

# Számítógépes szövegelemzés

Krauth Péter

**Tézis:** *A számítógépes szövegelemzés, nemcsak hogy integrálódik a vállalatok üzletiintelligencia-megoldásaival és ismeretgazdálkodási kezdeményezéseivel, kiterjeszti az informatika lehetőségeit új alkalmazások felé, és szerves részévé válik az alkalmazási rendszereknek, de általában is növelni fogja az ember-gép kapcsolat hatékonyságát.*

## 1. Témakör

Általános konszenzus van azon megállapítás mögött, hogy az adatbázisok és üzleti alkalmazások rohamos elterjedésének ellenére a keletkező információ túlnyomó többsége (egyebek szerint: 80-90%-a) továbbra is strukturálatlan információ. Fontos feladat ezért, hogy ezekből – további elemzéseket és feldolgozásokat lehetővé tevő – strukturáltabb adatokat lehessen előállítani.

**A számítógépes szövegelemzés<sup>1</sup> szöveges dokumentumokat elemez, hogy ezekből adatokat és metaadatokat nyerjen ki a legkülönbözőbb technikákkal.**

Lehetővé teszi – legalábbis részlegesen – a még strukturálatlan<sup>2</sup> szövegek strukturálttá alakítását („strukturálizálás”), amit aztán vizsgálni, mérni lehet és be lehet építeni további feldolgozásokba. A szövegelemzés, más néven szövegbányászat, több informatikai területtel is szoros kapcsolatban áll, mint pl. információkinyerés, adatbányászat, gépi tanulás, statisztika és számítógépes nyelvészet. Gyakori, hogy a különböző (pl. nyelvi, statisztikai) technikákat kombinálva együtt alkalmazzák.

A számítógépes szövegelemzés – tágabb értelemben – a még strukturálatlan szövegekben minták, kategóriák megtalálásán túlmenően, ezek trendjeinek a meglévő, strukturált adatokkal való összevetési feladatára, valamint az ilyen módon kapott adatok hagyományosabb módszerekkel való további elemzésére is kiterjed.

A strukturálizáló (kivonatoló) megoldásokat a szűkebb értelemben vett szövegelemzéstől függetlenül, nemcsak szövegekre, hanem egyre inkább más kevésbé strukturált adattípusra is kiterjesztik, és néhány kísérleti alkalmazásban felhasználásra kerülnek (pl. hanglenyomatok azonosítása felvételeken, arcfelismerés videókon, képeken). **Azonban ez a tanulmány csak a szövegek elemzésével foglalkozik.**

Említést kell itt tenni a *beszédfelismerés*ről is. Ez a technológia tulajdonképpen szöveget állít elő élő beszéd alapján, azaz a szövegelemzéshez „szállítja” a bemenő adatokat (magát a szöveget). A technológia jelenlegi fejlettségét jól mutatja, hogy kb. 80%-os szófelismerési pontosságot lehet már elérni magyar nyelven is. Ez még nem teszi lehetővé, hogy szintaktikai elemzéseket is lehessen végezni a beszédből meggenerált szövegen, de statisztikai módszerekkel már érdemben vizsgálható, és keresési célra is már jól használható. A

---

<sup>1</sup> Ez a tanulmány egyfajta kiegészítés is az NHIT Információs Társadalom Technológiai Távlatai (IT3) projekt keretében 2007-ben közzétett „Üzleti intelligencia” c. elemzéshez.

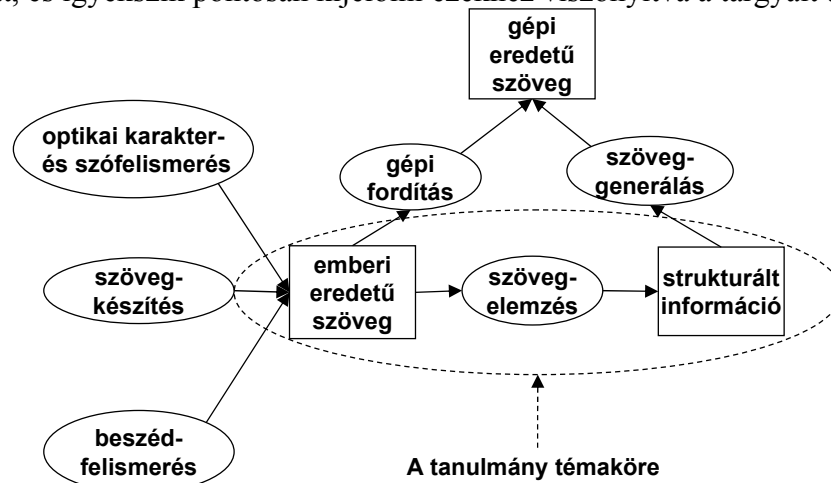
<sup>2</sup> Tulajdonképpen elég furcsa *strukturálatlannak* nevezni egy szöveget, amikor – ha nem is könnyen kezelhető módon, de – rendkívül mély nyelvi strukturákat tartalmaz, és nem szabad elfelejtkezni az egyéb pl. statisztikai, gyakorlati összefüggésekről és a nem nyelvi jellegű, formai szerkezetekről sem: táblázat, felsorolás, csoportosítás. Sőt, éppen ezek komplexitása okozza a problémát. Helyesebb lenne ezért a „strukturálatlan” helyett a „nehezen feldolgozható strukturákat tartalmazó” kifejezés. Ezt nehézsége miatt nyilvánvalóan nem célszerű használni, de a továbbiakban a „strukturálatlan” kifejezés erre fog utalni. Pontosabban: a szövegek strukturálatlanságán azt értjük, hogy sem dokumentumszinten, sem mondat szinten nincsenek a strukturális egységek és jelentéseik közvetlenül – és gépileg könnyen kezelhető – módon megadva. Dokumentumszinten pl. általában nem állnak rendelkezésre a szöveg nyelvére, témájára, stílusára stb. jellemző metaadatok, és gyakran nincs explicite megadva a szöveg belső szerkezete sem: címstruktúra, bekezdés- és mondatatórok. Mondatszinten csak erre a célra annotált dokumentumgyűjtemények esetén áll rendelkezésre a mondat belső (szintaktikai) strukturája.

tanulmány fontosnak tartja a beszédfelismerés technológiáját, de az előzőekben mondottakkal összhangban nem tekinti a témakörébe tartozónak.

További korlátozás, hogy nem terjed ki a tanulmány a *helyesírást*, ill. a *nyelvhelyességet ellenőrző és javító*, valamint a *szöveg készítését segítő* egyéb (pl. szakterület-specifikus szótárra és kifejezógyűjteményre épülő) eszközökre. Ez azzal indokolható, hogy egy adott célnak és kontextusnak megfelelő és adott tárgyról szóló szöveg *létrehozása* (és ennek támogatása), ill. egy létező szövegnek az *elemzése* (tkp. annak meghatározása, hogy miről szól) egymástól jól elválasztható feladatok. Míg az első feladatnak elsősorban a jó, helyes (ember által jól érthető) szöveg készítése a célja (érthetetlen szöveget viszonylag könnyű előállítani), a második esetben viszont akár helyesírási hibák, nyelvhelyességi problémák esetén is meg kell kísérelni az elemzést. Emellett a tanulmány terjedelmi kereteit is túlságosan kitérítaná mindkét feladat egyenrangú kezelése.

Külön esetet jelent a *gépi fordítás* területe, ami hasonló okok miatt maradt ki. Ez a szövegelemzés és -generálás együttes alkalmazását igényli (egy adott nyelvű szöveg elemzése, majd ez alapján új, más nyelvű szöveg létrehozása), természetesen nem mereven szétválasztva, hanem a szöveg különböző részeire és különböző szintjeire iteratív módon hajtva végre ezeket.

Az 1. ábra összefoglalja az említett, legtipikusabb számítógépes szövegkezelési technológiákat, és igyekszik pontosan kijelölni ezekhez viszonyítva a tárgyalt témakört.



1. ábra: szövegkezelési technológiák és a tárgyalt témakör

## 2. Jelenlegi helyzet

A múltban a strukturált és a strukturálatlan információk világa – leszámítva a könyvtári rendszereket – alig érintkezett egymással. Manapság azonban már a két terület technológiai között számos együttműködés valósul meg.

A *szövegelemzés* kifejezés ma egy sor nyelvi és lexikai elemző, mintafelismerő, információkinyerő, címkéző-strukturáló, kapcsolatelelemző és megjelenítő technikát ölel fel. A kifejezés azokra a tevékenységekre is kiterjed, amelyek ezeket a technikákat üzleti problémák megoldására alkalmazzák – akár jól-strukturált adatok elemzésével együtt, akár attól függetlenül. E technikák és tevékenységek azzal a képességgel rendelkeznek, hogy olyan ismereteket (pl. tényeket, üzleti szabályokat és kapcsolatokat) tárnak fel és jelenítenek meg, amelyek korábban gépi feldolgozás számára kezelhetetlen módon a szöveges dokumentumokba be voltak „zárva”. Lehetővé teszik ilyen módon, hogy a számítógépek az emberi kommunikációt annak eredeti formájában értelmezzék, és használják fel a további feldolgozás során.

## 2.1 Történeti áttekintés

Történetileg visszatekintve e fejlődés során három fázist lehet egymástól jól megkülönböztetni:

1. **1990-ig:** Az internet tömeges elterjedése előtt létrehozott, elsőgenerációs eszközök **kulcsszavakat használó teljes szövegkeresésén** alapultak. Az 1980-as évek vége felé létrehozott adatbázisok és dokumentumgyűjtemények ilyen funkciókat nyújtottak.
2. **1990-2000 között:** A technológia fejlődése később lehetővé tette e hagyományos technikák finomítását új funkciók irányában, azaz olyan keresőmotorok terjedtek el, amelyek a **kulcsszavakat már kombinálni képes statisztikai algoritmusokon** alapultak, és ezzel jóval célzottabb hozzáférést tettek lehetővé az információkhoz.
3. **2000-től:** A harmadik generációs technológia túllép ezeken a megközelítéseken. Az elmúlt évtizedben a termékek és eszközök olyan új osztályai jelentek meg, amelyek **nyelvi alapú** keresési és elemzési megoldásokat rutinszerűen alkalmaznak. Különösen fontosak ebben a kategóriában azok az eszközök, amelyek **mély nyelvi elemzésen** alapulnak (szemantikus motorok). A **szemantikus technológia** ugyanis **logikai és mondatközi elemzési** képességgel, valamint a többértelmű kifejezések kontextushoz illeszkedő **egyértelműsítésével** egészíti ki az alapszintű (nyelvtani és morfológiai) nyelvi elemzést.

A **nyelvi – különösen a szemantikai alapú – technológiák** képesek átlépni a hagyományos kereső eszközök korlátain úgymint:

- **pontatlanság**, azaz hogy képtelenek csak azokat a dokumentumokat meghatározni, amelyek lényegesek a keresési feltételek szempontjából (például egyetlen szóhoz kapcsolódó egymástól igen távol eső jelentések meg nem különböztetése);
- **alacsony találati arány**, vagyis hogy képtelenek minden lényeges dokumentumot megtalálni egy gyűjteményben; hagyományos megközelítés esetén lehetetlen például egyetlen kéréssel minden olyan dokumentumot meghatározni, amely foglalkozik egy adott fogalommal – kivéve, ha erre külön fel nem készítik, azaz a fogalom összes kifejeződését kulcsszóként egyenként bele nem illesztik.

A nehézség elsősorban abban áll, hogy ha a keresett információ benne is van egy szövegben, az legtöbbször csak „rejtve” van jelen. Ez igen bonyolulttá teszi „megtalálását”, mivel a hagyományos eszközök nem - vagy csak igen alacsony szinten - ismerik fel egy-egy szó vagy szókapcsolat jelentőségét a szövegben egy kérdés szempontjából, és a találatokat sem tudják kellően strukturált formában bemutatni. A nyelvi elemzésre építő eszközöket éppen az ilyen anomáliák csökkentésére találták ki. Ezek ugyanis lehetővé teszik a szövegek további „bányászatát”, mivel olyan gazdag annotációval látják el a szöveget, amely elég közel áll a szöveg létrehozójának eredeti – a szöveg létrehozását kísérő – mentális folyamataihoz. A szemantikára építő eszközök tovább növelik a szövegbeli információk kezelését automatikus kategorizálással (taxonómiák, tartalomosztályozási kritériumok és ontológiák használatán keresztül), a kontextus meghatározásával és a feltételezett kontextushoz tartozó háttérinformációkra is kiterjedő elemzéssel.

Az ilyen összetett, szemantikus-nyelvi technológiára alapuló eszközöktől várható csak, hogy a szövegértelmezést ne alapvetően a szavak általános gyakoriságai vagy véletlen előfordulásai határozza meg, hanem az adott kontextushoz illeszkedő specifikus jelentéseik. Mindez azt vetíti előre, hogy számítógépes eszközökkel is egységes és magasszintű értelmezéséhez lehet majd eljutni egy-egy szöveg tartalmának, és ezzel egyszer valamikor a – talán nem is olyan nagyon távoli – jövőben meg lehet majd közelíteni az emberek által már évezredek (egy-egy becslések szerint 100-200 ezer év) óta rutinszerűen használt beszéd- és szövegértési folyamatokat is.

## 2.2 Kapcsolódó területek

A szövegelemzés ilyen módon kapcsolódik a *tartalom- és ismeretkezeléshez*, amelynek keretében – többek között – a szöveges vagy más „strukturálatlan” adatforrások rendszerezésére törekszik, valamint módszereket és eszközöket ad a szervezeti tudás kinyerésére, tárolására, visszakeresésére és megosztására. A szövegelemzés kiterjeszti, de gyakran fel is használja a hagyományosabb szemantikai elemző és információkategorizálási módszereket, úgymint: üzleti területek terminológiáját rendszerező *tezauruszok*, szakterület-specifikus, hierarchikus, információosztályozást adó *taxonómiák* és objektumosztályok közti kapcsolatokat felállító *ontológiák*.

A szövegelemzés talán még ezeknél is szorosabban kapcsolódik a *kereséshez*. A keresési technológiák ma a szövegállományok rendszeres indexelésén és a keresett információk indexen keresztüli elérésén alapulnak. A keresett szavak szövegben való szereplésének általános logikai kapcsolatain („AND”, „OR” stb.) túl nem végeznek kapcsolatelemzést. A szövegelemzési technológia túlmegy a keresési technológián azzal, hogy statisztikai, adatbányászati és gépi tanulási megközelítések felhasználásával klasztereket, kategóriákat, összefüggéseket határoz meg, amelyek egyaránt vonatkozhatnak a forrásdokumentumokra és a bennük lévő objektumokra és fogalmakra.

## 2.3 Technológia

Abban az esetben, ha a szöveges információ jól be van ágyazva valamilyen strukturált környezetbe (pl. egy relációs adatbázis valamilyen attribútumaként jelenik meg), akkor a kontextus és az információ jelentése viszonylag egyértelmű. A szövegelemzésnek ekkor is van jól meghatározható hozzáadott értéke, de feladat lényegesen leegyszerűsödik.

Egy önálló, strukturálatlan szöveg természetesen egészen más eset. A szöveg formátuma, témája, nyelve stb. tetszőleges lehet. További nehézséget jelenthet a gyakran nem tökéletes helyesírás, a helytelen szóhasználat, a kontextus többértelműsége, stb.

A szövegelemzési technikák strukturált, számszerű adatok helyett dokumentumokkal és ezek tartalmával foglalkoznak: szavakkal, kifejezésekkel, valamint egyszerűbb és összetettebb objektumokkal, mint nevek, dátumok, események és egyéb fogalmak.

A legfontosabb cél, hogy lényeges mintákat és kapcsolatokat lehessen feltárni a forrásszövegben, strukturálni ezt az információt a további felhasználás érdekében, és az elemzés eredményeit felhasználva egyaránt lehetővé tenni a forrásszöveg interaktív vizsgálatát és automatikus kezelését.

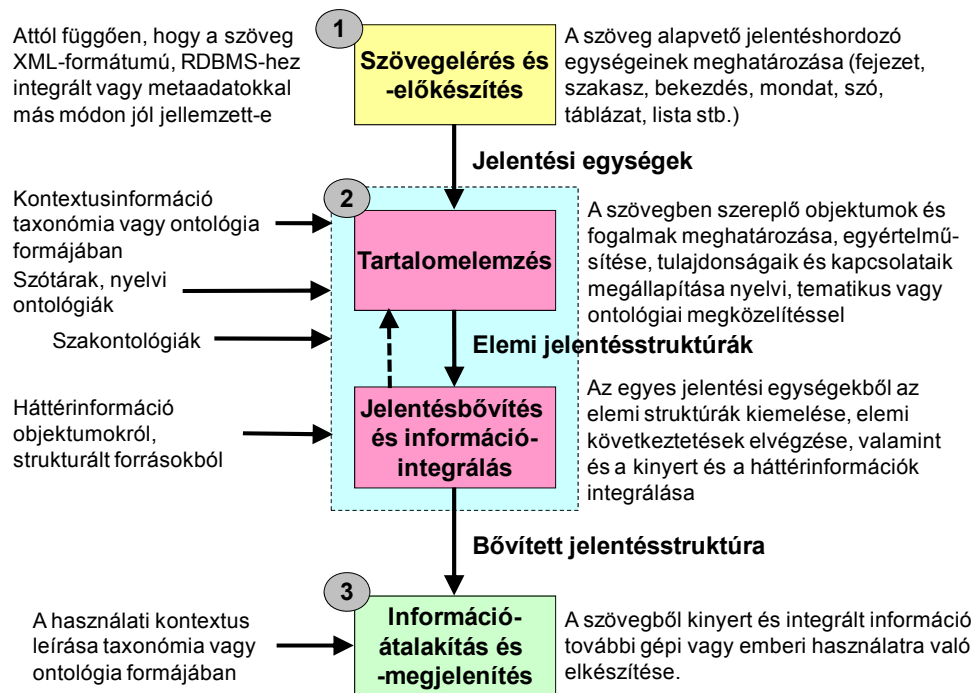
A szövegelemzés általában a **2. ábra** által mutatott három fő szakaszban kerül végrehajtásra.

### **1. szakasz: elérés és előkészítés**

Ez a szöveghez való konkrét hozzáférést jelenti fájlrendszerekben, tartalomkezelő, adatbázis- és e-mail-szervereken, vagy hálózati forrásokban, webhelyeken, webes adatbázisokban, sőt akár adatfolyamokban (data stream). Ennek keretében történik a szöveg jelentési egységeinek meghatározása (pl. szegmentáció, tokenizáció), és a szöveg esetleges formai átalakítása, vagy tartalmi módosítása is. A módosítás során megoldandó legfontosabb feladatok:

- Az adott kontextus(ok)ban lényegtelen „töltelék”-szövegek elkülönítése, és figyelmen kívül hagyása a további elemzések során.
- Az alapszavak (szótövek) meghatározása, azaz szótövezés, vagy komplexebb feldolgozás esetén teljes morfológiai elemzés. Ugyancsak idetartozik a tulajdonnevek és egyéb azonosító jelek kezdeti felismerése.
- Abban az esetben, ha nem történik később nyelvtani elemzés a szövegen, akkor célszerű a jelentést nem hordozó szavakat (kötőszavak, utaló szavak, névmások, névelők stb.) eltávolítani a szövegből.

További korrekciók szintén lehetségesek (pl. helyesírás-javítás), azonban mindez csak az előkészületekhez tartozik, és semmilyen jelentéssel még nem ruházza fel a szöveget.



2. ábra: A szövegelemzés fő szakaszai

## 2. szakasz: elemzés és integráció

Különböző speciális technikák, pl. lexikai és nyelvi elemzés – címkézéssel és információkinyeréssel együtt – alkalmazása a dokumentum tartalmának strukturálása céljából oly módon, hogy biztosítsa a már bevált adatbányászati technikákkal való feldolgozást. Önálló szövegek esetén az egyik legnagyobb probléma az általában igen magas dimenziószám, azaz, hogy sok különböző jellemzője van a szövegnek, amelyek különböző kontextusokat jelenthetnek, és amelyek szerint a szövegek különböző értelmezései állíthatók elő.

A tartalmi elemzésnél a következő három fő megközelítés van jelenleg használatban:

- *A nyelvi megközelítés* a strukturálatlan szöveget általában mondatokra bontja, és ezeket az adott természetes nyelv szabályai szerint mondatelemzésnek veti alá. A nyelvi megközelítést szerteágazó akadémiai kutatások támasztják alá, de így is rendkívül összetett, és – egyelőre – eléggé lassú. A számos természetes nyelv, és azok számtalan nyelvi finomsága miatt egy szöveg tisztán nyelvi alapokon történő elemzése nagyon nehéz. A különböző nyelvi megközelítések a használt számítógépes szótár és a mondatelemzéshez használt komplex algoritmusok jellege és tulajdonságai szerint térnek el egymástól.
- *A tematikus megközelítés* arra a feltételezésre épül, hogy egy dokumentum témájához (azaz, hogy miről is szól a szöveg) azok a szavak állnak a legközelebb, amelyek a leggyakrabban szerepelnek a szövegben. A szótöveket a szereplési gyakoriság szerint rendezik. Általában az így rendezett szótövek legfelső 10%-át tekintik a dokumentum központi témáinak. A szótövek sorrendje és egyes szótövek egymáshoz való közelsége alapján gyakran további következtetéseket is levonnak a dokumentum jelentésével kapcsolatban.
- *Az ontológiai megközelítés* az elemzendő strukturálatlan szöveget valamilyen, jól strukturált, lehetőleg „szakmailag elismert” szakontológiával (pl. üzleti ontológia) veti össze. A szakontológia fogalmai „szűrőként” viselkednek, és az ezen fennakadt

„találatok” számára és jellegére alapozva vannak le következtetéseket az eredeti szöveg tartalmára és jelentésére vonatkozóan.

A *statisztikai megközelítés* azért nem lett külön említve, mert tulajdonképpen nem is alkot önálló megközelítési módot annak ellenére, hogy statisztikai technikákat ma már mindegyik fenti megközelítésben felhasználnak (pl. szövegtörzsek<sup>3</sup> statisztikai elemzése, kollokációkeresés, gyakoriság-, klaszterelemzés, korrelációs számítás).

Könnyen úgy tűnhet, hogy pontos és teljes nyelvi elemzés nélkül nem lehet használható információt kinyerni egy szövegből. Ha azonban ez így lenne, akkor az emberek sem tudnának megtanulni beszélni. Az emberi elme ilyen átalakulását éppen az teszi lehetővé, hogy a hallott (de nem – vagy nem teljesen – értett) szövegekből többletinformációt (mintákat, kapcsolatokat) vagyunk képesek kinyerni, és ilyen módon egy még nyelv nélküli (vagy primitív nyelvi) állapotból tagolt beszédre képes állapotba tudunk kerülni<sup>4</sup>.

Ezt támasztják alá a felnőttkori – nem célzott és nem felügyelt – nyelvtanulás nehézségei, és az olyan helyzetek kialakulása is, amikor egy idegen nyelvi környezetben felnőttek nem képesek tagolt (nyelvtant használó) beszédre, hanem lényegében csak szavak, szószorozatok kimondására képesek (pidzsin-nyelvek). E jelenségek magyarázatára kézenfekvő feltételezés, hogy elménkben születéstől kezdve működik egy „szemantikus kereső motor”, egy olyan modul, amely egy hang, egy hangsor vagy egy szó hallatán képes felidézni az ehhez kapcsolódott emlékképeket, azaz viszonylagosan stabil „jelentést” tud adni ezeknek anélkül is, hogy a szavak finom – nyelvtani szerkezetekben megjelenő, szintaktikai – kapcsolódásait kezelni tudná. Ez utóbbit valószínűleg az elmében egy másik, teljesen csak az emberek kb. második életévére kialakuló „szintaktikai (nyelvtani) modul” végzi. Ilyen módon a szemantika, mint asszociációs képesség az elsődleges, míg a szintaktika csak erre ráépülve tud pontosabb, egyértelműbb jelentésszerkezeteket kialakítani – mondatok formájában.

Kicsit hasonló a helyzet a számítógépes szövegelemzésnél is. Lehetségesnek kell lennie, hogy teljes nyelvi elemzés<sup>5</sup> nélkül is információt tudjon kinyerni a számítógép: a tematikus és ontológiai megközelítések éppen erre építenek, de sok nem-teljes nyelvi elemző működőképessége is ezen alapul. Olyan technikák is idetartoznak emiatt, amelyek a szavak önálló és együttes előfordulásai gyakoriságának meghatározásával, de a nyelvtani szerkezetek figyelmen kívül hagyása vagy csak részleges felismerése mellett, esetleg más, pl. mondatok közötti kapcsolatok felismerésével tesznek lehetővé következtetések levonását a szöveg tartalmáról valamilyen adott alkalmazási szituációban.

Valószínűsíthető, hogy a legteljesebb információkinyerésre e technikák *együttes alkalmazásával* nyílik lehetőség.

### **2.3.1 Szövegelemzés nyelvtechnológiai alapelemei**

#### **2.3.1.1 WordNet**

---

<sup>3</sup> A korpusz általában olyan számítógépen tárolt írott vagy beszélt nyelvi anyagot jelöl, amelyen nyelvészeti elemzést végeznek. Gyakran olyan élőmunkával („kézzel”) annotált mondatgyűjtemények, amelyek minden bennük szereplő szóra és/vagy mondatra annotáció formájában tartalmazzák a nyelvészek által elvégzett, lexikai és szintaktikus elemzésük eredményét. Érdekes tudni, hogy a mai korpuszok már „elég” nagyok (1-100 millió szó): a gyermekkori nyelvelsajátítás során ugyanis az ember kevesebb mondattal találkozik.

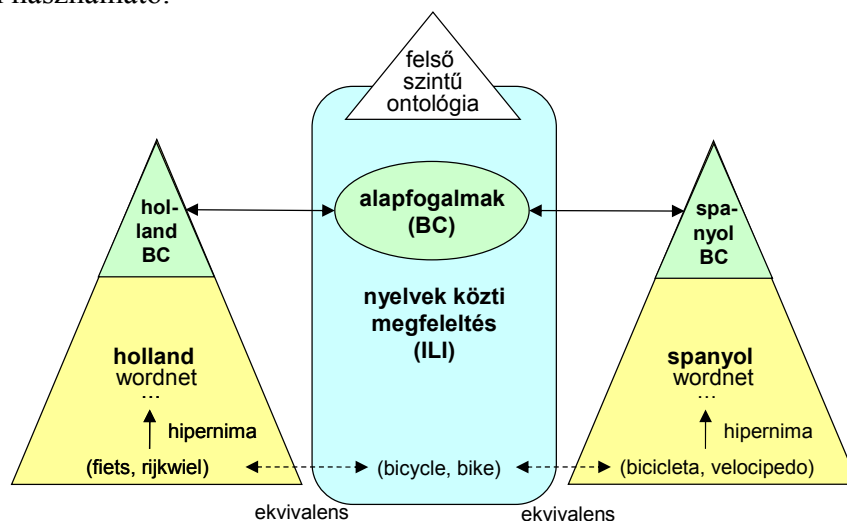
<sup>4</sup> Ez a folyamat elménknek az anyanyelvi környezetben hallott hangokra való „ráhangolódásával” veszi kezdetét, majd hangsorok azonosításával, (látszólag értelmetlen) ismételtetésével és „bevésésével” folytatódik, végül pedig a beszédben gyakran hallott szavak, tagoló hangsorok jelentésének és szerepének felismerésében teljessé válik ki, amikor ezeket már nemcsak kötni tudjuk emlékképeinkhez, hanem akaratlagos generálásukkal a kívánt hatást tudjuk elérni saját környezetünkben.

<sup>5</sup> Mondatok minden alkotó részére kiterjedő.

A WordNet eredetileg az angol nyelv számítógépes értelmező szótárának készült, amelynek kettős célja volt: egyrészt a szótár és a teaurusz (szinonimatár) funkciók kombinálásával intuitíven jól használható nyelvi lexikon létrehozása<sup>6</sup>, másrészt támogatni az automatikus szövegelemzést és a mesterségesintelligencia-alkalmazásokat.

Az eredeti WordNet-t a Princeton Egyetem Cognitive Science laboratóriumában George A. Miller pszichológiaprofesszor irányításával hozták létre, és tartják ma is karban. Az 1985-ben elindult fejlesztés eredményeként előállt adatbázis ma ingyen letölthető és megszorításokkal szabadon felhasználható, ill. interaktívan is böngészhető. 2006-ra ez az adatbázis 150000 szót, több mint 115000 szinonimacsoportot és 207000 szójelentést tartalmazott, amely tömörítve 12 MB méretű. Az adatbázishoz való programozási interfészt a Jawbone és a Natural Language Toolkit Lite projektek keretében Java, ill. Python programozási nyelven készült eszközök biztosítják.

Az EuroWordNet egy többnyelvű adatbázis számos európai nyelvre (eredetileg holland, olasz, spanyol, német, francia, cseh és észt), amely ugyanolyan elvekre épül és hasonló struktúrával rendelkezik, mint az előzőekben említett amerikai WordNet az angol nyelvre. Ez úgy értendő, hogy az egyes nyelvekre elkészített változatok ugyan egyedi, nyelvre specifikus<sup>7</sup> lexikai rendszerek, azonban össze vannak kapcsolva egy ún. Inter-Lingual-Index (ILI) nevű általános lexikkal is, amely a Princeton WordNet-en alapul. Ezen az indexen keresztül a nyelvek oly módon kapcsolódnak egymáshoz, hogy az egyik nyelv valamelyik szavából kiindulva meghatározhatjuk egy másik nyelv hasonló szavait. Az index legfelső szintje 63 nyelvfüggetlen jelentéscsoportot tartalmaz (**3. ábra**). Az 1999-ben befejeződött projekt sikeresen bemutatta, hogy információkinyerésre mind egynyelvű, mind többnyelvű környezetben használható.



3. ábra: EuroWordNet – a többnyelvű wordnet

Az EuroWordNet jól kialakult specifikációját 2000-től egyre több európai és Európán kívüli nyelvre kezdték el alkalmazni, és az elkészülő kompatibilis wordnet-eket az indexen keresztül fogják hozzákapcsolódni a többi wordnet-hez. Jelenleg svéd, norvég, dán, görög, portugál, baszk, katalán, román, litván, orosz, bolgár, szlovén és magyar nyelvekre van folyamatban a kiépítésük.

A Princeton WordNet és az EuroWordNet által inspirált együttműködés az ún. Global WordNet Association (GWA) keretében folytatódik tovább, amelynek célja – természetesen –

<sup>6</sup> A szavakat *szinonimacsoportokba* (synset) szervezi, és rövid, általános *meghatározásokat* ad rájuk, valamint rögzíti a szinonimacsoportok közötti jelentéshordozó (szemantikus) *kapcsolatokat*.

<sup>7</sup> A szükséges mértékben figyelembe veszik az adott nyelv sajátosságait, de igyekeznek a lehető legnagyobb mértékben kompatibilisek maradni a WordNet elveivel és felépítésével.

további wordnet-ek létrehozása, valamint további szabványosítás, mélyebb összekapcsolás, eszközfejlesztés és információterjesztés. A GWA 4. konferenciáját 2008 januárjában Szegeden tartották.

### 2.3.1.2 NooJ rendszer

A NooJ rendszer egy nagyon gyors, hatékony szövegelemző rendszer, amelyet *Max Silberstein* francia nyelvész készített, és ma már számos egyetemen használják tucatnyi nyelv elemzésére. Valójában egy *integrált nyelvelemző környezet* a programozásban használt integrált fejlesztői környezetek mintájára. Elődje az INTEX rendszer.

A NooJ nemcsak számítógépes nyelvészeknek való, hanem hasznos eszköz lehet mindenki számára, aki természetes nyelvű szövegeket kíván *bármilyen* céllal elemezni.

A NooJ nyelvfüggetlen rendszer, amelynek segítségével akár nulláról is felépíthető egy grammatika. A szövegek NooJ-re épülő nyelvtani elemzése során minden egységesen végesállapotú transzducerként<sup>8</sup> (tkp. jelátalakítóként) működik.

A NooJ rendszer gyors, robusztus és könnyen kezelhető (pl. a lexikon és a morfológia szövegfájlban szerkeszthető, a nyelvtanok gráfok formájában intuitív felületen készíthetők).

### 2.3.2 Szövegelemzés BI-technológiákkal

A mai strukturált adatintegrációs eszközök (pl. ETL<sup>9</sup>, replikáló és virtualizáló eszközök) csak erősen korlátozottan képesek strukturálatlan környezetekhez kapcsolódni (pl. tartalomkezelő eszközök), az adattárak tartalmát és a metaadatokat értelmezni, valamint a komplex, strukturálatlan információkat elemezni.

Mindazonáltal ezek azok az eszközök, amelyek a szövegelemző technikákat a leghamarább be fogják fogadni. Sőt azzal, hogy ezek az eszközök a háttér-információk egyre szélesebb köréhez képesek egyre rugalmasabban hozzáférni, be tudják kapcsolni ezeket magába a szövegelemzésbe is (pl. névelemek felismerése) növelve a hatékonyságot és eredményességet.

#### 2.3.2.1 ETL-eszközök

Elkülönülő adatforrásokhoz való hozzáférést és ezek integrációját támogatják, lehetővé teszik az ilyen adatok struktúrájának, formájának és tartalmának megváltoztatását, és ezeket különböző adatbázisok felé tudják szolgáltatni.

Az ETL-eszközök ma már kezdenek strukturálatlan adatok kezelésének képességével is rendelkezni, ha még erősen korlátozott mértékben is. A legtöbb, amire képesek, hogy kapcsolódni tudnak bizonyos típusú strukturálatlan adatforráshoz; el tudnak fogadni XML-struktúrákat; és szintaktikusan elemezni (parszolni) tudnak riportokat, dokumentumokat, pdf fájlokat. Ilyen pl. az Informatica-tól a *PowerCenter*, a Pervasive Software-től a *Data Integrator* és az IBM WebSphere-nél a *DataStage*.

Az ETL-eszközök képességei azonban ma még igen gyengék a jelentésfelismerés, az összetett struktúrájú dokumentumok feldolgozása, a kialakuló dokumentumkezelő szabványok felhasználása és az összetett médiatípusok kezelése területén.

#### 2.3.2.2 Adatvirtualizáció<sup>10</sup>

Az adatvirtualizációra épülő technológiák lehetővé teszik a felhasználók számára, hogy különböző forrásokból származó adatokat egységes (virtuális) nézetten keresztül láthassák.

---

<sup>8</sup> A transzducer olyan (általában elektronikai, elektronikus vagy elektro-mechanikai) készülék, amely az energiát különböző formái között átalakítja valamilyen célból, pl. mérés vagy információátvitel. Tágabb értelemben minden jelátalakító készülék transzducernek tekinthető.

<sup>9</sup> Extract-Transform-Load: az operatív adatbázisokból az adatokat *kinyerő*, szükség szerint *átalakító* és valamilyen adattárházba *betöltő* eszköz.

<sup>10</sup> vállalati információintegrációnak (enterprise information integration – EII) is nevezi néhány gyártó

Maguk az adatok a helyükön, az adott forrásban maradnak, de a (virtuálisan) integrált nézetük csak valamilyen, az alkalmazások által elért és használt memóriaterületen állítódik össze ideiglenes jelleggel.

Az adatvirtualizáló eszközök ma még többnyire strukturált adatforrásokkal (relációs adatbázisokkal) működnek együtt. Néhány ilyen eszköz azonban XML-alapú dokumentumokat és dokumentumtárakat is kezelni képes, és keresési funkciókat is kínál. Ilyen eszköz az IBM WebSphere *Information Integrator*-a, az Ipedo *XIP*-je vagy a *MetaMatrix Enterprise*.

Habár az adatvirtualizáló eszközök egyre többet kínálnak a dokumentumkezelő szabványok támogatása területén, továbbra is gyengék a szemantikus képességeik és az összetett médiatípusok kezelése tekintetében.

### **3. szakasz: átalakítás és megjelenítés**

A szövegelemzés utolsó lépése az olyan végtermék előállítás, amely kielégíti a felhasználó igényeit, mint pl. riport, diagram vagy valamilyen adatbázisba importált adatrekordok. Idetartoznak a *hálódiagramok* (ld. alábbi példát), amelyek az objektumkapcsolatokat és fogalom-összefüggéseket mutatják be, az olyan *böngészők* és *navigációs eszközök*, amelyek a kinyert információt visszavetítik a forrásszövegekre, és más olyan megjelenítő felületek, amelyek a szövegelemzés eredményeit interaktív vizsgálat számára teszik elérhetővé.

A felhasználónak ellenőrzést kell tudni gyakorolnia előlött a végtermék fölött. A szövegelemző szoftvernek emiatt biztosítania kell a manuális annotálás lehetőségét, ill. az elemek automatikus annotációjának megszüntetését. Szokásos megoldás az, hogy valamilyen szövegszerkesztőbe importálva az annotáció vizuálisan megjeleníthető és továbbszerkeszthető, vagy egy diagramkezelő, ill. más harmadik féltől származó eszközben az annotációk pl. linkdiagram formájában válnak kezelhetővé.

Ma már egyértelmű elvárás, hogy a szövegelemzés végtermékét XML-formátumban (is) elő kell állítani, valamint a megfelelő XML-sémaleírást dokumentálni kell, hogy más alkalmazások azt használhassák. Ugyanakkor azonban az operatív alkalmazási rendszerekkel való integráció ma még erősen ad hoc jellegű.

#### **2.3.3 Önszervező háló (Self-Organising Map – SOM)**

Alkalmazási szempontból az önszervező háló rendkívül hatásos diagrammatikus eszköz nagy dimenziószámú<sup>11</sup> objektumhalmazok megjelenítésére és interaktív vizsgálatára. A módszer alapja egy olyan neurális hálózaton alapuló tanuló algoritmus, amely a felügyelet nélküli tanulás egyik legnagyobb számításigényű fajtáját valósítja meg.

Az önszervező hálók általános célja az, hogy valamilyen adathalmazban egy rejtett, belső *hasonlósági struktúrát* fedjenek fel<sup>12</sup>. Ennek érdekében képesek:

1. nagy dimenziószámú objektumokat alacsony dimenziószámúban ábrázolni anélkül, hogy az adatok „lényege” elveszne;
2. az adatokat hasonlóság alapján úgy strukturálni, hogy a (hasonló) objektumokat geometriailag közel ábrázolja egymáshoz.

Az SOM-et bámulatosan sok területre alkalmazták már sikeresen, pl. hang- és kézírásfelismerés, génkifejeződési minták feltárása, és nem utolsónak szöveges dokumentumok elemzése. Az utóbbi esetben a gyűjtemény minden dokumentumát szavak vektoraként reprezentálják (ún. vektortérmodell, ld. **1. táblázat****Hiba! Érvénytelen könyvjelző-hivatkozás.** oszlopai), ahol a vektorelemek a gyűjteményben szereplő összes

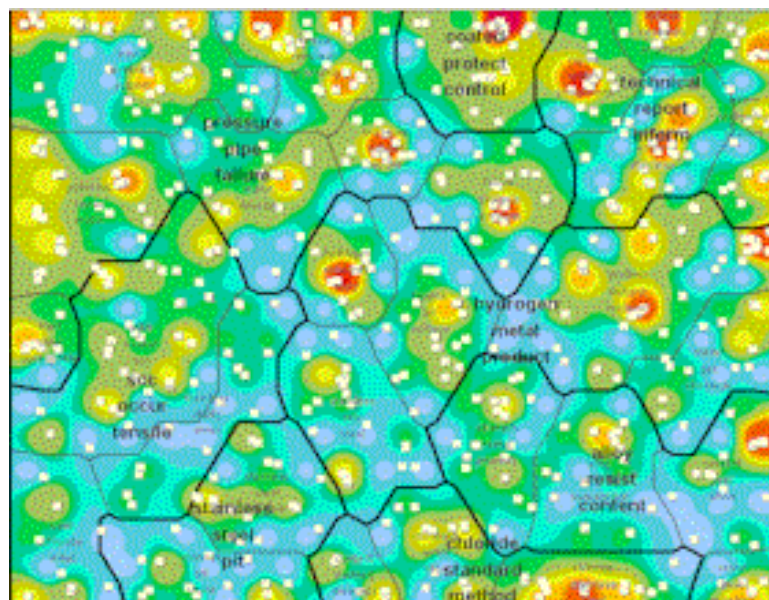
<sup>11</sup> Értsd: sok (valóban vagy látszólag független) adattal jellemezhető

<sup>12</sup> Az önszervező háló az emberi agynak azt a szerveződési mintáját próbálja modellezni egy egyszerű módszerrel, hogy a neuronok csoportos elrendeződést alakítanak ki, és a csoporton belüli neuronok kapcsolódása sokkal erősebb, mint a kapcsolódás a csoporton kívüli neuronokkal.

lényeges szó dokumentumbeli előfordulásainak gyakoriságát adják meg. Ezek a vektorok alkotják ebben az esetben a korábban említett „objektumokat”, és a szóelőfordulások gyakorisága az „adatokat”. A dokumentumok hasonlóságát a megfelelő vektorok között definiált távolságmérték alapján lehet számítani. Az 1. táblázat **Hiba! Érvénytelen könyvjelző-hivatkozás.**nak megfelelő önszervező háló (betanítás és futtatás után) a 4. ábra által mutatott áttekintő képet adja eredményül, ahol a sárgás-piros színű csoportok a nagyon hasonló „témaszervezetű” dokumentumokat fogják össze. A diagram mutatja a csoportokat képző és a témákat meghatározó domináns szavakat, és – megfelelő eszközzel – akár több szinten, fokozatosan finomodó felbontásban, interaktívan tudja mutatni a dokumentumcsoportokat, ill. megmutathatja magukat az odatartozó egyes dokumentumokat is.

1. táblázat: Egy dokumentumgyűjteményben a szóelőfordulások gyakorisága (példa)

	D1	D2	D3	...	D999	D1000
hydrogen	0.89					
metal	0.73					
product	0.75					
alloy		0.67				
resist		0.53				
content		0.58				
chlorid			0.44			
standard			0.67			
method			0.55			
stainless					0.45	
steel					0.96	
pit					0.87	
soc						0.76
occur						0.78
tensile						0.79
...						



4. ábra: Dokumentumcsoportok SOM-képe<sup>13</sup>

<sup>13</sup> A képen a dokumentumokat fehér pontok jelölik. A fekete vonallal körülhatárolt területekben lévő dokumentumok „hasonlóak” abban az értelemben, hogy a képen a területre írt szavak fordulnak elő gyakran

Az SOM által nemcsak, hogy fogalmat alkothatunk arról, mit is jelent a valóságban a csoportosítás (klaszterezés), de az eredményül kapott diagram térben is jól értelmezhető. Azaz nemcsak azt mutatja meg, hogy egy dokumentum milyen csoportba tartozik, de azt is, hogy mi a csoporthoz vagy annak középpontjához való „kötődésének foka”. Mindezen információt két dimenzióban színek segítségével ábrázolja (ld. **4. ábra**).

## 2.4 Alkalmazás

A különbözőféle szervezetek természetesen már régóta tudják, milyen betekintésre tehetnének szert a tevékenységükön belül, ill. kívül keletkező szövegek elemzésével. A közel múltig azonban az olyan technológia, amely ezt lehetővé tenné, nem volt kereskedelmileg életképes. Mára a helyzet jelentősen megváltozott, és a cégek a szövegelemzést már elég erős eszköznek tartják arra, hogy a rendelkezésükre álló, strukturálatlan adattömegekből közvetlenül használható, üzleti folyamataikban „bevethető” információt generáljanak. A technológia gyorsuló ütemben hagyja el a korai kísérletezések fázisát, és válik egyre érettebbé. A gyártók 30-50%-os növekedési ütemről számolnak be.

E növekedés mögött elsősorban a következő tényezők állnak: 1) a technológia jobb megértése, 2) a technológia beérése és 3) az a lehetőség, hogy az üzleti intelligenciába (BI) történt befektetéseket kiterjeszthetik a strukturálatlan információkra is. Az ilyen eszközök piaca meglehetősen dinamikus: a kisebb fejlesztő cégeket már felvásárolták a nagyobbak, és folytatódik a beintegrálódásuk a keresőmotorok, a BI és a tartalomkezelés piacaiba.

Egy piackutatás<sup>14</sup>, amelyet nemrégiben 118 nagy (200 millió dollár bevétel feletti) amerikai cég körében végeztek, megállapította, hogy:

- a szövegelemző megoldások üzembe állítását és az így előállított alkalmazásokat már **üzleti szempontok vezérlik**; a legtipikusabb alkalmazások ügyfélközpontúak: azoknak a megkérdezetteknek a 75%-a, akik tervezik ilyen technológia felhasználását, ügyfélkapcsolati területen tervezik azt;
- a **korai kísérletezők** a szövegelemzési technológiát adatbányászati és üzletiintelligenciamegoldásaikkal összefüggésben alkalmazták; tipikus, hogy a szövegből kinyert információkat strukturált információkkal integrálják, majd azt specifikus adatbányász vagy BI-megoldásokkal vizsgálják tovább;
- a korai kísérletek során **számos kihívással** szembesültek: 1) adatkezelési kérdések, mint pl. strukturálatlan adatokhoz való hozzáférés, adatkinyerésre való előkészítés és a tárolás legpraktikusabb módjai 2) a szokatlan funkcionalitásból adódó szoftverfejlesztési problémák 3) a taxonómia- és ontológiafejlesztés nehézségei;
- a meglévő kihívások ellenére a kísérleteket végzők növekvő és **egyre kézzel foghatóbb értéket** tudnak előállítani a szövegelemzési technológia segítségével: a megkérdezettek 73%-a úgy ítélte meg, hogy az ilyen jellegű elemzések nagy vagy kritikus fontosságúak.

A szövegelemzést végző szoftvereket jelenleg még elsősorban az új technológiákat általában is korán alkalmazó cégek használják, piaci elfogadottságuk más technológiákhoz képest még alacsony, de növekvő, és már számottevő azon üzleti funkciók köre, amelyek már ma is alkalmaznak szövegelemzést, pl.:

- Ügyfélkapcsolat-elemzés

**A szövegelemzés jelentős szerepet kezd betölteni olyan általános üzleti funkcióknál, mint az ügyfélkapcsolat-kezelés (Customer Relationship Management – CRM). A használt források közé tartoznak az ügyfél üzenetei és levelei, hívásközpontok feljegyzései és átiratai, és a CRM-rendszerek szöveges részeiben tárolt adatok. A cél**

---

bennük. Minél pirosabb egy-egy részterület, annál nagyobb a hasonlóság a részterületbe eső dokumentumok között. Az ábra egyes részterületeire kattintva részletesebb képet lehet kapni.

<sup>14</sup> The Road to Understanding Your Company's Unstructured Information (Fred Harper, Hurwitz & Associates), 2007.

**itt a termékek és szolgáltatások minőségével kapcsolatos kérdések azonosítása, a terméktervezés és -fejlesztés támogatása, valamint a kapcsolatfelvételi kísérletek megfelelő helyre történő irányítása.**

- Orvosbiológia: tünetek okozati hátterének kutatása  
A használt források közé tartozik a klinikai szakirodalom, gyógykezelési feljegyzések és jelentések, vegyi anyagok, gyógyszerek és orvosi tünetek adatbázisai, stb. A cél olyan kapcsolatok kinyerése, amelyek orvosbiológiai és kémiai anyagok, genetikai jelzők (pl. fehérjék, gének, bázisszekvenciák) valamint tünetek között állnak fenn – azaz olyan minták meghatározása, mint „A gátolja B-t”, „A aktiválja B-t” vagy „A összefüggésben áll B-vel”. Az ilyen alkalmazásnál az információkinyerés géneket, fehérjéket, kémiai anyagokat, tüneteket és tünetegyütteseket azonosít, valamint megkülönböztető jegycsoportokat (tkp. átvitt értelemben „aláírásokat”) orvosbiológiai és kémiai szövegekből felhasználva szakterületi szótárakat. Szükség van időbeli kapcsolatok, ill. más jellegű minták észlelésére is.
- Gyógyászat: diagnosztika javítása és ellenőrzése  
A használt források közé tartoznak klinikai-kutatási adatbázisok, betegnyilvántartás, biztosítási és jogszabályi nyilvántartások. A cél javítani a diagnosztikát, csökkenteni a téves diagnózisokat, biztosítani a megfelelő gyógykezelést, elősegíteni a minőségi szolgáltatást, növelni a kihasználást, csökkenteni a csalások, visszaélések számát, és ellenőrzés alá vonni a költségeket.
- Hírszerzés és terrorelhárítás  
A használt források közé tartoznak a hírek, vizsgálati jegyzőkönyvek, lehallgatott üzenetek, ügyiratok – különböző nyelveken. A cél szervezeti kapcsolatok és hálózatok felderítése, viselkedési és támadási minták azonosítása, veszélyelemzés, esemény-előrejelzés, stratégiai és taktikai értékelés.
- Rendészet  
A használt források közé tartoznak az ügyiratok, a bünyügyi és bírósági jelentések, jogi dokumentumok, földrajzi és demográfiai információk. A cél bűnelkövetési minták (időbeli, földrajzi eloszlás ill. személyek és szervezetek szerepe) feltárása, és a bünyügyi vizsgálatok és bünyvádi eljárások támogatása.
- Pénzügyintézetek: Értékpapírcsalások felderítése  
A használt források közé tartoznak a pénzügyi jelentések és hírek, vállalati ügyiratok és nyilvántartások, kereskedelmi és más tranzakciók feljegyzései. A cél az ún. „bennfentes” kereskedelem észlelése, beszámolás szabálytalanságokról, pénzmosási és más illegális tranzakciókról, valamint árképzési rendellenességekről.

További fontos alkalmazások: a jogszabályi következmények feltárása; szabadalmi vizsgálatok; alkalmazottak felvételénél az önéletrajzok feldolgozása; szabadszöveges válaszokat tartalmazó közvélemény-kutatások elemzése. E használati esetekben gyakran a szövegből kinyert információkat számszerű adatokhoz is hozzákapcsolják.

Végül a szövegelemzés új funkcionális alkalmazások létrehozását is elősegíti, mint pl. *ismertségértékelés*. Ez hírek, weboldalak, piaci jelentések, levelezések és más dokumentumok összegyűjtését, feldolgozását jelenti; majd információk kinyerését meghatározott fogalmak szerint, mint pl. vélemény, értékelési szempontok és súlyok; végül ezen információk elemzését. Szövegelemzésre képes eszközök nélkül az ismertségértékelés túlságosan költséges, lassú lenne, és az információknak csak nagyon kis körére terjedhetne ki. Hasonlóképpen a *társasági hálózatok elemzésére* szolgáló eszközök üzeneteket és a kommunikáció más elemeit, vállalati dokumentumokat, híreket vizsgálnak meg, hogy meghatározhassák a személyek és szervezetek kapcsolati rendszerét, és hogy kapcsolatfelvételi kísérleteiket a legjobb úton indíthassák el. Az ilyen jellegű elemzés lehetetlen lenne szövegbányászat nélkül.

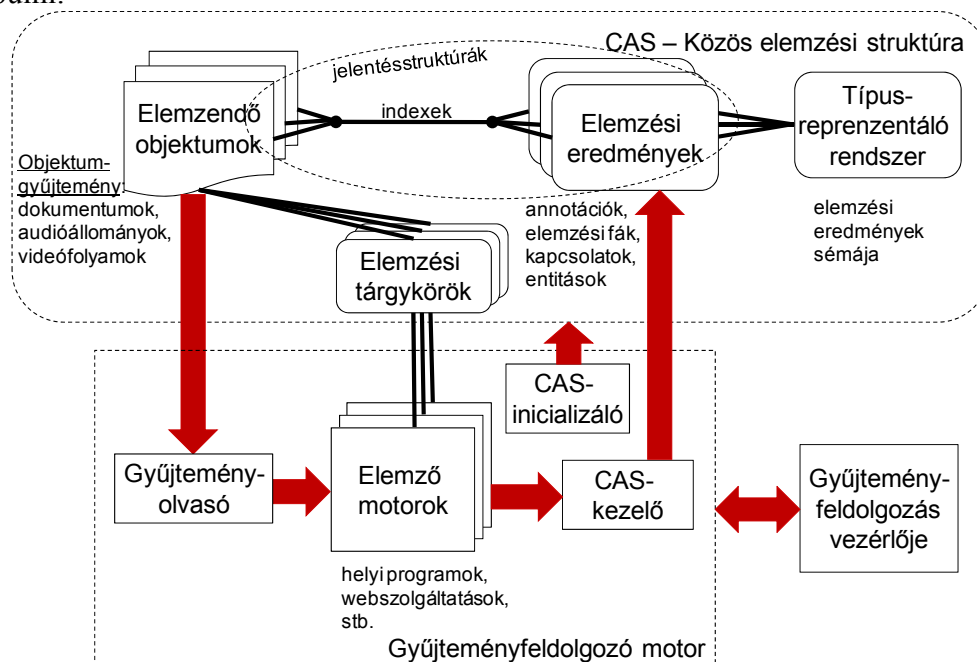
### 3. Folyamatban lévő fejlesztések

#### 3.1 UIMA – nyílt forráskódú fejlesztő rendszer szövegelemzéshez

Az UIMA (Unstructured Information Management Architecture) egy nyílt forráskódú fejlesztői keretrendszer, amely lehetővé teszi, hogy segítségével szövegelemző alkalmazásokat hozzanak létre. Az UIMA az IBM korábbi NLP-projektjeinek (Natural Language Processing) eredményeire épül<sup>15</sup>.

Az UIMA SDK (Software Development Kit) szabadon letölthető az IBM alphaWorks webhelyéről (<http://www.alphaworks.ibm.com/tech/uima>), a forráskód pedig a SourceForge webhelyén (<http://uimaframework.sourceforge.net>) férhető hozzá.

Az IBM-es indíttatás következtében nem meglepő, bár mindenképpen figyelemre méltó, hogy az UIMA elemző és annotációs moduljai különböző IBM-termékekben felhasználásra kerültek, pl. az IBM új, vállalati keresést (enterprise search) megvalósító megoldásaiban a WebSphere Information Integrator OmniFind Edition-ben és a WebSphere Portal Server-ben. A tervek szerint az IBM-termékekben használandó minden új elemzési technológia az UIMA-ra fog épülni.



5. ábra: Az UIMA fő architektúráis elemei

Az UIMA több területen is irányt szab a szövegelemzés élenjáró technológiáinak a kereskedelmi felhasználásnál:

- különböző köztesszoftver-környezetekben képes elosztott működésre,
- egyszerűen beágyazható különböző szoftvertermékekbe (alapvető követelmény kereskedelmi alkalmazásoknál),
- támogatja a multimodális – azaz szöveg-, audio- és videóállományokat feldolgozó – alkalmazások fejlesztését és működtetését,
- gyűjteménykezelési architektúrája szélesíti az alkalmazási lehetőségeket,
- metaadatok kiterjedt használatával lehetővé teszi az alkalmazáshoz szükséges komponensek dinamikusan meghatározását és konfigurálását.

Az UIMA több ún. *elemzési tárgykört* is támogat, azaz egy adott *elemzendő objektumra* több nézetet is meg lehet adni. Például, egy videófolyam esetén egy nézetet alkothatnak a

<sup>15</sup> Ld. IBM Systems Journal 43. kötet, 3. szám (<http://www.research.ibm.com/journal/sj/433/ferrucci.html>)

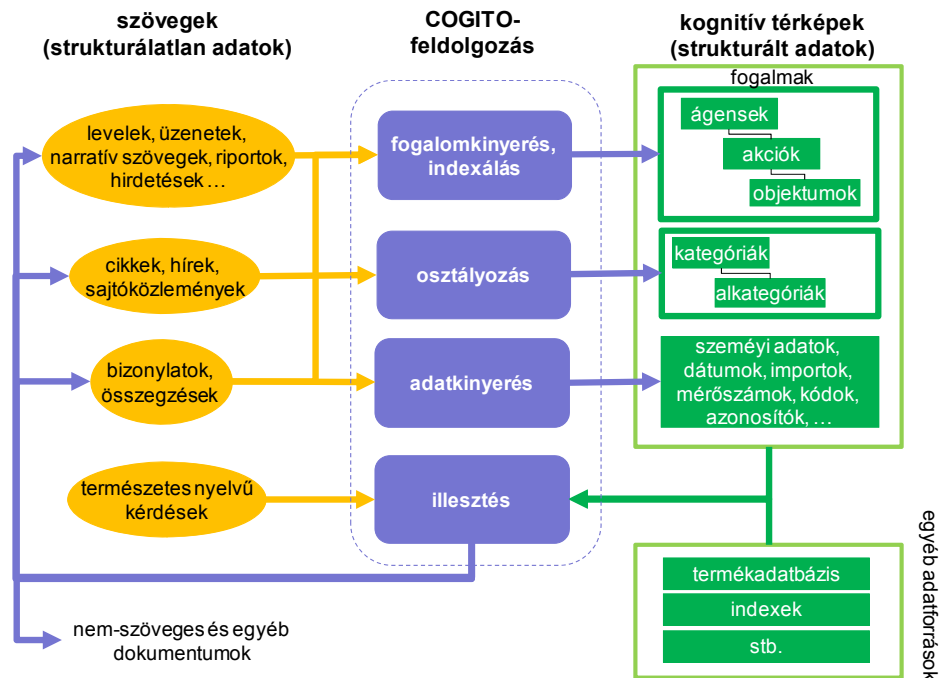
videóképek, egy másikat pedig a képaláírások szövegei, és mindkét nézetre külön *elemző motort* lehet alkalmazni.

A *CAS-kezelő* amellet, hogy összegyűjti az *elemzés eredményeit* egy gyűjteményben, képes az egy gyűjteményhez tartozó dokumentumok között állapotkövetésre is – általában valamilyen adatbázis- vagy kereső motor segítségével.

Az UIMA-keretrendszer biztosítja, hogy komponenseit, mint pl. az *elemző motorokat* és a *CAS-kezelőket* akár webszolgáltatásként, akár valamilyen jól-skálázható köztes szoftver részeként ki lehessen fejleszteni. Az UIMA-alkalmazások általában több dokumentumot egyszerre is elemeznek. Az UIMA-keretrendszer egy másik komponense, a *gyűjteményfeldolgozás vezérlője* számos olyan lehetőséggel rendelkezik, amelyekkel rugalmasan növelni lehet a teljesítményt és biztosítani a helyreállíthatóságot.

### 3.2 COGITO – kognitív térképek és adatforrások a szövegelemzésben

Az Expert System ([www.expertsystem.it](http://www.expertsystem.it)) cég COGITO szoftverrendszere 10 év és több száz emberév élenjáró nyelvtechnológiai fejlesztésének eredménye, amely a szövegek strukturális és lexikai egységeit felismerve képes azok jelentését meghatározni.



6. ábra: A COGITO feldolgozási folyamata

A COGITO jogilag védett szemantikai platformja a legkorszerűbb szövegelemző technológia, ami jelenleg kereskedelmi forgalomban kapható. Felhasználja a *Sensigrafo* nevű szemantikai hálót, amelynek képességei elengedhetetlenek a jelentésegyértelműsítéshez, és amely tartalmaz több mint:

- 3 500 000 kapcsolatot (pl. társítások, konstrukciók, témakörök és tárgykörök, idiómák, kifejezésmódok ...),
- 600 000 szemantikai fogalmat,
- 400 000 hiponima–hipernima kapcsolatot,
- 55 000 hipernima–troponima kapcsolatot,
- 370 000 kapcsolatot a korpuszhoz és több tízezer utalást a témakörök, objektumok, meronimák stb. között.

A COGITO által végzett feldolgozás (ld. **Hiba! A hivatkozási forrás nem található.** ábra) eredménye egy ún. kognitív-konceptuális térkép, azaz a bejövő strukturálatlan adatok

értelmezését lehetővé tevő, strukturált ábrázolás, amely mintegy vezérli a szöveg leglényegesebb elemeinek automatikus feldolgozását.

Ellentétben a szokásos egyszerűbb megoldásokkal a COGITO nem hanyagolja el a többnyire korán kiszűrt ún. STOP-szavakat, azaz nyelvtani minősítő és kötőszavakat sem (névelő, prepozíciók stb), és kombinálja a nyelvtani elemzést a szemantikai ismeretkezeléssel.

A COGITO-t nyílt és jól-skálázható felépítése miatt elég egyszerűen lehet különböző típusú tartalmak irányában testre szabni: pl. műszaki dokumentáció, enciklopédikus anyagok, pénzügyi információk, újsághírek stb. feldolgozására. A COGITO-t kiemelkedő elemzési képessége és komponensei közti integráltsági szint teszi napjaink egyik vezető szövegelemző eszközévé. Fő komponensei:

- **a nyelvi elemző: morfológiai, grammatikai és szintaktikus elemzése a mondatnak**
- **a lexikon: szavak és alapjelentéseik felismerése szemantikai háló segítségével**
- **a memória: az előző elemzések eredményeinek nyomon követése, kontextusmeghatározás a jelentésegértelműsítés és optimális jelentéshozzárendelés érdekében**
- **az ismeretkezelő: való világbeli, mindennapi ismeretek ábrázolása leíró szabályok formájában, amelyek közvetlenül felhasználhatók az elemzés során**
- **a tartalomábrázoló: a szövegek tartalmának ábrázolása kognitív térkép formájában, amely többféle célra is felhasználható, pl. indexálás, osztályozás, összegzés, fordítás.**

Jelenleg az angol, olasz, francia nyelveket támogatja, de hamarosan megjelenik a német, spanyol és portugál verziója is.

### 3.3 Cyc-NL – szemantikus értelmezés a szövegelemzésben

Széles körben elfogadott, hogy a Cyc (<http://www.cyc.com/cyc/technology/whatisycyc>) a legnagyobb és technikailag legjobban felépített szemantikus rendszer. Van nyílt változata is ([www.opencyc.org](http://www.opencyc.org)), amelynek fogalmait nemrégiben elkezdték összekapcsolni a Wordnet szinonímacsoportjaival is.

A Cyc természetes nyelvű alrendszere (Cyc-NL) a legegyszerűbben három fő komponensén, a *lexikonon*, a *szintaktikus elemzőn* és a *szemantikus értelmezőn* keresztül írható le.

A lexikon angol szavakról tartalmaz szintaktikus és szemantikai információkat. Amikor a Cyc-NL a szöveg mondatait feldolgozza, akkor először azt ellenőrzi le a lexikonban, hogy lehetséges-e a szöveg egyes elemeit a lexikon szavaihoz hozzárendelni. Ekkor még sok szó esetében igen bizonytalan, hogy mi a jelentése vagy milyen szerepet tölt be a szövegben. Ez a feladat a szintaktikai elemzőre hárul, azaz hogy eldöntse, az egyes szövegrészek mely hozzárendelése megfelelő, valamint, hogy olyan mondat szerkezetet építsen fel, amelyet a szemantikai értelmezőnek át lehet adni.

A Cyc-NL szemantikus értelmezője a szintaktikai szerkezetet olyan ún. CycL-formulákká<sup>16</sup> alakítja át, amelyek már összevethetők a Cyc ismeretbázisával vagy egy SQL-generátoron keresztül kérdésként feltehetőek valamilyen relációs adatbázisnak.

A Cyc-NL szemantikus komponense az értelmezési folyamat szinte minden szintjén használja az ismeretbázist. Például, az „*A fiú meglátta a fényeket a távcsővel.*” mondatban ez a komponens ellenőrzi az ismeretbázisban, hogy vajon a távcsövet használják-e inkább eszközként a látás során, vagy éppen fordítva a fény az egy olyan dolog-e, amelynek rendszerint távcsőve van. Az ellenőrzés eredményeként a szemantikus értelmező természetesen ez utóbbi feltevést el fogja vetni, és az első esetnek megfelelő CycL-formulát állítja elő. Ezzel pont ellentétesen „*A férfi meglátta a nőt a kalappal.*” mondatban a Cyc-NL a második értelmezést fogja valószínűsíteni, azaz hogy olyan nőt látott meg, akinél kalap volt,

---

<sup>16</sup> A CycL (Cyc Language) a Cyc ismeretprezentációs nyelve.

és férfi nem egy „kalap segítségével” látta meg a nőt, mivel a látás során nem szoktak kalapot eszközként felhasználni.

Az ilyen mindennapi, az értelmezési folyamatot vezérlő, egyszerű ismeretek használata teszi lehetővé, hogy a Cyc-NL a természetes szövegekben mindig jelenlévő többértelműségeket feloldja anélkül, hogy kizárólag statisztikai technikákra kelljen hagyatkoznia.

### 3.4 ConceptNet – praktikus kontextuális következtetés a szövegelemzésben

A ConceptNet a legnagyobb szabadon hozzáférhető, mindennapi tudást tartalmazó ismeretbázis és természetes nyelvű elemző eszköz, amely sok gyakorlati, következtetési feladatot tud támogatni a szövegelemzésnél, pl. témagenerálás, analógiakészítés stb. A ConceptNet a legfontosabb számítógéppel feldolgozható erőforrás, amit az Open Mind Common Sense projekt (OMCS - <http://openmind.media.mit.edu>) által létrehozott korpuszból előállítottak. 2004-ben már 1,6 millió kapcsolatot tartalmazott 300 000 fogalmi csomópontja között, amelyek kiterjednek a mindennapi élet térbeli, fizikai, közösségi, időbeli és pszichológiai vonatkozásaira.

Míg a WordNet és a Cyc „tudásmérnökök” hosszú éveken, évtizedeken keresztül végzett kézműves tevékenységének eredményeként jöttek létre (és fejlődnek folyamatosan), addig a ConceptNet-et automatikusan hozzák létre az OMCS-korpuszból, amelynek tartalmát azonban a legnagyobb nyilvánosság bevonásával állítják össze. Az ötlet az, hogy a laikusok is érdemben hozzá tudnak járulni a „józan ész”-nek megfelelő ismeretek összegyűjtéséhez, nem kell a szakértőkre várni. Az OMCS-projektben kb. 14000 szerző webes együttműködése mellett 700 000 mondatot gyűjtöttek be és rendszereztek.

Ilyen módon a ConceptNet ismeretbázisát az OMCS-korpusz félig strukturált, angol nyelvű mondataiból egy automatizált folyamat hozza létre, amelynek során az ismeretbázis minden állítása aszerint értékelődik, hogy hányszor lett az adott állítás a korpuszban megfogalmazva, és milyen mértékben lehetett közvetett módon származtatni a ConceptNet-ben lévő egyéb, más állításokból.

A ConceptNet-nek a beépített ismeretbongésző mellett egy olyan természetes nyelvű szövegfeldolgozó motor is része, amely szövegekben kontextus-orientált *következtető feladatokat* tud ellátni, mint pl. témagenerálás, témakiemelés, jelentés egyértelműsítés, szemantikus osztályozás, érintettség meghatározás, analógiakészítés stb. Ez a ConceptNet-be integrált nyelvfeldolgozó modul a MontyLingua (<http://web.media.mit.edu/~hugo/montylingua/>) természetes nyelvű felismerő rendszernek az adott célra adaptált változata. A MontyLingua a platformfüggetlen Python nyelven készült, de Java-könyvtárként is rendelkezésre áll.

A ConceptNet-et mérete és terjedelme összehasonlíthatóvá teszi a két legismertebb, nagyméretű szemantikus ismeretbázissal a WordNet-tel és a Cyc-kal. Mindazonáltal lényeges különbségek vannak köztük. Míg a WordNet a lexikai osztályozásra és a szóhasonlóságok meghatározására van optimalizálva, addig a Cyc a formalizált logikai következtetésre, a ConceptNet pedig arra, hogy praktikus, kontextusalapú következtetéseket lehessen végrehajtani valós szövegekben.

A WordNet-hez képest a következő 3 területen vannak a fő eltérések:

- nem csak tisztán lexikai egységeket (szavakat, ill. elemi jelentéssel bíró egyszerű kifejezéseket) tartalmaz, hanem összetettebb pl. igéből és egy vagy több kapcsolódó argumentumból álló kifejezéseket is (pl. „ételt vásárol”, „üzletbe megy”), ami lehetővé teszi, hogy a mindennapi élethez közeli fogalmak jóval szélesebb körét tudja kezelni
- nem csak 3 szemantikus alaprelációt tud kezelni (*hasonló, része és alesete*), hanem húsz különböző szemantikai viszonyt is (pl. *okozata, aleseménye, képes vmire, indítéka, helye, tulajdonsága*), ami lehetővé teszi, hogy a mindennapi élet jelenségeit és helyzeteit közvetlenebbül tudja kifejezni

- nem csak formális taxonomikus igazságokat tud kezelni (kutya -> kutyaféle -> húsevő -> emlős), hanem a mindennapi életben gyakran igaznak bizonyuló, de nem féltétlenül mindig igaz viszonyokat is (pl. a kutya hobbiállat vagy házőrző)

A Cyc a mindennapi tudást formalizált, logikai keretben ábrázolja, ezért a pontosan és egyértelműen leírható helyzetekben körültekintő, deduktív következtetésekre képes. A ConceptNet ezzel szemben a bizonytalan, nem jól leírható helyzetekben is képes valamifajta következtetések levonására.

A ConceptNet ismeretbázisában lévő állítások túlnyomó része (kb. 4/5-e) ún. általános konceptuális kapcsolatoknak (*k-line*) van szentelve. Az ezek által kódolt kontextuális, mindennapi tudást használó következtetések jól alkalmazhatóak a szöveges információk kezelésében, mivel lehetővé teszik, hogy a számítógép nagy vonalakban jellemezni tudja a szöveget különböző, érdeklődésre számotartó dimenziók mentén, mint pl. téma, érintettség. Továbbá azt is lehetővé teszi, hogy a számítógép új, ismeretlen fogalmakat is feldolgozzon az ún. strukturális analógiák alkalmazásával, azaz olyan kontextusokba helyezve ezeket a fogalmakat, amelyek már ismertek.

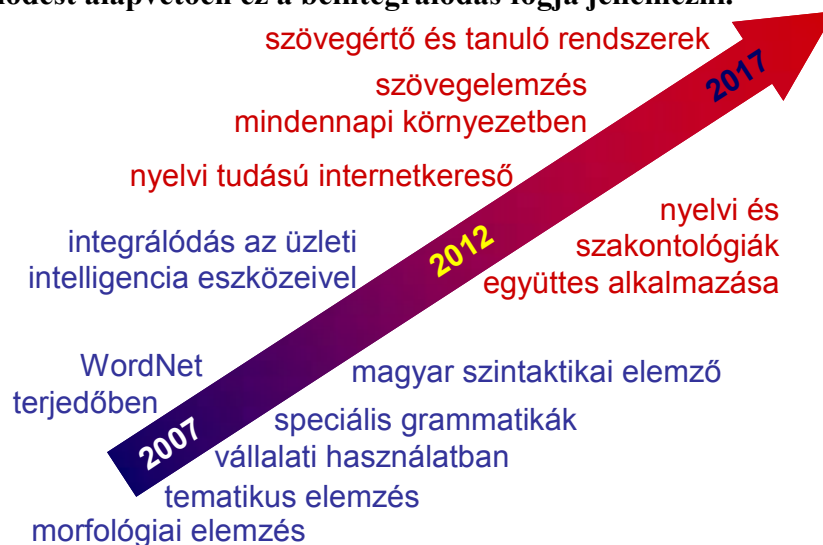
#### 4. A várható fejlődés

A strukturálatlan információk kezelése a jövőben általános használatba kerül, hogy mélyebb betekintést lehessen nyújtani az üzleti tevékenységek során bekövetkezett és rögzített eseményekbe, és magyarázatot lehessen adni rájuk – a piac- és versenyelemzés, ügyfél- és partnerkapcsolat-kezelés a gyártás, a fogyasztói javak, a pénzügy és az élettudományok területén.

A szövegelemzés vonzereje abban áll, hogy jelentősen kitágítja azon források körét, amelyekből használható információkat lehet kinyerni, és azokat további feldolgozásra alkalmas módon ábrázolni. Ilyen módon alaptermék – mindazonáltal nem helyettesíti azokat a jól bevált operatív és menedzsmentszintű informatikai megoldásokat, amelyek az adatok eleve strukturált formában történő felvitelét, és a már strukturált adatok elemzését végzik (pl. adatbázisok, ERP/CRM-rendszerek, üzleti intelligencia).

Az alaptermék jellege inkább azt jelenti, hogy egy eddig – kényszerűen – elhanyagolt terület, a szövegekben rejlő, felszínesen strukturálatlan, valójában azonban mélyen – de gépi feldolgozás számára hagyományosan nehezen kezelhető módon – strukturált információk kinyerése beépülhet potenciálisan minden alkalmazási szintű technológiába.

**A várható fejlődést alapvetően ez a beintegrálódás fogja jellemezni.**



7. ábra: Várható fejlődés a szövegelemzés terén

Más oldalról a szövegelemzést ma alkotó technikák finomodnak, miközben egyre függetlenebbé válnak szakterületektől, nyelvektől – pontosabban: a kapcsolatuk ezekkel egyre modulárisabbá, szabványosabbá, és ezen keresztül lazábbá is válik. Fontos mérföldkő lesz a már nyelvi – bár még egyszerű – képességekkel is rendelkező internetes keresők megjelenése és elterjedése.

A számítógépes szövegelemzés fejlődésének ugyanakkor van egy további – és hosszabb távon talán legjelentősebb – iránya: kiegészül és ötvöződik ismeretreprezentációs és következtetési technikákkal, amelynek hatására kézzel foghatóvá válik, és gyakorlati alkalmazásba kerül a szövegek számítógépes értelmezése (gépi szövegértés).

#### **4.1 Integrálódás kapcsolódó technológiákkal**

A szövegek megfelelő elemzésével és a kinyert információk körültekintő integrálásával a strukturált és strukturálatlan információforrások zökkenőmentesen együtt kezelhetők, és ezzel új dimenzió adható az üzletiintelligencia-megoldásokhoz. Elsősorban a gépi tanulás, az információ-visszakeresés és a természetes nyelvek statisztikai elemzése azok a technikák, amelyek fejlődése lehetővé teszi, hogy nagy mennyiségű szövegből értékes üzleti információkat nyerjenek ki.

Az üzleti intelligencián belül különösen az adatintegrációs eszközöknek lesz kiemelt szerepe, mivel a strukturált és strukturálatlan információk egységes kezelése irányában fejlődnek, és ennek eredményeként a jövőben már szemantikus képességekkel<sup>17</sup> is rendelkezni fognak.

Az információkereső eszközöknek jobban együtt kell tudni működniük a strukturált integrációs eszközökkel, hogy a strukturálatlan tartalmakban rejlő információkat ki tudják aknázni.

#### **4.2 Szakterület-független elemzés**

Fokozatosan olyan eszközök jelennek meg, amelyek azokat a területeket célozzák meg, ahol a jelenlegi adatintegrációs eszközök gyengék, viszont a szövegelemző eszközök hagyományosan erősek pl. jelentésfelismerés, metaadatok meghatározása, kapcsolatok és leképezések kikövetkeztetése.

Ahogy ezek a technológiák piacra kerülnek, szakterület-független szemléletet hoznak magukkal, és már nem szükségképpen irányulnak csak strukturált, vagy csak strukturálatlan adatra: a két világ összekötésére törekednek. Ez a technológia még korai fázisában van, de már vannak gyártói: pl. a Unicorn és a Metatomix. A különböző típusú és eredetű metaadatok egyesítését és konszolidálását lehetővé tevő képességek különösen értékesek lesznek a teljes körű adatgazdálkodást folytató szervezetek számára.

A szakterület-független technológiák megjelenését támogatja a szakontológiák terjedése és felhasználása a szövegelemzésben. A szakterületi fogalmak a szakontológiára korlátozódnak, amelytől működésében teljesen elkülönül maga a szövegelemzés.

#### **4.3 Nyelvi tudással rendelkező internetes kereső**

A 2001-es WAC (Web As Corpus) workshopon fogalmazódott meg nyilvánosan először az a gondolat, hogy tulajdonképpen a web maga egy nagy korpusz: a valaha volt és a valamikor létrejövő legnagyobb korpusz.

A web elképesztő tömegű szöveget tárol, amely rendkívül gyorsan nő. A lehető „legdemokratikusabb” médium: a beszélők minden eddiginél szélesebb körét reprezentálja. Annak ellenére, hogy gyakran teljesen bizonytalan eredetű (akár nem anyanyelvi) szövegeket is tartalmaz, bizonyos célokra így is jó, ahogy van (pl. szavak gyakorisági elemzése).

---

<sup>17</sup> Azaz értelmező, jelentésmeghatározó, -kinyerő képességekkel

A web nyelvi korpuszként való kezelése történhet pl. úgy, hogy a Google-t használják a szövegminták begyűjtésére egyfajta menet közbeni (on-the-fly) korpuszkészítéssel és -vizsgálattal egybekötve (pl. WebBootCat, <http://corpora.fi.muni.cz/bootcat> instant korpusz). A web nagysága új szemléletmódot és lehetőségeket is behozhat. Korábban felvetett, egyszerű ötletek éppen e nagy méret miatt már működhetnek: pl. összetett szó fordítása először tagszavanként egyenként, majd gyakoriságvizsgálattal a weben (Grefenstette 1999); párhuzamos szövegek keresése: szokásos link- és HTML-struktúra összevetés (Resnik 1999). Természetesen a web heterogén, nem kiegyensúlyozott, állandóan változó jellege gondot is fog okozni, ennek ellenére várható, hogy a fentiekhez hasonló megoldások az évtized végére elvezetnek egy teljes körű, önálló, nyelvi tudású internetes kereső létrehozásához, amely rendelkezik olyan nyelvi operátorokkal, mint pl. kifejezésre, tagmondatra, szótőre keresés<sup>18</sup>.

#### 4.4 Gépi szövegértés

A szövegek *számítógépes* elemzésével kapcsolatban lehetetlen szót nem ejteni arról, hogy vajon hogyan hasonlítható mindez a szövegek *ember általi* megértéséhez.

Magától értetődőnek tűnik az a megállapítás, hogy az ember igen mélyen képes megérteni a szövegek tartalmát: a szövegben megjelenő utalásokat messzire képes követni, rejtett célzásokat képes „kihámozni” a szövegből, távoli szövegrészek közti kapcsolatot képes felismerni, és mindenek előtt „értelmet” tud tulajdonítani egyszerű szimbólumsorozatoknak, azaz kapcsolatba tudja hozni tudatának mindenkori tartalmával.

Minderre a számítógép tudvalevőleg nem képes. Mindannyian nap mint nap tapasztaljuk, hogy a mindenféle – szinte hihetetlen – képességeik ellenére mégis mennyire „buták” a számítógépek: megakadnak a feldolgozásban, amikor pedig a szükséges információ – számunkra nyilvánvaló módon – rendelkezésükre áll, viszont akkor folytatják a feldolgozást – és jutnak téves eredményre –, amikor tudnivaló, hogy valami lényeges információt még nem kaptak meg.

Egy másik fontos megállapítás a „megértéssel” kapcsolatban, hogy a szövegeket az egyes emberek – kicsit vagy nagyon, de – másként értelmezik. Folyamatosan szembesülünk azzal, hogy nincs egyetlen, abszolút értelmezése semmilyen szövegnek, de legalábbis kisebb-nagyobb bizonytalanságok és különbségek vannak a szövegek értelmezésében. A szövegértelmezés azonban nemcsak az értelmezést végző személy saját és sajátos ismereteitől és tapasztalataitól függ, de függ – sőt, talán elsősorban – azoktól a kultúráktól is, amelyben a szövegértelmező személy felnevelkedett, vagy amelyben él és dolgozik. Nagyon sok mindent (személyes és kulturális háttér) kell tehát tudni a szövegek megértéséhez<sup>19</sup>.

Mégis, el lehet-e valahogy képzelni, hogy mi, emberek *milyen módon* vagyunk képesek arra, amit általában úgy szoktunk nevezni, hogy „megértés”? Erre azért fontos válaszolni, mert, ha igen, akkor talán egy ilyen elképzeléshez hasonló módon működő rendszert létre is lehet hozni. Ha nem, akkor viszont tényleg – legalábbis a belátható időben – áthidalhatatlan a távolság ember és számítógép között ebben a vonatkozásban is.

Ehhez érdemes magunkba nézni és megvizsgálni, hogy mire is gondolunk, amikor azt mondjuk, hogy valamit „megértettünk”. Általában az készlet erre minket, amikor sikerül „hozzákötni” szövegbeli (vagy hallott) információkat korábbi ismeretekhez, tapasztalatokhoz, eseményekhez vagy egyszerűbb, elfogadott tényekhez. Amíg nem történik meg ez a

---

<sup>18</sup> A távolabbi jövőképebe tartozik viszont az olyan gépi fordító, ami automatikusan épít szótárat a webről, és így tölti ki a saját szótárának hiányait – fordítás előtt.

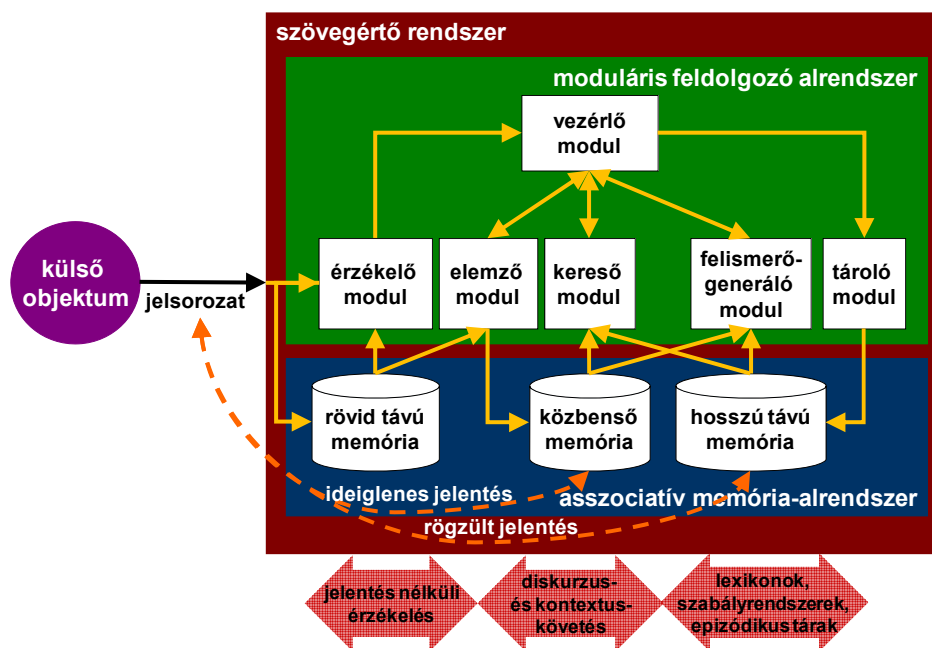
<sup>19</sup> Hogy a szövegértési képesség mennyire nem magától értetődő még az embereknél sem, mutatják azok a nemzetközi felmérések is, amelyeket összehasonlítási célból ma már a magyar iskolákban is elvégeznek. Ezek egyik fő megállapítása, hogy a magyar oktatási rendszer nem támogatja kellőképpen a magas szintű szövegértési képesség kialakulását a diákokban. Ennek javítása érdekében hoztak létre egy ún. szövegértési és –alkotási projektet, amelynek eredményei (új szemléletű tankönyvek) azonban még csak kipróbálási fázisban vannak.

*hozzákapcsolás* – vagy azért, mert nincs mihez, mert annyira újszerű az információ, vagy azért, mert még nem találtuk meg azt elménkben, amihez köthetnénk, vagy “csak” az odavezető utat nem –, addig nem igazán szoktuk azt mondani, hogy értünk valamit. Érdeemes ezzel összefüggésben arra is figyelni, hogy az emberek ősidőktől kezdve az érzékszervi észlelés útján létrejövő belső kapcsolatteremtést a “megértés”, a “felfogás” és a “tapasztalat” előfeltételének tekintették<sup>20</sup>.

Számítógépes rendszerek esetében valami hasonló történik akkor, amikor a szövegben elkülönített információkat összeköti a rendszer egy taxonómia vagy valamilyen ontológia fogalmával vagy egy adatbázisban szereplő objektummal (azonosító-, névegyezés, névhasználat, hasonló tulajdonságok, kapcsolatok stb. alapján). És ha ez kellő bizonyossággal megtehető egy szöveg tartalmát alkotó szavakkal, szó szerkezetekkel, mondatokkal ill. egyéb más jelentésegységekkel, akkor talán joggal jelenthető ki, hogy a számítógépes rendszer “megértette” a szöveget. Persze ez a megértési szint jóval alacsonyabb, mint amit, az embernél megszoktunk, és természetesen egy kicsit más jellegű is, de maximum ez az, amire a “gépi megértés” kifejezést használni célszerű – legalábbis az eljövendő évtized során.

#### 4.4.1 A gépi szövegértés referenciarendszere

Ez a fejezet egy olyan hipotetikus rendszer felépítését igyekszik bemutatni (8. ábra), amely a fentiekben mondottakkal összhangban és azok értelmében képes lehet valamilyen minimális szintű szövegértésre. Várható, hogy 10-15 éven belül ilyen architektúrára épülő szövegértő rendszerek jelennek meg kereskedelmi forgalomban is.



8. ábra: Egy „szövegértő” referenciarendszer elvi felépítése

Egy “szövegértő” rendszer működésében az előzőekben említettek miatt alapvető fontosságú az érzékelt és a memóriában valamilyen formában ábrázolt és tárolt fogalmak közötti *kapcsolatteremtési* képesség (kereső és felismerő-generáló modul). A már stabilan

<sup>20</sup> Erre utal az is, hogy a magyarban e szavak eredete is a „hozzáér”, a „megfog” ill. a „tapogat” szavakra vezetődik vissza, azaz, hogy valamilyen hasonlóság van a között, ahogy a valóságban megérintjük, megfogjuk, letapogatjuk ill. ahogy, általában érzékeljük a dolgokat (vagyis, hogy fizikai, érzékelési kapcsolatba kerülünk velük), és a között, ahogy ésszel felfogjuk, tudatunkba emeljük, „Jelki szemünk előtt látjuk”, azaz végső soron és valójában „megértjük” ezeket a dolgokat. Ilyen összefüggések más nyelvekben is fennállnak (vö. a latin ‘videre’ – látni –, és a német ‘wissen’ – tudni – igék).

rendszerezett (hosszú távú) és a még csak ideiglenesen feldolgozott (közbenső) memóriák tartalma közötti *kapcsolatok* felismerése, előkeresése ill. új kapcsolatok létesítése így egy gépi rendszer megértő képességének alapfeltétele.

Az *érzékelő modul* a bejövő (észlelt) jelek (hangok, karakterek stb.) felismerésével és sorokba (értelmezhető szimbólumokba) rendezésével foglalkozik. A bejövő jelek egy rövid távú memóriába kerülnek, és ott maradnak legalább addig, amíg valamilyen szimbólumhoz nem lesznek kötve. Már itt említést kell tenni a *vezérlő modulról*, amely mindegyik modulból elérhető, és például ebben az esetben a jelek képeinek *hosszú távú memóriában* tárolt példányai előkeresését ill. az azokkal való összehasonlítást vezérli a *kereső modullal* együttműködve.

Az *elemző modul* a szimbólumok és sorozataik „jelentésének” meghatározását végzi, azaz egy ún. *közbenső memóriában* eltárolja, és a *kereső modul* segítségével felidézi azokat a memóriatartalmakat, amelyek közvetlen vagy közvetettebb módon, de kapcsolódnak az adott szimbólumhoz. A *felismerő-generáló modul* mintákat (szabályosságokat) ismer fel a szimbólumsorozatok között, és újabb kapcsolatokat, szabályokat következtet ki közöttük, ha ez lehetséges. A *közbenső memória* rendkívül fontos, mert egy beszélgetés vagy egy hosszabb szöveg során előfordult összes lényeges szimbólum, ezek konkrét, adott kontextusbeli jelentése, szerepe mintegy valamilyen magas szintű „munkamemóriában” tárolódik, és jóval közvetlenebbül és így gyorsabban elérhető a beszélgetés vagy szövegolvasás során.

A *közbenső memória* azon elemeit, amelyek jól illeszkednek a *hosszú távú memória* mindenkori tartalmához, esetleg különös hangsúlyt (megerősítést) kapnak vagy ismételten előfordulnak, a tároló *modul* felismeri és beépíti a hosszú távú memóriába. A *tároló modulon* keresztül válik az egész rendszer rendkívül dinamikussá, mert egyrészt ilyen módon tanulást valósít meg, másrészt állandó változtatásokat végez a *hosszú távú memóriában*, és ezzel a rendszer képes összhangban tartani a külső objektumokban (állapotában, tulajdonságaiban stb.) bekövetkező változásokkal a belső memóriát.

## 5. Befolyásoló tényezők

### 5.1 Technológia

Bárhova is nézünk a szövegelemző technikák terén, mindenütt rendkívül számítás- és táríge-nyes feladatokba ütközünk. A számítási és tárolókapacitások eddigi növekedési ütemének fennmaradása ezért előfeltétel a szövegelemzés alkalmazásának kiteljesedéséhez. A növekedési ütem csökkenése negatív hatással lenne a terület fejlődésére.

Az *OCR-technológia* (Optical Character Recognition) és a *beszédfelismerési technológiák* további fejlődése és terjedése, és az ettől függetlenül kibontakozó *Web 2.0*<sup>21</sup> újabb nagyságrenddel növelheti az elemzésre váró szövegek körét, és új kérdéseket vehet fel: például miről szólnak még egy-egy témához kapcsolódó blogbejegyzések? Ezen belül a *webes kereséssel* szembeni növekvő felhasználói elvárások (pontosabb találati lista<sup>22</sup>) önmagában is elősegítik a nyelvtechnológia fejlődését. Számos technológia mutat tehát ugyanabba az irányba, nevezetesen hogy a feldolgozandó *szövegek mennyisége* továbbra is jelentősen növekedni fog, és ez önmagában nagy nyomást fog kifejteni a szövegelemzési technikák fejlődésére.

Emellett jelentkezik az is, hogy a környezetünkben egyre több „intelligensnek” mondott eszköz jelenik meg, amelyeknél egyre inkább elvárás lesz a beszéddel, kézírással vagy más módon bevitt rövid szövegek, utasítások feldolgozási képessége. Az ember-gép

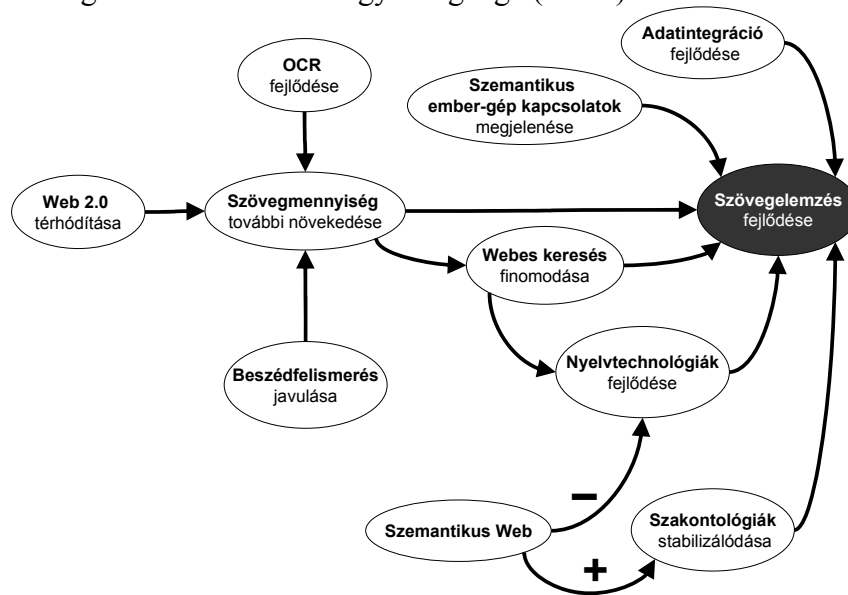
---

<sup>21</sup> A tipikus internetfelhasználó egyben tartalom-előállítóvá is válik. Lásd még az NHIT Információs Társadalom Technológiai Távlatai (IT3) projekt keretében 2006-ban közzétett „Web 2.0 (és ami mögötte van)” c. elemzést.

<sup>22</sup> Pl. „kérem azokat az anyagokat, amelyekben valamilyen adott minősítéssel rendelkező szervezetek teljes listája található” vagy „kérem a vállalateladásokról és -felvásárlásokról beszámoló híreket az elmúlt féléből”.

kapcsolatoknak ilyen értelemben vett rugalmassága, *szemantikai szintjének megnövekedése* és finomodása szintén a szövegelemzési technikák elterjedésére és fejlődésére fog kihatni. Nem szabad megfeledkezni azonban egy jelentős negatív hatásról sem. Az ún. *szemantikus web*<sup>23</sup> és más kezdeményezések azt tűzik ki célul, hogy a szövegek eleve ne csak sima, hanem tartalmi címkékkel ellátott szöveggént jelenjenek meg (XML formában vagy elektronikus formaként) ill. más módokon is a tartalomra utaló információk szöveggészítéssel egyidejű megadására ösztönöznek (pl. kategóriabesorolás, szógyakoriság-térkép). **Mindez nem teszi szükségtelemmé a szövegelemzést, de csökkenti ennek során az igényt teljes nyelvi elemzésekre.**

Végül a szövegelemzés fejlődését segítik a háttér-információknak az elemzésbe való becsatlakozását biztosító, fejlett *adatintegrációs technológiák* és *szakontológiaiák* kialakulása. Ez utóbbiak kialakulását és elfogadottságát – paradox módon – éppen a szemantikus web kezdeményezése segíti elő a Web Ontology Language (OWL) létrehozásával és terjesztésével.

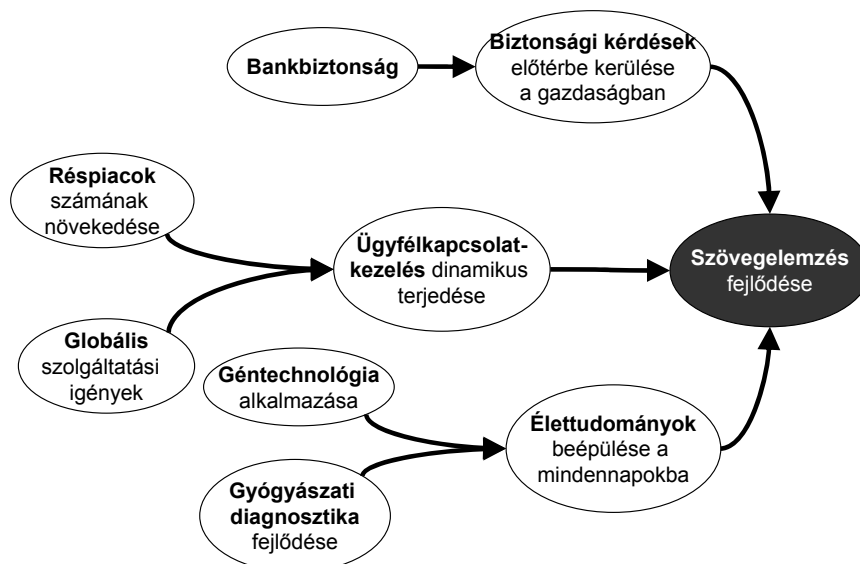


9. ábra: Technológiai tényezők hatása a szövegelemzés fejlődésére

## 5.2 Gazdaság

A gazdasági oldalon a korábban körvonalazott alkalmazási lehetőségek adják a fő hajtóerőt. Ezek közül is elsősorban az *ügyfélkapcsolat* kezelésének finomodása (pl. az ilyen rendszerek által rögzített szöveg- és hanganyagok felhasználási lehetőségének felmerülése), valamint a különböző szintű *biztonsági kérdések* (nemzetbiztonság, rendészet, bankbiztonság stb.) előtérbe kerülése teszik szükségessé a szövegekben rejlő információk minél teljesebb kinyerését.

<sup>23</sup> Lásd még az NHIT Információs Társadalom Technológiai Távlatai (IT3) projekt keretében 2005-ben közzétett „Jelentésalapú technológiák” c. elemzést.



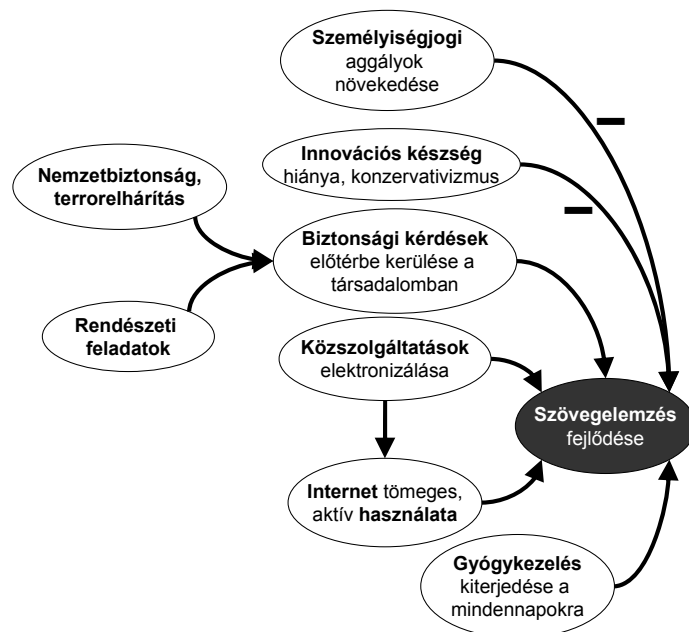
10. ábra: Gazdasági tényezők hatása a szövegelemzés fejlődésére

Negatív hatásként mindenképpen számolni kell azonban a több oldalról is megnövekvő személyiségjogi aggályokkal: szabad-e az emberek által kimondott, leírt szövegeket más célra is felhasználni, mint amire a szerzője szánta. Mivel ennek az egész kérdéskörnek az ellenőrzése – éppen a technológia fejlődése miatt – egyre nehezebbé válik, várható, hogy esetleg radikális tiltása is felmerül a szövegelemzés gazdasági és egyéb felhasználásában.

### 5.3 Társadalom

Társadalmi szinten az *internet tömeges, aktív használata* az a tényező, ami leginkább és legáltalánosabban előtérbe helyezi a szövegelemzési technológiák terjedését, és igénylik azok fejlődését. Egyre több ember kerül kétoldalú kapcsolatba digitális tartalmakkal (fogyasztja és létrehozza), és az elektronikus szövegek elfogadása – függetlenül attól, hogy elektronikus formokban előfeldolgozottan vagy szabad szövegek formájában vannak – általános társadalmi elvárás lesz. Ezt fogja jelentős mértékben elősegíteni a már említett *Web 2.0* közösségi hálózata, a *sávszélesség dinamikus növekedése* és a *közszolgáltatások elektronizálásának megvalósulása* is.

A *gyógykezelési eljárások* várhatólag egyre jobban kiterjednek a mindennapi élet és munkavégzés idejére, ahogy az egészségügyi rendszerben a megelőzés és a kórházi kezelés alternatívái (járóbetegellátás, napos műtétek stb.) egyre nagyobb hangsúlyt kapnak. Mindez növeli a távdiagnosztika, valamint a távolból és közvetetten megvalósuló tanácsadási tevékenységet, és egyben kiterjeszti azoknak a körét, akik ilyen módon gyógyászati célú informatikai rendszerekkel kapcsolatba kerülnek. Reálisan csak az várható, hogy a megnövekedő gyógyászati célú ember-gép kapcsolatok, tranzakciók egyre nagyobb mértékben rá fognak szorulni a beszéd és végső soron a szöveg minél teljesebb gépi feldolgozására.



11. ábra: Társadalmi tényezők hatása a szövegelemzés fejlődésére

Negatív hatást jelenthet ugyanakkor egyes társadalmi rétegekben az *innovációs készség esetleges hiánya*. A szövegelemzés használata ugyanis némi fogékonyságot és toleranciát igényel a használó oldaláról: a szövegelemzéssel kinyerhető információk belátható időn belül csak részlegesek és időnként tévesek lesznek, ezért aktív közreműködésre van szükség a használó részéről, hogy egy-egy adott szituációban a legmegfelelőbb segédeszközöket tudja megkeresni és alkalmazni.

Egy másik negatív hatás tulajdonítható a várhatóan megnövekedő *személyiségjogi aggályoknak*. Az az (egyébként helyes) adatvédelmi elv, hogy a személyes információkat csak az előre meghatározott és ismert célra szabad felhasználni, nyilvánvalóan nehezíti az olyan szövegek automatikus feldolgozását, ahol személyes adatok is találhatóak<sup>24</sup>.

Paradox módon azonban a *biztonság* témaköre általában is a fókuszba kerül, és ez viszont erősíteni fogja az igényt arra, hogy az interneten (pl. blogokban, e-mail-ekben) megjelenő potenciálisan veszélyes tartalmakat és azok készítőit minél előbb azonosítani lehessen. Ez a hatás meghatározó módon függ a társadalom tényleges veszélyérzetétől, és megjelenése csak akkor várható, ha a társadalomra veszélyes emberek, csoportok tevékenysége markánsan megnövekedik.

## 6. Várható hatások

### 6.1 Technológia

#### 6.1.1 Gépi fordítás tökéletesedése

A szövegelemzési technológia azzal, hogy a szövegek tartalmi elemzése során adatbázisokban, adattárházakban tárolt és egyéb adatforrásokból (pl. XML, webszolgáltatás, Excel) elérhető háttér-információkat a nyelvi elemzés számára igény szerint rendelkezésre tudja bocsátani, javíthatja a gépi fordítás eredményességét különösen a nagy változatosságot mutató névelemekkel (név, kód, azonosító jel, mozaikszó stb.) jelölt objektumok (tárgyak, személyek, intézmények stb.) azonosításánál.

<sup>24</sup> Várható ezért, hogy olyan *anonimizálási eljárások* is megjelennek és általánosan beépülnek a szövegelemző rendszerekbe, amelyek kellő biztonsággal ki tudják szűrni a személyes adatokat, és azokat nem valós adatokra le tudják cserélni. Erre példa és iránymutató a gyógyászatban a kórlapok, leletek automatikus statisztikai feldolgozása előtt azok anonimizálására irányuló fejlesztések.

Emellett a *szakterületi* – tehát nemcsak nyelvi – *ontológiák* bevonása a szövegek elemzésébe szintén javíthatja a nyelvi elemzés hatékonyságát.

### **6.1.2 Multimodális adatelemzés elősegítése**

A számítógépes szövegelemzés sikeressége és egyes (nem nyelv-) technológiai eredményei tovaryűrűző hatásként erősíteni fogják más strukturálatlan adatfajták elemzésének, kezelésének és integrációjának technológiáit.

Olyan multimodális információs rendszerek kialakulását és gyakorlati elterjedését vetíti ez előre, amelyekben szöveg, hang (beszéd és zene), álló- és mozgókép valamint a strukturált adatok elemzése összehangoltan és egymást erősítve történik.

## **6.2 Gazdaság**

### **6.2.1 Teljes körű információkezelés felé**

A gazdaság működésében egyre nagyobb szerepet kap az információkezelés minősége és hatékonysága, a vállalati információellátási ciklus eredményessége. Ma a strukturálatlan adatok üzleti folyamatokban történő használata és nem kellő szintű feldolgozása között súlyos ellentmondás húzódik, amely gátat szab az információkezelés további terjedésének vállalati szinten.

A számítógépes szövegelemzés lehetőséget teremt, hogy a vállalatok információkezelési tevékenységüket kiterjesszék e strukturálatlan adattömeg egyik legfontosabb fajtájára, az emberek közti kommunikáció természetes és leghatékonyabb formájára a beszédre és írásos dokumentumokra.

## **6.3 Társadalom**

### **6.3.1 Papírmunka további visszaszorulása**

A számítógépes szövegelemzés (keresés) a dokumentumkezelő rendszerekkel (tárolás), a digitális aláírással (hitelesítés) és az OCR-rel (digitalizálás) együtt a technológiák olyan csoportját képezi, amely már alkalmas arra, hogy kiváltsa a papírdokumentumot. Jóllehet nem várható még a „papírmentes iroda” rövid időn belüli elterjedése, azonban azzal számolni kell, hogy nemsokára már a közigazgatásban is jórészt mentesül az ügyintézői munka a papírdokumentumok létrehozásának kényszerétől, és ezzel egyidejűleg a felszabaduló időt inkább az ügyfelek érdemi kiszolgálására fordíthatják.

### **6.3.2 Soknyelvűség fenntarthatósága**

Az angol nyelv globális használatát nem túl kockázatos megjósolni, de minél gyorsabban, minél pontosabban és minél finomabban akarjuk gondolatainkat, érzelmi állapotainkat másokkal közölni, úgy látjuk szükségét más, „mélyebben beágyazott” nyelv használatának is (anyanyelv, szaknyelv, csoportnyelv).

Különböző embercsoportok, különböző időszakokban ugyanis mást és mást tartanak fontosnak, más céljaik vannak, a dolgok közötti viszonyok más és más vonatkozásait hangsúlyozzák. Ezek a különbségek pedig megjelennek a szemléletmódjukban, a fogalomkészletükben és a használt nyelvtani (szintaktikai) szerkezetekben. A nyelvek jelentősen eltérnek abban is, hogy mit fejeznek ki, támogatnak meg szintaktikailag (nyelvtani konstrukciókkal), és hol, milyen mértékben hagyatkoznak az emberek kombinatív értelmező – jelentést kikövetkeztető – képességére. Az egyes emberek által birtokolt nyelvi képesség tehát a kultúra terméke, ugyanakkor a kultúra alapját a közösségi nyelv adja.

A társadalom alappilléreit pedig azok a kultúrák alkotják, amelyek szokásrendszere, viselkedési, gondolkodási és nyelvi kifejezési módja sokáig meghatározza a felnövő emberek életét, és felnőttként is keretet ad az életvitelükhöz. De túl ezen a kultúrák sokfélesége és

változatossága a társadalom megújuló képességének is a záloga. A kisebb kultúrák azonban jelentős hátrányban vannak a nagyokkal szemben.

**Az információs társadalom fontos célja kell, legyen ezért, hogy biztosítsa a nyelvek – és ezen keresztül a kultúrák – sokféleségének fenntarthatóságát a digitális korszakban is.**

A szövegelemzés számítógépes támogatásának technológiai keret adnak, amelyekbe az egyes nyelvi rendszerek az eddigi tapasztalatok szerint be tudnak illeszkedni. A számítógépes szövegelemzési lehetőségek kiterjedése egy-egy újabb nyelvre tulajdonképpen az adott nyelv – és kultúra – beemelését is jelenti az információs társadalomba.

### **6.3.3 Hatékonyabb ember-gép kommunikáció**

A technológia fejlődésének, és ennek az informatika fő vonalát alkotó fejlődési tendenciákba való bekapcsolódásának hatásaként egyfajta „mesterséges intelligencia mindenütt” állapot fog kialakulni az ember-gép kapcsolatok területén. Ez nem azt jelenti, hogy a belátható jövő informatikai rendszerei emberi képességekkel rendelkeznek, azaz tényleg elérnék az emberi intelligencia szintjét, hanem azt, hogy viszonylag egyszerű, de életszerű szituációkban egyre több gépi eszköz rendelkezik majd olyan minimális „intelligenciával”, ami a mainál hatékonyabb és értelmesebb kommunikációt tud biztosítani ember és gép között<sup>25</sup>.

A pusztán érzékszervi képességeinkre és ezek minél teljesebb kihasználására alapozó kommunikációs módszer ugyanis kezdi elérni a határait. A számítógépek ma már csodálatos audiovizuális képességekkel rendelkeznek, a tapintási és szaginformációk talán hamarosan szintén bekerülnek majd a „képbe” – legalábbis bizonyos speciális szituációkban. És természetesen az érzékszervi információk feldolgozása tovább gyorsul, és mindez még növelni fogja az ember-gép kommunikáció lehetőségeit.

Azonban ezen az úton áttörés várhatóan már nem érhető el – és nemcsak az ember korlátai miatt, hogy nem tud több információt befogadni egy adott időegység alatt, hogy nem tud több filmet megnézni, több zenét meghallgatni, több mobilhívást megválaszolni, és több időt tölteni az Interneten kereséssel, mint amennyit ma tud. A számítástechnikával lassan elérjük az ember érzékszervi feldolgozási határait, szerencsére azonban vannak még más tartalékok. Az „okos” ember ugyanis több információt „lát meg” az érzékszervi információkban, mint amennyi közvetlenül azokban látszik, valamint több és mélyebb ismeretet tud leszűrni ezekből, mert jobban össze tudja kapcsolni korábbi ismereteivel. Különösen, sőt hatványozottan igaz ez a beszédre és általában a szöveges információk befogadására.

Amit tehát a számítógépes szövegelemzés fejlődése előtérbe állít az a „mentális” kommunikáció lehetősége a rendszerekkel. Ezen az értendő, hogy ahhoz, hogy egy adott – esetleg nagyon leszűkített – szituációban hatékony kommunikációt lehessen folytatni egy rendszerrel (pl. sok átfogó információ időegység alatt; közvetett és részletes információk igény szerint) az kell, hogy magának a rendszernek is legyen egy megfelelő modellje az adott szituációról (időpont, témakör, előtörténet stb.). Pontosan ezek azok a kontextust leíró (meta)információk, amelyek a szövegelemzést is segítik abban, hogy minél több és relevánsabb információt tudjon a számítógép a szövegekből kinyerni.

**A számítógépes szövegelemzésben a tudásreprezentációs és következtetési képességek megjelenése és fejlődése azt is elősegíti, hogy az ember-gép kommunikáció a jövőben további áttöréseket tudjon elérni a hatékonyság tekintetében.**

Ha belegondolunk, az emberek is azért „beszélnek el” időnként egymás mellett, mert adott dolgokról egészen másként vélekednek, máshogy képzelik el, mást tételeznek fel, más a természetes számukra: azaz a szituációról (öntudatlanul) alkotott modelljük jelentősen eltér egymástól. A hipotézis az, hogy minél közelebb van a kommunikáló felek „mentális” (tkp.

---

<sup>25</sup> Ez az „intelligencia” bizonyos értelemben folytonos tartományának feltételezését vonja maga után, amelyben helyük van bizonyos „mentális” képességekkel rendelkező mesterségesen alkotott rendszereknek – a tartomány egyik végén –, és persze az embereknek a tartomány másik (nagyon távoli) vége felé.

belső) modellje a szituációról, annál könnyebben és hatékonyabban tudnak kommunikálni. Ezért kell tehát tudásreprezentációs és következtetési képesség. A tudásreprezentáció szintje meghatározza az ilyen belső modellek minőségét és részletezettségét, míg a következtetési képesség a származtatott (közvetett) információk kinyerésének mértékét – vagy egyáltalán lehetőségét.

Gyakori, hogy a beszélgető partnerünk tud valamit, amit mi nem, és ez akadályozza a kommunikációt. Ez a hiányzó információ vagy véletlenül említésre kerül a beszélgetés során (ekkor ezt fel kell ismerni - mintafelismerés), vagy nekünk úgy kell a beszélgetést vezetni, hogy kiderüljön, mi az, ami hiányzik a „képből” (célvezérelt kommunikáció), vagy a saját eddigi ismereteinkből származtatni kell, és a kommunikáció számára közvetlenül használhatóvá (explicitté) kell tenni. Mind a három említett, kommunikáció közbeni mentális tevékenység egy-egy példa a szükséges következtetési képességre.

A mai számítógépes rendszerek zavaró „butasága” is hasonló okokra vezethető vissza: nincs megfelelő (értsd: kellően összetett, részletes és rugalmas) szituációs modelljük. Ennek javításához pedig azok a tudásreprezentációs és következtetési képességek kellene, amelyek fokozatos kialakulása és használatba vétele a szövegek egyre terjedő számítógépes elemzése folyamán fognak stabilizálódni.

## **7. Hazai helyzet**

### **7.1 Jelenlegi helyzet**

#### **7.1.1 Nyelvtechnológia**

A nyelvtechnológiák területén 3 szervezetnek van meghatározó jelentősége: az MTA Nyelvtudományi Intézete (NYTI), a Morphologic Kft. és a Szegedi Tudományegyetem (SZTE) Nyelvtechnológiai Csoportja (NyTCs). Mellettük még fontos szereplőként említendő meg a BME Média Oktatási és Kutató Központ (MOKK), Alkalmazott Logikai Laboratórium Kft. (ALL) és a Signum Kft. Szövegelemzési és szövegbányászati feladatokkal foglalkoznak még a BME Távközlési és Médiainformaticai Tanszékén (TMIT) is.

E szervezetek tevékenységének köszönhetően a lexikai elemzés területén elég jó a helyzet, mert jelenleg van 3 teljes körű morfológiai rendszer (HUMOR – Morphologic Kft., HUNMORPH – BME MOKK, Elekfi-rendszer – MTA NYTI). Ide kapcsolódóan említendő meg a két lexikai adatbázis is: az igei vonzatkeret adatbázis (NYTI) jelenleg kb. 30 ezer igét tartalmaz (esetleírásokkal, szemantikai jegyekkel ellátva), a névszói adatbázis kb. 100 ezer névszót tartalmaz (nyelvtani, szemantikai jegyekkel ellátva).

A szintaktikai elemzés területén – a komplexitásából adódóan érthető módon – kevésbé jó a helyzet, itt még csak részeredményekről lehet beszámolni. A Morphologic a már magyar WordNet-re épülő Metamorpho rendszere kívánkozik első helyre, de már ma is említésre méltó a NYTI névszói és melléknévi szerkezeteket, valamint tagmondatokat felismerő rendszere, továbbá a MOKK HUNPARS rendszere is.

A szintaktikai elemzéshez kapcsolódnak az ún. korpuszok (szintaktikai adatbázisok). A korpuszok alapján gépi tanuló algoritmusok (pl. neuronhálók) segítségével szintaktikai elemzőket vagy névelem-felismerőket (named entity recognizer) lehet fejleszteni. Jelenleg az SZTE készítette a legnagyobb magyar korpuszt (az ún. Szeged korpuszt), amely mintegy 1,2 millió szót tartalmaz, és legújabb 3. változata (treebank) már a mondatok szintaktikai fáit is tartalmazza.

A szemantikai elemzés területén nagy előrelépés várható a magyar WordNet létrejöttétől és szabad hozzáférésétől, valamint a névelemek (pl. tulajdonnevek, mozaikszavak, azonosítók) felismerése terén a HUNNER projektől (MOKK, SZTE, NYTI).

### **7.2 Fejlesztések és várható fejlődés**

### 7.2.1 Magyar WordNet (HUWN)

A magyar WordNet létrehozására irányuló szakmai tevékenység legalább 2000-ig vezethető vissza. Azóta folyik szakmai körökben a feladat előkészítése, a megközelítések, stratégiák kialakítása. 2004 elején még arról kellett beszámolni, hogy nem sikerült állami támogatást kapni egy ilyen projektre, de végülis 2005-ben beindulhatott a magyar WordNet létrehozását célzó projekt, amelynek eredményeképp egy kb. 40 ezer szinonimacsoportot tartalmazó magyar nyelvi adatbázist jött létre a 2007-ben lezárult fejlesztés során. A magyar WordNet 2008-tól várhatóan ingyenesen elérhető lesz az interneten, azonban továbbfejlesztése jelenleg nincs napirenden.

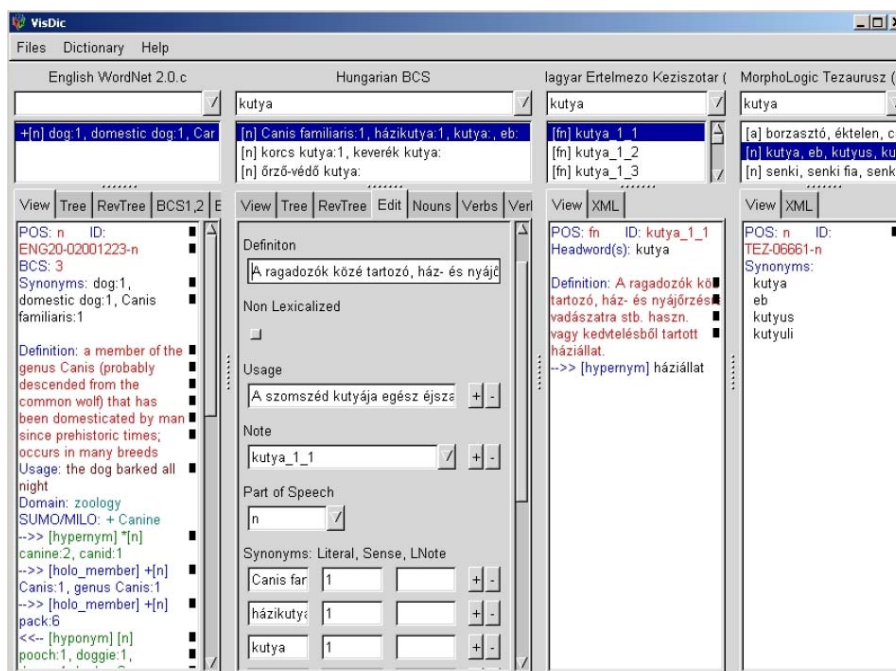
### 7.2.2 Magyar NooJ

A Nyelvtudományi Intézet elkészítette azt a minimális eszköztárat, amellyel a NooJ rendszer magyar nyelv elemzésére is alkalmassá vált. A magyar NooJ nemcsak számítógépes nyelvészeknek való, hanem hasznos eszköz lehet mindenki számára, akinek magyar nyelvű, bármilyen témájú szövegek elemzése a célja.

A magyar NooJ közreadásával az Intézetnek az a célja, hogy támogatást nyújtson a magyar nyelv korszerű technológiával történő, empirikus kutatásához, és hogy egy aktív magyar felhasználói közösség alakuljon ki, akik a jövőben megosztják egymás között eredményeiket. A szótár jelenleg a Magyar Értelmező Kéziszótár mintegy 80 ezer címszavát tartalmazza, ami e szókészlet teljes körű ragozásán keresztül kb. 130 millió szóalak felismerését teszi lehetővé. Az eszköz által nyújtott szótármodulok folyamatos fejlesztés alatt állnak (pl. optimalizálás), és elérhetővé tételük rövidesen megkezdődik. 00

Mindezzel az Intézet egy olyan program elindulásában bízik, amelynek keretében a magyar nyelvre különböző, cél- és kontextusspecifikus nyelvtanok alakulhatnak ki. Ennek eredményeként pedig a magyar nyelvű szövegek korlátozott, de kontextusra optimalizált és hatékony számítógépes nyelvi elemzése kézzel fogható közelségbe kerülhet.

A NooJ-alapú nyelvtani elemzők nem teszik szükségtelessé egy általános magyar szintaktikai elemző elkészítését, azonban konkrét igényeknek jól megfelelő, hatékony alternatívát, kompromisszumos megoldást nyújthatnak.



12. ábra: A magyar WordNet készítéséhez használt erőforrások

### 7.2.3 Egészségügyi célú szövegelemzés

### 7.2.3.1 Automatikus kórlapkitöltés

Az Alkalmazott Logikai Laboratóriumban folyó kutatás célja egy kórlapkitöltő rendszer készítése, amely a kórházi panaszfelvételek során lejegyzett magyar nyelvű szövegeket egységes kórlap-reprezentációvá alakítja. A kórlapnak tartalmaznia kell a beteg adatait, a rá vonatkozó panaszfelvételek körülményeit, és az egyes panaszfelvételek során leírt tüneteket, panaszokat, azok tulajdonságait, fellépésük és megszűnésük idejét, ill. az esetleg fontos egyéb körülményeket.

A szöveg nyelvi előfeldolgozását a MorphoLogic MetaMorpho nevű morfológiai és szintaktikai elemzője végzi, amely a szintaktikailag elemzett szöveget XML formátumú elemzési fa formájában adja vissza.

Az alkalmazás keretében kezelt orvosi szövegek esetében különös jelentősége van a szövegnormalizálásnak. Kezelni kell a szakterületre jellemző idegen (főképp latin) szavakat, a rövidítéseket ill. ezek legkülönfélébb változatait, a számokat, és a szöveg sietős lejegyzése miatt feltűnően gyakori hibákat, elgépeléseket.

A rendszer tartalmazza a nyelvi előfeldolgozót (normalizálás, morfológiai és szintaktikai elemzés), a jelentésreprezentációt létrehozó szemantikai elemzést, és a reprezentációból a kórlap egyes mezőinek megfelelő információt kiolvasó kórlapkitöltőt.

A jelenleg működő demonstrációs változatnak még számos korlátja van. A továbblépéshez az szükséges, hogy a szintaktikai és szemantikai elemzés ne egymás után, hanem egymással párhuzamosan történjen. Ehhez alkalmas nyelvtan kidolgozása szükséges. Ez a megoldás javíthatná a többértelműségek feloldását is. A szemantikai elemzésbe jelenleg procedurálisan beépített szabályok helyett leíró szabályokat célszerű használni.

### 7.2.3.2 Leletek automatikus címkézése

A számítógépes ontológiák, valamint strukturált szótárak fejlesztése terén tapasztalható fejlődés ellenére a legtöbb kórházban az adatok tárolása részben továbbra is szöveges formában történik. Szövegelemzési (pl. nyelvtechnológiai) eszközökkel azonban lehetséges e nagy mennyiségű, szöveges adatban a rejtett struktúra felfedezése, és a tárolt információ elérhetővé tétele.

Az Egyesült Államokban a kórházi ellátás során elvégzett vizsgálatok, kezelések dokumentálására az ICD-9-CM (International Classification of Diseases, 9th Revision, Clinical Modification) kódjait használják<sup>26</sup>. Az egészségbiztosítók kifizetései a klinikai dokumentumokhoz rendelt ICD-kódok alapján történnek, amelyeket a kezelés után, utólag rendelnek a kórházi dokumentációhoz.

A leletek automatikus osztályozását lehetővé tevő számítógépes eljárások fejlesztése gazdasági szempontból azért alapvető fontosságú, mert orvosi témájú szöveges dokumentumok kódolására, illetve az ennek során keletkező hibák javítására pl. évi mintegy 25 milliárd dollárt fordítanak az Egyesült Államokban.

A Szegedi Tudományegyetem Mesterséges Intelligencia Kutatócsoportja amerikai kórházak és kutatóintézetek által 2007-ben rendezett verseny kapcsán készített el egy, angol nyelvű leletek automatikus címkézését végző rendszert. A rendszer fő jellegzetessége, hogy kombinálja a szabályalapú kódolást a statisztikaival, azaz a gépi tanulási modell felhasználásával. A pusztán szabályalapú megközelítéssel kb. 83%-os pontosságot tudtak elérni, míg a statisztikai elemzés kiegészítő használatával ezt sikerült kb. 89%-ra növelni. Ezzel az eredménnyel a versenyben 2. helyezést értek el.

A versenyen egyébként 50 résztvevő indult, és a beküldött megoldások 15%-89% közötti pontosságot mutattak, 77%-os átlagos érték mellett. Összesen 21 rendszer ért el 80% feletti eredményt, azaz ért el olyan pontosságot, amely – korlátozott mértékben – összevethető az

---

<sup>26</sup> Magyarországon hasonló módon az ICD-nek megfelelő ún. BNO (Betegségek Nemzetközi Osztályozása) kódokat használják.

emberi annotáció pontosságával. A legjobb megoldások (így a szegedi rendszer is) az emberi címkézés pontosságát közelítik, azaz az eredmények azt mutatják, hogy valóban reális célkitűzés a klinikai szövegek elfogadható pontosságú gépi feldolgozása.

A szegedi megoldás abból a szempontból különösen előremutató, hogy versenyben rendelkezésre bocsátott kódolási útmutatókból többé-kevésbé *automatikus* konverzióval állították elő a kezdeti szabályrendszert, majd ezt címkézett adatok felhasználásával gépi tanulási modellekkel *teljesen automatikusan* fejlesztették tovább. Ez a megközelítés biztosítja, hogy a rendszert az eredmények érezhető romlása nélkül a versenyben használnál jelentősen nagyobb számú kódra is adaptálni, skálázni lehessen. Kijelenthető tehát, hogy az SZTE szabályalapú és a gépi tanulási modellek kombinációján alapuló, valamint nagy mértékben automatizálható megoldása versenyképes eredmény tud elérni a klinikai dokumentumok osztályozásában.

A kutatás-fejlesztés távlati célját ezek alapján egy olyan hazai alkalmazás jelentheti, amely magyar nyelvű klinikai szövegek, nagy számú kód alapján történő gépi osztályozása során 90%-t stabilan meghaladó pontosságot tud elérni.

#### **7.2.4 Pszichológiai szempontú szövegelemzés**

A Pécsi Tudományegyetem Pszichológiai Intézetében folyó kutatás célja a szövegek automatikus elemzésének kidolgozása különböző pszichológiai dimenziók mentén. A kutatás konkrétan egy olyan *szövegelemző program* kialakítására irányul, amely képes automatikus azonosítására a *történeti jellegű* szövegekben kifejezésre kerülő:

- **személy- és csoportértékeléseknek;**
- **szubjektív időélménynek;**
- **szándékoknak (intencionalitás);**
- **érzelmi vonatkozásoknak;**
- **okási kapcsolatoknak;**
- **személyek egymásközi viszonyában bekövetkező változásoknak;**
- **belső tudattartalmaknak;**
- **a történésekhez való aktív-passzív viszonyoknak;**

A szövegelemző programot a NooJ-rendszer segítségével fejlesztik. Az általános megközelítés az, hogy a történeti szövegek minden egyes (fentiekben felsorolt) tartalmi vonatkozásához a NooJ-rendszeren keresztül egy-egy ún. lokális nyelvtant fejlesztenek ki, amelynek segítségével tudják a szövegből a kívánt információt kinyerni.

A kutatás 2007 folyamán a Magyarországon használt történelemkönyvek narratív pszichológiai elemzésével foglalkozott, amelyben differenciált kép kialakítására törekedtek a magyar történelemábrázolásban előforduló érzelmekről (barát-ellenség), a nemzeti identitás alakulásáról (empátia, aktivitás-passzivitás), történelmi alakokról és korszakokról stb. azaz általában a történelem szociális reprezentációjáról.

A kutatási projekt egyik nagy kihívása a különböző szempontú és egyelőre külön ágakon futó elemzések eredményeinek integrálása, és ilyen módon integrált következtetések levonása. A másik ígéretes irány a környező országokban hasonló kutatások indukálása lehet, hiszen ennek megvalósulása esetén objektív összehasonlító elemzésre nyílna lehetőség a hazai ill. külföldi történelemszemléletek között.

#### **7.2.5 Egyéb fejlesztések**

##### **7.2.5.1 Nyelvfüggetlen tulajdonnév-felismerő rendszer**

A Szegedi Tudományegyetem Informatikai Tanszékén folyó fejlesztés eredményeként egy számos alkalmazásban kiemelkedő pontosságot elérő statisztikai tulajdonnév-felismerő rendszer kialakítása van folyamatban.

A tulajdonnevek azonosítása (és kategorizálása) folyó szövegben meghatározó fontosságú a számítógépes szövegelemzés során. Az információkinyerés esetén a tulajdonnevek általában jelentős információhordozó szerepet töltenek be a szövegben.

A rendszert eddig három merőben eltérő feladaton vizsgálták meg:

magyar nyelvű gazdasági szövegek feldolgozása a Szeged Treebank korpusz gazdasági rövidhíreit használva;

angol nyelvű újsághírekben (sport, politika, gazdaság) szereplő tulajdonnevek felismerése egy adatbázis alapján;

angol nyelvű orvosi zárójelentések anonimizálása (a páciensek, doktorok, kórházak stb. neveinek felismerése és véletlenszerű azonosítókkal való lecserélése).

A rendszer – elsősorban az összegyűjtött nagyméretű tulajdonsághalmaz, illetve az abban rejlő lehetőségek hatékony kiaknázásának köszönhetően – több összehasonlításban is versenyképesnek bizonyult a hasonló problémákra ismert legjobb módszerekkel.

### 7.2.5.2 Szövegosztályozási alkalmazások

A BME Távközlési és Médiainformatikai Tanszékén már 2002 óta foglalkoznak a szövegelemzés egyik klasszikus feladatával, a szöveg témájának meghatározásával. A kutatások során kifejlesztettek egy olyan kategorizáló eljárást, amely igen hatékonyan képes hierarchikus kategóriarendszerbe (taxonómiába) való osztályozásra. A módszert eredményesen alkalmazták többek közt:

- szabadalmi szövegek automatikus és félautomatikus osztályozására;
- néhány szavas internetes keresőkifejezések kontextusának meghatározására, amellyel a keresés témaköre leszűkíthető;
- híryananyagok tematikus besorolására;
- valamint más módszerekkel kombinálva numerikus (azaz nem szöveges) objektumok osztályozására, amelyet orvosi diagnosztizálást segítő döntéstámogatásnál alkalmaztak.

### 7.2.5.3 Magyar, internetes, gazdasági témájú tartalmak keresése

Szintén a BME TMIT-en folyó fejlesztés célja egy olyan keresőszolgáltatás kiépítése, amely az interneten magyar nyelven elérhető gazdasági témájú tartalmak lehető legteljesebb körét egy helyen kereshetővé, és – amennyiben a tartalomszolgáltató ill. jogtulajdonos részéről ennek nincs akadálya – elérhetővé is teszi. A keresőszolgáltatás tematikus, valamint szemantikus elveket és újfajta vizualizációt alkalmaz.

A szolgáltatás keresési funkciói a következők:

- **Szabadszavas kérdés-vezérelt keresés;**
- **Mintadokumentum alapú keresés;**
- **Tematikus böngészési lehetőség rögzített témastruktúrában;**
- **Keresésfinomítási lehetőség a találatok kulcsszavai alapján.**
- **Ugyancsak az internetes kereséssel kapcsolatos a TMIT-en az internetes adatbázisok tartalmában, a „mélyhálóban” való keresés technológiáját megalapozó kutatás, amely lehetővé teszi, hogy a keresőnek természetes nyelven (magyarul) adjon a felhasználó keresőkifejezéseket.**

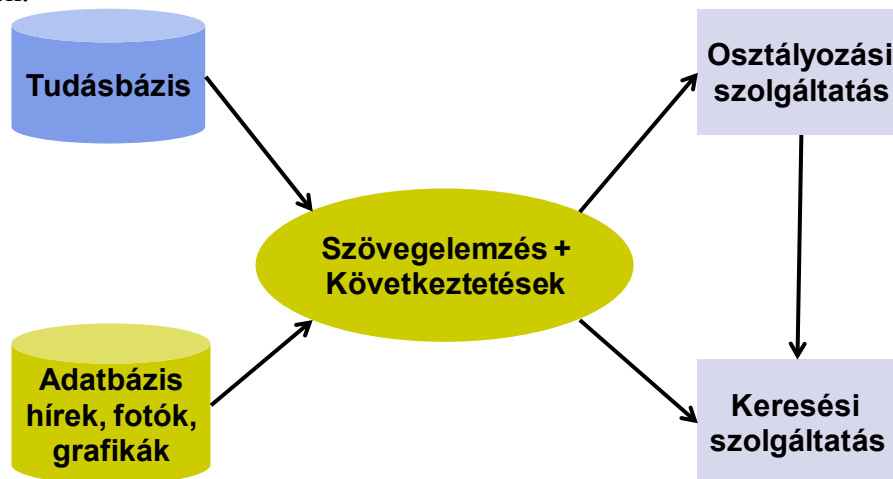
### 7.2.5.4 Dokumentumok automatikus kódolása tudásbázis alkalmazásával

A nagy mennyiségű dokumentumokat tartalmazó rendszerek létrehozása az esetek döntő többségében – a megőrzés mellett –, szolgáltatási, visszakeresési céllal jön létre. Egyre fontosabb, hogy a visszakeresés során a különböző típusú – szöveg, kép, grafika – dokumentumok egységes kezelése megvalósuljon. A mai technika természetesen ezekben az esetekben a képekhez, grafikákhoz kapcsolódó szövegek kezelésére és nem a látványra terjed ki.

A szövegfeldolgozás – a dokumentumok megfelelő számú és tartalmú metaadattal történő ellátása –, nem öncélúan, hanem az eredményes visszakeresés céljából készül. Ez azt jelenti, hogy a keresett dokumentumok – nem több és nem kevesebb – egyszerű lekérdezéssel megtalálhatóak és/vagy szolgáltatások erre a képességre alapozhatóak.

Az MTI Zrt-ben az Ovitás kft-vel együttműködve kialakításra került egy tudásbázison alapuló automatikus kódoló rendszer, amely a többszintű szűkítő keresést is támogatja.

Az automatikus kódolás előnye: az emberi – szubjektív – tényező a kódok elhelyezésénél elmarad, így az egységessé válik, jelentős kódolói kapacitás megtakarítható. Olyan tartalmilag megbízható szolgáltatások alakíthatóak ki, amelyek emberi kódolás esetén nagy „zajjal” működnének.



13. ábra: A szövegelemzés beépülése az MTI-szolgáltatásaiba

A technika alkalmazásának hátránya azonban, hogy jelentős szellemi kapacitást igényel a tudásbázis kiépítése, karbantartása. Az ilyen rendszerek üzleti megtérülése a többszörös használat és a sokirányú ráépült szolgáltatásokkal érhető el.

A tudásbázis alapja egy háromszintű hierarchikus, tematikailag szinte minden területet lefedő mintegy 1800 kódot tartalmazó rendszer, amelyhez strukturált hívófogalmak kapcsolódnak – intézmények, funkciók, személyek kötött hármasa; események. Az intézményekhez, eseményekhez földrajzi kódok is tartoznak. Mind a kódrendszerhez közvetlenül, mind az intézményekhez, eseményekhez olyan fogalmak kapcsolhatók eltérő erősséggel, amelyek jellemzőek rájuk és a szövegben előfordulva nagy valószínűséggel az adott kódra vagy szervezetre lehet következtetni.

Az automatikus kódoló a dokumentumok szövegében „felismeri” illetve a szövegben található fogalmak alapján „következtet” a tudásbázisban megtalálható személyekre, intézményekre, eseményekre, földrajzi helyekre és a tartalmilag illeszkedő kódokra. Ezeket a szöveg, kép, grafika rekordjaira „elhelyezi”, de mindig csak a legalsó szintet.

A tudásbázis tartalmaz szinonimákat, felismeri a rövidítéseket – másképpen kezelve azokat, amelyek több jelentésűek –, „megkülönbözteti” az azonos nevű, de különböző személyeket, és a félreértésre okot adó földrajzi kódokat, intézményneveket is figyelme veszi. A tudásbázis „ismeri” az idő fogalmát, amely lehetővé teszi, hogy a szövegben felismert személy esetén „tudja”, hogy az időponttól függően milyen funkció, intézmény kódját tegye az adott dokumentumra. Követi az intézmények, országok „életútját”, azaz a szövegeket az adott időszakra jellemzően kódolja.

A kódolás folyamatában a tudásbázist használja hasonlóan a kereséshez. A keresés során a rendszer természetesen a rekordot nemcsak a ráakott legalsó szintű kódokra, hanem az összes felette található kódra rákérdezve is megtalálja.

A keresés sajátossága, hogy a megtalált halmaz több szempont szerint – az adott halmazban található földrajzi helyek, témakörök, személyek, intézmények, események, valamint az idő és

a dokumentum formája szerint is szűkíthető. A halmazban található kódokat a kereső előfordulási relevancia szerint ajánlja fel.

A rendszer a fentiekkel összhangban képes pl. a kiadott hírek, grafikák és fotók egyedi tartalmának profil szerinti, valós idejű szolgáltatására.

### 7.3 Befolyásoló tényezők és hatások

A számítógépes szövegelemzés hazai elterjedésében meghatározó szerepet játszik azoknak a nyelvspecifikus technikáknak a megléte és széles körű rendelkezésre állása, amelyek elkészítése elsősorban a hazai nyelvész és informatikus szakembergárda feladata lehet. Ilyen technikák pl. az általános és specifikus magyar nyelvi elemzők, a magyar nyelvi ontológiák, a jó minőségű nyelvi annotációt tartalmazó, magyar szövegtörzsek és a szakontológiák magyar nyelvi változatai.

A hazai kis és középvállalkozások sem anyagi lehetőségeiknél, sem innovációs készségükönél fogva nem alkalmasak az ezekhez szükséges fejlesztések finanszírozására. A hazai nagyvállalatok túlnyomó többsége multinacionális, ezért nem különösebben érdekelt a magyar nyelv számítógépes technológiáinak kidolgozásában, bár mint kísérleti helyszín a fejlesztés egyes szakaszaiban reálisan bevonhatók. A hazai tulajdonú nagyvállalatok többnyire el vannak foglalva multinacionális versenytársaikkal folytatott piaci küzdelemmel. Sok függ tehát a mindenkori magyar kormányzat előrelátó képességén, azon, hogy tud-e és akar-e olyan finanszírozási kereteket teremteni, amelyek biztosítják az említett fejlesztési feladatok végrehajtását méghozzá olyan időkeretek között, amely nem növeli, hanem csökkenteni a magyar nyelv „digitális” távolságát a vezető európai és a hasonló helyzetben lévő regionális nyelvektől.

Ha a szükséges fejlesztéseket sikerül időben elvégezni, akkor olyan *nyílt forráskódú, de legalábbis alacsony költségű technológiák* állhatnak elő a magyar nyelvhez, amelyek nagyban hozzájárulnának és jelentősen felgyorsítanák a vállalatok és intézmények információgazdálkodási rendszereinek kiterjesztését a strukturálatlan információkra. A hazai vállalatok és intézmények sem maradnának így le pl. az ügyfelekkel folytatott kommunikáció hatékonyságát és eredményességét illetően, de a hazai egészségügyi, bűnüldöző és igazságszolgáltató szerveknek sem kellene a magyarul készült dokumentumok (zárójelentések, vallomások, beadványok, periratok stb.) számítógéppel segített elemzését mellőzniük.

## 8. Összegzés

A gazdaságban és közigazgatásban naponta keletkező információk túlnyomó többsége ma is strukturálatlan (szöveg, hang, kép), amelynek jelentős része szöveges. Az információtechnológia eddigi és előrevetíthető fejlődési üteme mellett nagyon valószínű, hogy ezen strukturálatlan információk feldolgozása és célzott elemzése belátható időn belül szervesen beépül a vállalatok és intézmények mindennapos információgazdálkodási tevékenységébe. A számítógépes szövegelemzés szerteágazó technológiai bázisra épül. Nem korlátozódik a nyelvtechnológiákra, bár ezek kétségtelenül rendkívül fontos részét képezik. Folyamatosan jelennek meg újabb kísérletek nyelvi elemzés nélküli, vagy – kompromisszumos megoldásként – részleges nyelvi elemzést használó megoldásokkal. Ezek a megközelítések többnyire a szógyakoróság elemzésén ill. szakontológiákra való illeszkedés vizsgálatán alapulnak.

A várható fejlődést alapvetően az említett beintegrálódás fogja jellemezni. Más oldalról a szövegelemzést alkotó technikák kapcsolata a különböző szakterületekkel, nyelvekkel egyre modulárisabbá, szabványosabbá válik. Fontos mérföldkő lesz a már nyelvi képességeket is magába foglaló internetes kereső megjelenése és elterjedése. A hazai nyelvtechnológiai fejlesztések eredményeként már középtávon várható, hogy különböző magyar nyelvi elemzők

kerülnek használatba, amelyek képességei a teljes mondatelemzéstől a részleges mondat-, ill. mondatközi elemzésig terjednek.

A számítógépes szövegelemzés fejlődésének ugyanakkor van egy további – hosszabb távon valószínűleg jelentősebb – hatása: gyakorlati alkalmazásba kerül és várhatóan mindennapjaink részévé válik a szövegek számítógépes értelmezése.

A számítógép azonban máshogy fogja „beszélni” a nyelvet, mint mi emberek. Nem fog rendelkezni azzal az emberi kultúrákban gyökerező kognitív háttértudással, amelynek nyomai – a gép számára zavart okozóan – minduntalan fellelhetők a szövegekben, és amelynek egységes digitális kódolása nem lehetséges, mert éppen egyéni, csoport- és alkalomfüggő jellege és hajlékonysága adja a lényegét. Rendelkezni fog azonban olyan „háttértudással”, amelyet on-line hozzáféréseken keresztül elérhető adatbázisok alkotnak – még hozzá egy-egy területre vonatkozóan átfogó és naprakész információkat szolgáltatni tudó adatbázisok.

### **Köszönetnyilvánítás**

A szerző szeretné kifejezni köszönetét *Tikk Domokos*nak a tanulmány előzetes változatának alapos nyelvi-szakmai lektorálásáért, *Talyigás Judit*nak az MTI dokumentumkódoló rendszerének bemutatásáért és ismertetéséért, valamint a *számítógépes nyelvészettel foglalkozó magyar szakemberek* közösségének az általuk használt és kifejlesztett technológiákról a rendszeres éves szegedi konferenciájukon közzé tett információkért.

