text{maxsúly})$.

Szabályelemek: Rövidítés típusa vagy reguláris kifejezés, ld. később.
Attribútumai:

- típus: rövidítés | reguláris kifejezés — a rövidítések listában megadott fix értékek, a reguláris kifejezésekkel mintákat lehet megadni (pl. web, ftp, e-mail címek).
- pozíció: a mondathatároló jelhez képesti pozíció, nem nulla egész szám.

Rövidítéslista: Rövidítéselemekből áll.
Rövidítéselem: Értéke a rövidítés szövege.
Attribútuma:

- típus: egyszerű | név előtt | név után | szám előtt | szám után — a felismerés helyességének javítása érdekében érdemes megkülönböztetni a rövidítések előfordulási típusát (ld. még szabályelemeknél). A típuslista bővíthető.

Ha külön kezeljük pl. a cégneveket és a személyneveket, akkor más szabály vonatkozik a Dr. Tóth Pál-ra és az XYZ Kft.-re. Egy rövidítés többször is szerepelhet a listában különböző típussal.

A fenti szabálykészlettel kell leírni a lehetséges mondathatároló-jelet tartalmazó, de más funkciót betöltő karakterláncokat. Ezeknek egy bő listája:

- webcímek, pl. `szovegbanyaszat.typotex.hu`, akár http(s)-el,
- ftp-címek, pl. `ftp.dante.de`, akár ftp előtaggal,
- e-mail címek, pl. `szovegbanyaszat@typotex.hu`
- IP-címek, pl. 154.66.254.168,
- dátumok:
  - csak számokból álló, pl. YYYY.MM.DD formátum,
  - kiírt dátum, pl. 2006. július 20.,
  - rövidített dátum, pl. 2006. júl. 20.,
  - részleges dátum, pl. 2006. júl. vagy 07.20.,
- tizedesponttal jelzett törtszámok, pl. 3.141,
- római számok, pl. II. világháború, XVI. Benedek, XX. század stb.,
- általános sorszám, pl. 24. helyen végzett, 8. rendelet, 7. osztályos tankönyv,
- egyszerű rövidítés, pl. szerk., ui., ill.,
- tulajdonnevek rövidítés előtaggal, pl. dr. Tóth Vilma (itt ügyelni kell, mivel a következő szó nagybetűs),
- tulajdonnevek rövidítés utótaggal, pl. Nyereség Kft.