Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék

# Heterogén adatbázisok lekérdezése strukturált nyelvi szerkezetek komplex szemantikai feldolgozásával

Tézisfüzet

**Kardkovács Zsolt Tivadar**

Tudományos vezető:
Dr. Magyar Gábor (BME–TMIT)
Dr. Gajdos Sándor (BME–TMIT)

Budapest, 2007

# 1 Bevezetés

A tipikus természetes nyelvű adatbázis-interfészek (NLIDB) architektúrája négy rétegből áll [3]; a természetes nyelvű kérdést szintaktikailag elemzik, majd szintaktikai elemző eredményét egy ún. szemantikai elemzésnek vetik alá, amely az NLIDB-hez csatolt, ún. támogató adatbázisokra leképezi a szintaktikai elemző által előállított formális mondatot, a szemantikai elemző által létrehozott lekérdezést végrehajtatják a támogató adatbázis-kezelőkkel; végül a kapott eredményt a rendező összefésüli. Az NLIDB központi eleme a szemantikai feldolgozást végző fokozat, amely tehát egy jól definiált formális struktúra és az NLIDB tudásbázisaként rendelkezésre álló támogató adatbázisok között egyfajta hidat képez.

A valós világ formális, egyszerűen és könnyen bővíthető reprezentációja elegendhetetlen feltétel az NLIDB esetében, hiszen a nyelvtanilag helyes, jólformált bemenetek megadásával az eredeti jelentéstartalommal azonos formális, számítógéppel könnyen feldolgozható lekérdezést kell előállítaniuk. Ezt a fajta bővíthetőséget a szakirodalom az NLIDB hordozhatóságának [15, 11, 3], illetve hordozható reprezentációnak nevezi. A hordozható reprezentáció legfőképpen a szemantikai elemzést támogatja, vizsgálataim kizárólag erre a területre korlátozódtak az értekezésemben.

Az eddigi munkák során az NLIDB megvalósítására négy lényegesen különböző paradigmával tettek kísérletet [11, 3]. Az egyes eljárások a nyelvet, a nyelvi reprezentációt helyezik előtérbe, míg a lekérdezendő adatbázist – az egyed-kapcsolati–modellen (ER–modell) [9] alapuló köztes lépcsős megoldások [16, 17] kivételével – mint erőforrást veszik figyelembe annak eldöntésére, hogy a nyelvi reprezentációban szereplő névelemek milyen adatbázisbeli előfordulással, típussal bírnak. Az ER–modell alapú megoldás ugyanakkor meglehetősen érzékeny az adatszerkezetek változtatására, ráadásul a koncepcionális modellezés nem veszi figyelembe az adatbázisoknál tipikus, redudancia csökkentése miatt szükséges sémadekompozíciót sem. Tegyük hozzá, hasonló problémákat rejt egy ontológia alapú köztes lépcső kialakítása is, hiszen a legismertebb ontológiát leíró nyelvek segítségével a világot általában reprezentálhatjuk, és ezeken következtetéseket lehetünk képesek megfogalmazni, ugyanakkor ezek közvetlen, uniform leképezése egészen más koncepció szerint létrehozott adatbázis-szerkezetekre, illetve a relációkon értelmezett műveletsorozatokra szinte lehetetlen vállalkozás.

Az NLIDB hatékonyságának növelése érdekében az adatbázisban tárolt világmodell közvetlen felhasználása, illetve kifejezetten relációs adatmodellek [10] segítségével leírt ontológiaszerű adatbázis-technológia kidolgozása kívánatos. A relációs adatmodellen alapuló megoldás azt garantálja, az „ontológia” kellően egyszerű lehet, ráadásul a rajta értelmezett műveletek közvetlenül alkalmazhatóak az NLIDB-t támogató adatbázisok adatszerkezeteire.

Értekezésemben arra keresem a választ, hogy hogyan lehet kizárólag a relációs adatmodell eszközeivel olyan kompakt reprezentációját kialakítani az NLIDB által átfogott világnak, amely egyfelől közvetlenül alkalmas a támogató adatbázisokra való leképezésre, másfelől

pedig hordozhatóságot biztosít. Eredményeimet három téziscsoportba foglaltam össze.

Az első téziscsoport a relációs adatmodell alapján létrehozott reprezentációval foglalkozom. A hordozhatóság fogalma döntően a relációs szétválaszthatóság [8] fogalmára vezethető vissza a relációs adatmodell alapú reprezentáció környezetében, ezért a szétválaszthatóság kritériumát felhasználva bemutatok egy olyan adatszerkezetet, amely tetszőleges relációs adatszerkezetből előállítható, a létrehozott adatszerkezetre megfogalmazott lekérdezés pedig a leképezés alapjául szolgáló adatszerkezetre algoritmikusan átfordítható.

A második téziscsoport az általam bemutatott reprezentáció előnyeit, sajátosságait mutatja be egy konkrét, komplex jelentéstartalommal bíró nyelvi szerkezet, a birtokos szerkezet [5, 13] formális, nyelvfüggetlen leírásának új, algoritmikus feldolgozása kapcsán. Megmutatom, hogy a vizsgált reprezentáció és az arra épülő algoritmus a birtokos szerkezetek gépi feldolgozásában jelentős előrelépés, hiszen az alanyiságot kifejező, illetve többszörös láncolással előálló birtokos szerkezeteket egyedül ez a megoldás képes egyelőre megfelelően leképezni adatbázis-lekérdezésekre.

A harmadik téziscsoport az általam vizsgált reprezentáció nyelvi többértelműségre, heterogenitásra és jelentéstartalmi átfedésre vonatkozó kiterjesztéssel foglalkozik. Arra ad választ, hogy egy megfelelő fizikai szervezéssel, indexeléssel hogyan lehet a nyelvi pontatlanságokat kezelni, illetve hasonló, a felhasználói kérdésnek leginkább megfelelő válaszokat adatbázisban megtalálni. A téziscsoportban vizsgált eredmények mindig nyelvfüggőek abban az értelemben, hogy a nyelvi átfedések, többértelműségek legfeljebb egy nyelvet fedjenek le, ugyanakkor kellően általánosak ahhoz, hogy lényegében bármely nyelvre illeszthetőek legyenek.

## 2 Kutatási célkitűzések

Értekezésem célja, hogy egy skálázható, hordozható, a támogató adatbázisok adatszerkezetében már kódolt információkat teljes egészében felhasználó, a természetes nyelvi szerkezetek szemantikai szintű feldolgozását széleskörűen biztosító eljárásokat és megoldásokat adjon a relációs adatbázisok technológiájának felhasználásával. Az értekezésnek ugyanakkor nem célja és nem tárgya

**(a)** a természetes nyelvű szövegfeldolgozás szintaktikai, nyelvészeti aspektusainak vizsgálata,

**(b)** az információkinyerés adatbázistechnológiákon túlmutató problémái, beleértve a webkeresés különféle aspektusait is,

**(c)** valamint az elosztott és szövetséges adatbázisok [20] tanulmányozása sem.

A reprezentáció kidolgozásánál törekedtem arra, hogy a megoldás minél szélesebb körben alkalmazható legyen, lehetőség szerint tárgyterülettől, nyelvtől és konkrét adatszerkezetektől függetlenül. Döntéseimet, illetve a konstrukció létrehozásában jelentősen befolyásolta, hogy

az elérni kívánt eredményeket úgy alakítsam ki, hogy ne csak közvetlenül az NLIDB, hanem általában és hosszabb távon az információkinyerési és a természetes nyelv számítógépes „megértését” (NLU) [1] igénylő területeken közismert, viszonylag olcsó technológiákkal is alkalmazható legyen.

## 3 Módszertan

Kutatásom során új reprezentációs modellt dolgoztam ki az NLIDB hordozható szemantikai rétegének relációs adatmodell alapokon való felépítéséhez. A reprezentációs modell sajátosságainak igazolására felhasználtam analatikus módszereket, halmazelméleti, valamint a relációs adatmodell-elmélet terén elért eredményeket.

A reprezentációs modell kifejezőerejének tesztelését konkrét nyelvi szerkezet analitikus elemézésével, szimulációval igazoltam. A szimulációt indokolta, hogy a vizsgált rendszer általánossága miatt formális leírásuk általában nem adható meg a rendelkezésre álló analitikus eszközök segítségével, továbbá a vizsgálandó birtokos nyelvi szerkezet jelentéstartalmának formális, zárt alakban való megadása jelenleg nem ismeretes.

A jelentéstartalmi átfedés területén elért eredményeket, a felépített új modellre vonatkozó algoritmusok helyességét analitikus eszközökkel igazoltam.

## 4 Új eredmények

Az általam elért új eredményeket három téziscsoportba rendeztem. Az első téziscsoport egy új, szemantikai feldolgozást támogató adatreprezentációval, illetve annak sajátosságaival foglalkozik. A második téziscsoportban a természetes nyelvi birtokos szerkezetek jelentésének leképezését és algoritmikus feldolgozásának módját foglaltam össze. A harmadik téziscsoport a nyelvi átfedések, illetve az adatbázisban a felhasználói kérdésnek leginkább megfelelő adatok keresését lehetővé tevő indexelési eljárást tárgyalom.

Az egyes fontosabb állítások, modellek és definíciók mellett zárójelben a disszertáció vonatkozó oldalszámait jelenítettem meg a könnyebb átláthatóság és kereshetőség érdekében.

### 4.1 NLIDB adatszerkezet kialakítása

- Kapcsolódó publikációk: [J6, J3][C12, C15, C3].
- Ismert függelten hivatkozások száma: 0.
- Elismerések: HTE Pollák–Virág díj, 2006.

Az NLIDB megvalósításai két fontos paraméter között próbálnak egyensúlyt teremteni a széleskörű használhatóság, a hatékonyabb alkalmazás érdekében: a mondatok „megértésének” pontossága (feldolgozási pontosság) és a feldolgozási algoritmus adaptivitása (hordozhatósága) között.

Az eddigi munkák az NLIDB megvalósítását négy lényegesen paradigma mentén valósították meg. De ezek közül is kizárólag az adatbázis adatszerkezetéből visszafejtéssel előállított ER–modellt felhasználó megoldások [16, 17] használják fel az NLIDB tudásbázisaként működő támogató adatbázisok adatszerkezetében kódolt információkat. Az adatbázisban tárolt világmodell felhasználása az NLIDB működésében kézenfekvő és kívánatos megoldásnak tűnik, mindazonáltal logikusnak látszik nem egy adatbázis-specifikus ER–modell, hanem egy olyan általánosabb megoldás kidolgozása, amely az adatbázis-szerkezetet univerzális formában, egyfajta ontológiaként használja és értelmezi – majd ezt az univerzális szerkezetet képezi le a támogató adatbázisok egyedi adatszerkezeteire.

Szükség van tehát arra az adatmodellben kódolt információra, hogy

- melyek a konkrétan lekérdezhető individuumok, fogalmak és tulajdonságok,
- milyen közvetlen fogalmi alá-fölé rendeltség viszony áll fenn ezen individuumok között,
- milyen funkcionális függőségek adottak az attribútumok között,
- illetve mely attribútumok mely attribútumokkal köthetőek össze „értelmesen” pl. egy illesztés (join) keretei között.

Utóbbi kritérium azért fontos, mert például az 1038 mint szám szerepelhet történelmi évszámként is, irányítószámként is és egy település lakosainak számaként is – ennek megfelelően az egyes értelmezéseknek megfelelő attribútumok mentén különböző sémákat is lehet illeszteni a relációs adatmodell szabályai szerint. Az ad hoc egyezés hatására „értelmetlen”, rosszabb esetben kezelhetetlen méretű eredményre vezethetnek.

A téziscsoport eredményei az adatbázisban tárolt tudás újrafelhasználásának eddigi eredményeit nem csak általánosítja, hanem jelentősen ki is terjeszti azáltal, hogy lehetővé teszi tématerületekhez tartozó adatbázisok csatolását a korábbi modellek szerkezeti módosítása nélkül. Az általam létrehozott modell az első olyan sikeres próbálkozás, amely ontológiák, adatbázisok és természetes nyelvi reprezentációk között teremt kapcsolatot. A modell megalkotását a Szavak hálójában c. projekt (NKFP-0019/2002) keretében teszteltem, működőképességét a projekt sikeres lezárása fémjelzi. A modellre épülő rendszer megalkotásáért a Hírközlési és Informatikai Tudományos Egyesület a Pollák–Virág-díjat adományozta.

**I. Tézis**

*Megalkottam a tématerületi hordozhatóság matematikai modelljét olyan NLIDB rendszerek számára, amelynek szemantikai elemzéséhez használt reprezentációs rétege relációs adatmodellre épül, továbbá algoritmikusan eldönthetővé tettem a hordozhatóság eddig csak intuitív fogalomalkotással meghatározott tulajdonságait. A megalkotott modellről igazoltam, hogy a modellt alkotó kritériumok egyes komponsei tagonként külön-külön eldönthetőek. Megmutattam, hogy a relációs szétválaszthatóság [8], és ennek*

*megfelelően a relációs adatmodell alapú NLIDB hordozhatósága is független a relációs adatbázisok függésőrző tulajdonságától. Kiterjesztettem a Chan–Mendelzon-tételt [8] nem függésőrző sémaszerkezetekre is, és bebizonyítottam az így kapott általánosabb tételt.*

A tématerületi hordozhatóság fogalmát eddig csak intuitív módon határozták meg [15, 11, 3], így majdnem minden megoldásra általában elmondható, hogy hordozható, hiszen egyáltalán nem ellenőrizhető ez a tulajdonság. A hordozható NLIDB fogalma az alábbi tulajdonságokat nevesítették eddig:

- Az NLIDB tématerülettől nem függő nyelvi szabályok szerinti feldolgozással él.
- Tématerülettől független leíráson alapuló szemantikai feldolgozást valósít meg.
- Általános jellegű lexikonnal rendelkezik.
- Nem csak egyetlen adategységből, hanem az adatszerkezeten értelmezhető műveletek sorával előállítható információkat is képes megtalálni, felismerni, kezelni.
- Mélyebb nyelvi, programozói vagy más, informatikai jellegű tudást a felhasználótól az NLIDB kezelőfelülete nem igényel.

A továbbiakban a szakirodalomban szokásos jelöléstechnikákat használom, valamint olyan ismert fogalmakat mint a reprezentatív példány [19], a konzisztens adatbázis [19], független [19] és szétválasztható [8] adatszerkezet. Ezek formális leírása és sajátosságainak elemzése bővebben a disszertációmban olvasható (13–18. oldal). A szemantikai feldolgozást segítő relációs alapokon nyugvó NLIDB esetében a hordozhatóság kritériumai a következőképpen foglaltam össze (20–21. oldal):

- a relációs $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma>$ adatbázisnak, ahol $\mathbf{R}$ az adatbázis sémáinak, $\mathbf{r}$ a relációinak, továbbá $\Sigma$ az adatbázison értelmezett funkcionális és tartalmazási függőségek összessége, konzisztensnek és szétválaszthatónak kell lennie,
- bármely tématerületi bővítés nyomán létrejött $\mathcal{DB}' = <\mathbf{R}', \mathbf{r}', \Sigma'>$ adatbázis továbbra is konzisztens és szétválasztható úgy, hogy $\mathbf{R} \subseteq \mathbf{R}'$, $\mathbf{r} \subseteq \mathbf{r}'$ és $\Sigma \subseteq \Sigma'$ is teljesül,
- továbbá a relációs lekérdezőnyelven megfogalmazott lekérdezés előállítására kizárólag egy rögzített szintaxissal rendelkező, a rendszer bemeneti nyelvének jellemző szintaktikai jegyeit definiáló, formális, Γ-nyelvtant használhatunk fel.

A relációs adatbázisban a konzisztencia a reprezentatív példányon [19] alkalmazott `KÖVET` (chasing) eljárás segítségével véges sok lépésben eldönthető [19]. A szétválaszthatóság igazolására az ún. Chan–Mendelzon-tételt [8] alkalmazhatjuk, amely szerint, ha $\mathbf{R}$ függésőrző és független a $\Sigma$ függéshalmazt tekintve [19], és van olyan két különböző $R, S \in \mathbf{R}$ séma, amelyekre $R^+(\Sigma) \supseteq S$ – ahol $R^+(\Sigma)$ az $R$ séma mint attribútumhalmaz $\Sigma$ függéshalmaz szerinti lezártja –, $\mathbf{R}$ akkor és csak is akkor szétválaszható, ha létezik olyan nem triviális $\{X \rightarrow A\} \in \Sigma$, amelyre $X$ az $S$ séma szuperkulcsa és $A$ az $S$ egyik attribútuma. A tétel

segítségével véges számú attribútumhalmaz-lezárással és halmazművelettel egyértelműen eldönthetjük, hogy a $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma>$ adatbázis-szerkezet szétválasztható-e. Természetesen, a függésőrzőség, valamint a $\mathbf{R}$ függetlensége $\Sigma$ függéshalmaztól egyaránt véges sok lépésben igazolható [19].

Felmerül azonban a kérdés, hogy nem függésőrző lehetnek-e szétválaszthatóak? Egyszerű konstrukciók segítségével beláttam (18. oldal), hogy van olyan nem függésőrző relációs sémaszerkezet, amely szétválasztható, de nem függésőrző; illetve függésőrző, de nem szétválasztható, azaz a két fogalom független egymástól.

Legyen pl. $\mathbf{R} = \{R(A), S(B)\}$, ahol $R$ és $S$ két egyetlen attribútumból álló, valamint legyen $\Sigma = \{A \rightarrow B\}$, ahol $A \rightarrow B$ egy érdemi funkcionális függőség. Bármely $\mathbf{r}$ relációhalmaz esetén $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma>$ definíció szerint nem függésőrző, ugyanakkor mindig konzisztens és szétválasztható, mivel a reprezentatív példányban $A$ és $B$ értékei sosem vehetik fel kétszer ugyanazt azt értéket a relációk halmaztulajdonsága miatt. Ennek megfelelően a `KÖVET` eljárás során az $A \rightarrow B$ funkcionális függőséget sosem lehet felhasználni, tehát a reprezentatív példány vetítése az $\mathbf{R}$ sémaszerkezet elemeire mindig meg kell egyezzen az $\mathbf{r}$ relációhalmaz elemeivel. Ezzel beláttam, hogy szétválaszthatóságnak nem feltétele és nem következménye a függésőrző tulajdonság.

A függésőrző tulajdonság nem vonja maga után a sémaszerkezet szétválasztható tulajdonságát még akkor sem, ha a sémaszerkezet veszteségmentes és független egy adott $\Sigma$ függéshalmaztól. Legyen ugyanis $\mathbf{R} = \{R(ABC), S(ABD)\}$, $\Sigma = \{A \rightarrow D, AB \rightarrow C\}$ és $\mathbf{r} = \{r(R) = \{<a, b_1, c>\}, r(S) = \{<a, b_2, d>\}\}$. A $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma>$ adatbázis függésőrző, veszteségmentes, és $\mathbf{R}$ független $\Sigma$ halmaztól, hiszen a sémaszerkezet konzisztenciáját az $r(S)$ reláció konzisztenciája garantálja. Ugyanakkor a szerkezet nem szétválasztható, mivel a reprezentatív példány előállítása során a `KÖVET` eljárás az $A \rightarrow D$ függőséget felhasználva létrehoz egy $t[ABCD] = <a, b_1, c, d>$ elemet, amelyet az $S(ABD)$ sémára vetítve új elemet eredményez az $r(S)$ relációhoz képest. Igazoltam tehát, hogy a függésőrző és a szétválaszthatóság tulajdonságok egymástól függetlenek (18–19. oldal).

Mivel a függésőrzőség és a szétválaszthatóság független fogalmak, a Chan–Mendelzon-tételben szereplő függésőrző kritérium a tétel alkalmazhatóságát csökkenti; nem függőségőrző sémaszerkezet esetében még nem tudunk nyilatkozni sem a szétválaszthatóság, sem a hordozhatóság eldönthetőségéről.

Tegyük fel, hogy adott $\Sigma$ függéshalmaz és egy $\mathbf{R}$ sémaszerkezet, amelyre igaz, hogy $\mathbf{R}$ független a $\Sigma$ függéshalmaztól, de $\mathbf{R}$ nem függésőrző a $\Sigma$ halmazra nézve, azaz létezik olyan $X \rightarrow A \in \Sigma$ függőség, amelyben szereplő attribútumok együttesen nem fordulnak elő egyetlen sémában sem. Mivel $\mathbf{R}$ független a $\Sigma$ halmazt tekintve, ezért tetszőleges $\mathbf{R}$ sémaszerkezetre illeszkedő $\mathbf{r}$ relációhalmazra igaz kell legyen, hogy a belőle képzett reprezentatív példány konzisztens. Ennek megfelelően az $X \rightarrow A$ függőséghez sem lehet találni két olyan $t_1, t_2$ elemet valamely reprezentatív példányban, amelyre $t_1[X] = t_2[X]$ és $t_1[A] \neq t_2[A]$ teljesülne. Ez viszont csak úgy lehetséges, hogy ha bármely az $\mathbf{R}$

sémaszerkezetre illeszkedő $\mathbf{r}$ relációhalmazra igaz, hogy a belőle képzett reprezentatív példányban a reláció elemei az $X \cup \{A\}$ attribútumhalmazon nem teljesen kitöltöttek, vagyis $X$ kitöltése nincs hatással az $A$ attribútumok kitöltéseire. Amennyiben ugyanis valamely reprezentatív példány eleme az $X \cup \{A\}$ attribútumok mindegyikén nem `NULL` értékű, akkor szerkeszthető hozzá egy olyan Armstrong-reláció [4], amelyben az $\mathbf{R}$ sémaszerkezet által megőrzött legbővebb $\Sigma' \subset \Sigma$ függőshalmaz fennáll, de $X \to A$ nem teljesül – ez pedig ellentmondásban van $\mathbf{R}$ és $\Sigma$ függetlenségével.

A Chan–Mendelzon-tétel általánosításával kimondtam (18–19. oldal), hogy valamely $\mathbf{R}$ sémaszerkezet akkor és csak akkor szétválasztható egy $\Sigma$ függéshalmazt tekintve, ha $\mathbf{R}$ független a $\Sigma$ halmaztól, továbbá bármely két különböző $R, S \in \mathbf{R}$ séma, amelyekre $R^+(\Sigma') \supseteq S$ – ahol $\Sigma' \subseteq \Sigma$ a $\Sigma$ azon legbővebb függéshalmaza, amelyet az $\mathbf{R}$ megőriz –, létezik olyan nem triviális $\{X \to A\} \in \Sigma'$, amelyre $X$ az $S$ séma szuperkulcsa és $A$ az $S$ egyik attribútuma.

**II. Tézis**

*Megalkottam egy relációs adatmodellel reprezentálható metamodellt a relációs adatbázisokban tárolt szemantikai információk leírására. A metamodellről igazoltam, hogy helyes és teljes az érdemi kapcsolatokra nézve [J3][C3]. Megalkottam a természetes adatbázisok modelljét, amely a természetes nyelvben jelenlevő fogalmi összekapcsolásokat és többértelműségeket modellezi a relációs adatmodell felhasználásával [J3][C2, C3]. Konstruktív bizonyítást adtam arra, hogy létezik szétválasztható természetes adatbázis [C3].*

A természetes nyelvek reprezentálhatósága, algoritmikus feldolgozása szempontjából hasznos ontológiák létrehozása nagyon költséges, ugyanakkor az adatbázisokban nagyon hasonló formában rendelkezésre áll már ontológiai tudás. Az ontológiával szemben a relációs adatbázisoknak előnye, hogy matematikailag egységesebb, egyszerűbb és garantáltan számítható modellel dolgozhatunk, a kifejezőerő csökkenése árán. Ugyanakkor az adatbázisok merev, összetett szerkezetei, ugyanarra a problémára adott sokszínű megoldásai nem kedveznek a nyelvi jelenségek adatbázisbeli megfeltetésére. Nem véletlen, hogy az NLIDB területén létrehozott eddigi megoldások nem használták ki az adatszerkezetben rejlő szemantikai információkat.

A relációs adatmodellre épülő NLIDB tudásbázisának három követelménynek kell eleget tennie: hordozhatónak kell lennie, az NLIDB által lekérdezhető adatbázisok szemantikai leírását kell adaptív módon felhasználnia a megfelelő adatbázis-lekérdezések előállítására, végezetül pedig természetes nyelvi jelenségek megfelelő modellálására kell alkalmas legyen. A hordozhatóságot, amint korábban már rámutattam, eddig csak intuitív formában vizsgálták, a relációs adatmodellre épülő NLIDB esetében egyáltalán nem. A nemzetközi szakirodalom az adatbázisban tárolt szemantika leírásával foglalkozik kizárólag érdemben, nevezetesen az adatbázisokban tárolt sémaszerkezet egyed-kapcsolati modelljének visszafejtésével nyertek ki adatbázisokból szemantikai információkat [16, 17]. A megközelítés azonban nem alkalmas több adatbázis összefogására, hiszen az egyed-kapcsolati modellek

integrációja általában nagyon költséges feladat, ráadásul az adatbázisok sémaszerkezeteinek módosítására kifejezetten érzékeny. A természetes nyelvi jelenségek modellálására tipikusan ontológiákat vagy környezetfüggő nyelvtanokat alkalmaztak eddig [3][J3], a lekérdezendő adatbázisban már eleve kódolt tudást nem, bármennyire is kézenfekvőnek látszik [2]. A relációs adatmodellre épülő NLIDB esetében tehát az is kérdéses, hogy egyáltalán létezik-e olyan relációs adatmodellre felépíthető struktúra, amely lehetővé teszi a három követelmény együttes teljesítését.

A relációs adatmodell csak értékek közötti relációk alapján kapcsolhat össze adatokat, de kizárólag a felhasználó ismerheti, hogy az értékek mentén összeköthető relációk közül melyek hordoznak releváns információt. Az NLIDB számára ugyanakkor létfontosságú, hogy ezen kapcsolatok formalizáltan rendelkezésre álljanak, hiszen ezek nélkül nem garantálható, hogy az NLIDB szemantikailag helyes, összetett adatbázis-műveleteket hozzon létre. Szükség van tehát az adatbázisban tárolt szemantikai információk metamatikai modelljére, azaz arra az információra, hogy mely sémára illeszkedő relációk kapcsolhatóak össze illesztésekkel (join), más szavakkal: mely attribútumok között áll fenn érdemi kapcsolat. Az adatbázisok gyakorlati tapasztalatai alapján elmondható, hogy azon attribútumok között bizonyosan fennáll érdemi kapcsolat, amelyek között referenciális vagy hivatkozási integritás definiálva van. Márpedig két ilyen típusú kapcsolat van jelen relációs adatbázisokban: az idegen kulcs alapú hivatkozások és az `is-a` kapcsolatokat leíró hivatkozások.

Éppen ezért az attribútumhivatkozások (pl. külső kulcsok) jelölésére bevezettem a $\lambda$ hivatkozási függvényt (21–22. oldal) [J3][C8, C2, C7], amely attribútumhalmazokat képez le attribútumhalmazokra az alábbi feltételek mellett:

- az $R \in \mathbf{R}$ sémában $\lambda(X) = X$, amennyiben $X \rightarrow R$,
- a $\lambda$ függvény csak olyan $R \in \mathbf{R}$ sémában levő $X$ attribútumhalmazokra értelmezett, amelyekhez létezik $S \in \mathbf{R}$ séma úgy, hogy $\lambda(X)$ az $S$ séma kulcsa,
- továbbá $X \dashrightarrow \lambda(X)$ mindig teljesül bármely $X$ attribútumhalmazra.

A hivatkozási függvény mellett bevezettem az `is-a` bináris kapcsolatot reprezentáló bináris $\Xi$ relációt (21. oldal). A $\Xi(R, S)$ valamely $R, S \in \mathbf{R}$ sémákra csak abban az esetben igaz egy $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma>$ adatbázisban, ha $R$ és $S$ valamely alkalmas $X$, illetve $Y$ attribútumhalmazára $\Sigma \models \{X \rightarrow R, Y \rightarrow S, X \dashrightarrow Y\}$. Szemantikát leíró adatbázisnak neveztem el azt a $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma, \lambda, \Xi>$ ötöst, amelyben a $\Xi$ bináris reláció minden sémapárosra értelmezett, továbbá minden $A$ attribútumra igaz, hogy létezik egy $A \subseteq X$ attribútumhalmazt tartalmazó séma, amelyre $\lambda(X)$ értelmezve van (21–22. oldal).

Az érdemi bináris kapcsolatok jelölésére ezért bevezettem az $\varepsilon(X, Y)$ relációt (22. oldal), amely egy $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma>$ adatbázisban pontosan akkor igaz, ha létezik olyan $R, S \in \mathbf{R}$ séma, amely tartalmazza rendre $X$, illetve $Y$ attribútumhalmazt, továbbá $\Sigma \models \{X \dashrightarrow Y, Y \rightarrow S\}$ is teljesül, ahol $X \dashrightarrow Y$ tartalmazási függőséget jelöl. Mivel $\lambda$ függvény a szemantikát leíró adatbázisban ha értelmezett, akkor szükségszerűen kulcs az értéke,

továbbá definíció szerint $X \dashrightarrow \lambda(X)$ mindig fennáll, így a $\lambda$ függvény helyes az érdemi függőségekre nézve. Mivel a $\Xi$ reláció eleve csak kulcsok között értelmezett kapcsolatokra igaz, ezzel igazoltam, hogy a $\Xi$ is érdemi kapcsolatokat fejeznek ki. Következésképp, a szemantikát leíró adatbázis helyes az érdemi függőségekre nézve. A szemantikát leíró adatbázis teljes is az érdemi függőségekre nézve, ugyanis bármely kellően nagy számosságú attribútumokat tartalmazó sémaszerkezethez mindig konstruálható olyan Armstrong-relációhalmaz, amely megsérti valamely $X \dashrightarrow Y$ tartalmazási vagy az $Y$ attribútumhalmaz szuperkulcs tulajdonságát valamely $S$ sémában, ha ezek a $\Sigma$ halmazból nem következnek [7, 12]; ezzel pedig bebizonyítottam, hogy a szemantikát leíró adatbázisban nincs olyan $\varepsilon(X, Y)$ kapcsolat, amelyet az adatbázis függéseiből nem levezethető (22–23. oldal).

A természetes nyelvi jelenségek relációs adatbázisbeli modellezésére egy olyan szemantikát leíró modellt alkottam (24–26. oldal), amelyben az adatbázis sémáinak kulcsa minden esetben természetes kulcs (24. oldal) is egyben [J3][C8, C7]. A modellt természetes adatbázisnak neveztem el. Mivel a természetes kulcs mindig egyszerű az élő nyelvekben, hiszen alapvetően mindent elnevezünk, így a továbbiakban feltételezzük, hogy az adatbázisban minden kulcs egyszerű és a hivatkozási függvény minden attribútumra értelmezett. A természetes kulcs sajátossága, hogy nem egyértelmű, azaz több valós világbeli individuumnak is lehet ugyanaz az adatbázisbeli reprezentációja, így az adatbázisban megjelenik a többértelműség. A többértelműség azonban pontosan akkor áll fenn, ha az a természetes nyelvben is fennáll, az egyértelműsítésért a nyelvben megjelenő, illetve a nyelvből az adatszerkezetre leképezett kontextus meghatározása felelős [J6][C12, C15, C3].

A természetes adatbázis definiálására ekvivalens állításokat tettem (25. oldal). A szemantikát leíró adatbázis akkor és csak akkor természetes, ha $\lambda$ hivatkozási függvény értékkészlete is természetes kulcsokból áll. Ugyanis, minden $K$ kulcsattribútumhalmazra $\lambda(K)$, ennek megfelelően az állítás ekvivalens a természetes adatbázis definíciójában szereplő állítással. Hasonlóan igazoltam az érdemi kapcsolatok és a szemantikát leíró adatbázisok definíciója alapján, hogy egy szemantikát leíró adatbázis akkor és csak akkor természetes, ha bármely rajta értelmezett érdemi $\varepsilon(X, Y)$ kapcsolatban $Y$ természetes kulcs.

A természetes adatbázis azonban nem feltétlenül szétválasztható, ennek megfelelően a hordozhatósága is kétséges. Mivel a hordozhatóság esetében a sémák, a relációk és a függőségek halmaza is csak bővülhet, így heterogén környezetből származó sémák esetében a leguniverzálisabb leírási módra kell törekedni, azaz minden sémában legfeljebb annyi attribútumot szabad meghagyni, amely a világ modellezésére feltétlenül szükséges. Ebben az esetben ugyanis több attribútum garantáltan nem hagyható el, ezzel ténylegesen is csak a bővítésre ad lehetőséget. E gondolatok mentén beláttam, hogy a természetes kulcsok sosem hagyhatóak el egy sémából, viszont minden más attribútum igen – ezek ún. kapcsolótáblákon, idegen kulcsként természetes kulcsokra hivatkozó sémákon keresztül kötődnek a természetes kulcs által jelölt individuumokhoz. Következésképp célszerű megkövetelni, hogy a természetes kulcsok kizárólag önállóan, egyetlen attribútumból álló

sémákban forduljanak elő. Megjegyzem, ez a szemlélet analóg az ontológiák esetében alkalmazott leíró logikai fogalom (concept) és szerep (role) alapú felépítéssel azzal a különbséggel, hogy a szerep nem feltétlenül bináris reláció ebben az esetben.

A fentiek alapján normalizáltnak neveztem el az alábbi tulajdonsággal rendelkező $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma, \lambda, \Xi>$ természetes adatbázist (NNDB) [C3] (32–34. oldal):

1. Bármely természetes kulccsal rendelkező $R \in \mathbf{R}$ séma egyetlen attribútumot, a természetes kulcsát tartalmazza. Ezeket nevezzük elsődleges sémáknak.
2. Bármely természetes kulccsal nem rendelkező, másodlagosnak nevezett $R \in \mathbf{R}$ sémának legalább két attribútuma van.
3. Bármely két különböző $R, S \in \mathbf{R}$ sémára $R \cap S = \emptyset$.
4. Nincs két olyan különböző másodlagos $R, S \in \mathbf{R}$ séma, amelyre $R^+(\Sigma) \supseteq S$.
5. Bármely másodlagos $R \in \mathbf{R}$ séma Boyce-Codd normálformában van [10, 14].
6. Bármely $X \dashrightarrow Y$ tartalmazási függőség esetén $Y$ természetes kulcs.

Az NNDB szétválasztható (34. oldal) – ez a definíció több állításából is következik. Egyrészt megköveteltük, hogy különböző sémában szereplő attribútumok különböző jelöléssel legyenek reprezentálva; ennek a 3. tézisben lesz jelentősége. Ha minden attribútum különböző, akkor bármely sémának mint attribútumhalmaznak a lezártja az $\mathbf{R}$ adatszerkezet által megőrzött $\Sigma' \subseteq \Sigma$ szerint szükségszerűen önmaga. A Chan–Mendelzon-tétel általánosítása alapján már emiatt is szétválasztható. Hasonlóan, mivel $R^+(\Sigma') \supseteq S$ nem fordulhat elő az NNDB esetében, az NNDB emiatt is szétválasztható ugyancsak az általánosított Chan–Mendelzon-tétel alapján. Amennyiben az attribútumok nem különbözőek minden sémában, akkor a $R^+(\Sigma') \supseteq S$ megkötésnek az a jelentősége, hogy minimalizálja a potenciálisan felírható sémák számát, egyetlen sémába tömörítve az összetartozó elemeket. Fontos megemlíteni, hogy az NNDB nem feltétlenül függésőrző, de nyelvfeldolgozáshoz használt tudásbázisként erre nincs is közvetlenül szükség.

Ha az NNDB szétválasztható, ebből következik az is, hogy a formális Γ-nyelvtant adatbázis-lekérdezésre leképező NNDB alapú NLIDB szükségszerűen hordozható, ha bármely tématerületi módosítás eredményeképpen bővített tudásbázisa továbbra is NNDB marad. Egy tetszőleges adatbázis NNDB-re való leképezésének egy lehetséges, általam megalkotott algoritmusa részletesebben a disszertációban olvasható (26–31. oldal).

### 4.2 Jelölt birtokos szerkezetek algoritmikus feldolgozása

- Kapcsolódó publikációk: [J3][C8, C2, C7, C3].
- Ismert független hivatkozások száma: 0.
- Elismerések: NLIDB 2005 szakkonferencián a [C2] közleményt a legjobb öt cikkbe választották.

A birtokos jelzős szerkezetek nagyon sokféle szemantikai kapcsolatot fejezhetnek ki (lásd 4.1 táblázat és [6]), ráadásul a birtokos- és birtokszerepek is felcserélődhetnek eltérő szövegkörnyezetben (pl. a könyv szerzője, a szerző könyve), így algoritmizált feldolgozásuk, formalizálásuk korántsem egyszerű feladat. Ennek illusztrálására pedig lásd 4.2 táblázatot.

| Birtokos típusok | Példakifejezések (magyar és angol) |
|---|---|
| származás-, forrásleírás | Moszkva küldötte (men of Rome) |
| anyagleírás | – (ring of gold) |
| rész–egész viszony | a tanszék vezetője (head of department) |
| mennyiségi leírás | húsnak kilója (pound of beer) |
| (állandósult) kapcsolat | Péter felesége (Pam's address) |
| birtoklás | Sára sapkája (John's coat) |
| alanyiság | Verdi operája (dramas of Shakespeare) |
| tárgyiasság | Budapest látképe (portray of Elisabeth II) |
| cél- és szándékleírás | dolgozók iskolája (school of girls) |
| láncolás (halmozás) | Ábel apjának barátja (name of Tom's wife) |

**4.1 táblázat**. *Birtokos szerkezetek típusai*

| Birtokos kifejezés | Ekvivalens példa SQL–kifejezés |
|---|---|
| Bizet Carmenje | `SELECT cim FROM operak`<br>`WHERE szerzo = 'Bizet' AND cim = 'Carmen'` |
| Shakespeare drámái | `SELECT cim FROM dramak`<br>`WHERE szerzo = 'Shakespeare'` |
| Edit címe | `SELECT cim FROM cimek`<br>`WHERE nev = 'Edit'` |
| könyvek szereplői | `SELECT szereplo FROM szerepek`<br>`WHERE darab IN ( SELECT cím FROM konyvek )` |
| vállalat vezetői | `SELECT fonok FROM vallalat` |
| Petőfi anyjának neve | `SELECT nev FROM szemelyek`<br>`WHERE nev IN ( SELECT anya FROM csaladfa`<br>`WHERE gyermek = 'Petőfi' )` |

**4.2 táblázat**. *Birtokos szerkezetek és velük ekvivalens SQL-kifejezések*

Birtokos szerkezetek általános feldolgozása – akár a nyelvspecifikus, akár a nyelvfüggetlen változatot nézzük – jelenleg hiányzik a nemzetközi szakirodalomból. A birtokos szerkezetek feldolgozásának hiánya egyrészt az indexelők, osztályozók, másrészt az ismert NLIDB megoldások magas hibaszázalékát is indokolja. Ennek egyik jellemzője, hogy pl. a „Mikor születtek I. János Károly gyermekei?" kérdésre a találatok döntő többsége nem a gyermekek,

hanem I. János Károly király születéséről szólnak. A disszertációmban, illetve rangos hazai és nemzetközi fórumokon bemutatott, általam kidolgozott eljárás egy nagyon fontos, igen gyakori nyelvi szerkezet első átfogó, általános célú algoritmikus feldolgozására ad lehetőséget NNDB környezetben. Az általam javasolt (V)ISA-algoritmus, illetve a legismertebb, a szakirodalomban a legtöbbet hivatkozott megoldások közötti kifejezőerőbeli különbséget mutatja a 4.3 táblázat.

| Megvalósítás | származás | anyagleírás | partitív viszony | mennyiségleírás | állandó kapcsolat | birtoklás | alanyiság | tárgyiasság | célleírás | láncolás |
|---|---|---|---|---|---|---|---|---|---|---|
| Practice | n | n | i | n | i | n | n | n | n | n |
| START | i | n | n | n | i | n | n | n | n | n |
| SQ-HAL | n | n | n | n | i | n | n | n | n | n |
| NL for Cindi | n | n | n | n | i | n | n | n | n | n |
| Masque/SQL | n/a | n | i | n | i | n | n | n | n | n |
| NChiql | n | i | i | n | i | n/a | n | n/a | n | n |
| KID | n | n | n | n/a | i | n | n | n | n | n |
| (V)ISA | i | n | i | n | i | i | i | i | n | i |

**4.3 táblázat**. *Birtokos szerkezettípusok feldolgozás implementációk szerint*

Az algoritmus első részletes leírását publikáló cikkemet, az egyetlen, kizárólagosan NLIDB tématerülettel foglalkozó szakkonferencián a legjobb öt cikk közé választották 2005-ben.

### III. Tézis

*Megalkottam a szerepviszonyaiban jelölt, adatbázisban ábrázolható és lekérdezhető birtokos szerkezetek matematikai modelljét, az ún. (V)ISA-modellt, NNDB alapú NLIDB rendszerek számára. A modellről igazoltam, hogy a szakirodalomban megjelent NLIDB rendszerek közül elsőként alkalmas láncolt (összetett) birtokos szerkezetek modellálására, és jelentéstartalmuk meghatározására. Algoritmust adtam jelölt birtokos szerkezetek adatbázis-lekérdezéssé alakítására. Az algoritmusról igazoltam, hogy helyes a (V)ISA-modellre nézve. Az algoritmus és a modell jóságát teszteredményekkel igazoltam.*

A (V)ISA–modell és a rá épülő (V)ISA-algoritmus nyelvfüggetlen abban a tekintetben, hogy a modellezett jelenség nem függ annak nyelvi környezetétől, ugyanakkor tartalmaz egy nyelvfüggő relációt, amely a különböző nyelvekben esetlegesen eltérő, birtokos szerkezetekkel nem reprezentálható konkrét kapcsolatokat kizárja. A nyelvfüggés, illetve az annak

modellezésére megalkotott reláció nem saját eredmény (lásd [6, 5, 21]), ugyanakkor a reláció kiterjesztése az értelmezhetetlen nyelvfüggetlen birtokos kifejezések modellálására saját, önálló, új megoldás.

A birtokos szerkezetek SQL lekérdezéssé alakításakor az esetek döntő többségében beágyazott lekérdezést eredményeznek az általános NLIDB esetén is. A beágyazott lekérdezések tetszőleges mélységű kezelése meghaladják a ma ismert NLIDB rendszerek képességeit. Ennek egyik oka, hogy a nagyon gyakori birtokos szerkezetek jelentéstartalmának relációs adatmodellre való leképezésére eddig nem állt rendelkezésre kellően általános modell. Az általam megalkotott modell az első kísérlet arra, hogy algoritmikusan feldolgozhatóvá, kezelhetővé tegyük ezt a kiemelten fontos nyelvi szerkezetet.

Az általam létrehozott modell könnyebb átláthatósága érdekében az eddig alkalmazott $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma, \lambda, \Xi>$ jelölésrendszer kiegészítéseképpen bevezetjük az $\mathcal{A}$ és $\mathcal{I}$ szimbólumokat, amelyek rendre a sémák attribútumainak összességét, illetve az adatbázisban található attribútumok $\mathrm{DOM}(\mathcal{A})$ értékkészletét, az adatbázis individuumait jelölik. Jelölje továbbá $\alpha{\twoheadrightarrow}\beta$ azt a birtokos szerkezetet, amelyben $\alpha$ a szintaktikai értelemben a birtokos szerepű tag, míg $\beta$ a birtok.

A birtokos szerkezetek birtok szerepű tagja jellemzően individuum (természetes kulcsérték), individuumhalmaz vagy elsődleges séma. Az elsődleges sémák NNDB esetében azonban felfoghatóak individuumok halmazaként, ennek megfelelően az általános, adatbázisban leírható birtokos szerkezeteket úgy modelláltam, hogy az $\alpha{\twoheadrightarrow}\beta$ birtokos kifejezés birtokos szerepű tagja minden esetben nem üres individuumhalmaz, míg a birtok szerepű tag egyaránt lehet séma, attribútum vagy nem üres individuumhalmaz (40–42. oldal). Látni kell azonban, hogy bizonyos birtokos szerkezetek előfordulhatnak egy adott nyelvben, bizonyosak nem [5, 18, 22], ilyen értelemben a birtokos szerkezetek érvényessége nyelvfüggő; be kell vezetnünk egy $\Pi$ bináris predikátumot az érvényes birtokos szerkezetek jelölésére. Azt mondjuk, hogy $\Pi(\alpha, \beta)$ akkor és csak akkor igaz, ha létezik a természetes nyelvben olyan $\alpha{\twoheadrightarrow}\beta$ birtokos szerkezet, amely adatbázisban reprezentálható (42–43. oldal).

Mit értsünk viszont birtokos szerkezet szemantikai jelentése alatt, azaz milyen adatbázis-lekérdezés feleltethető meg a birtokos szerkezet jelentéstartalmának?

Jelölje $\|A\| = \{B \mid B \in \mathcal{A} \wedge \lambda(B) = \lambda(A)\}$ az $A \in \mathcal{A}$ attribútum NNDB-beli ekvivalenciaosztályát. Mivel az NNDB esetében minden hivatkozási függvény értéke természetes kulcs, itt valóban ekvivalenciaosztályokról beszélhetünk. egyen $\Psi : 2^{\mathcal{A}} \rightarrow 2^{\mathbf{R}}$ egy olyan leképezés, amely attribútumhalmazokhoz azt a maximális sémahalmazt rendeli, amelynek elemeire igaz, hogy az attribútumhalmaz legalább egy elemét tartalmazza. Azaz

$$\Psi(X \in 2^{\mathcal{A}}) = \{R \mid \exists A \in X \wedge A \in R\}.$$

Vezessük be a $\psi : \mathcal{I} \rightarrow 2^{\mathcal{A}}$ egy olyan leképezés, amelyre

$$\psi(I \in \mathcal{I}) = \{A \mid I \in \mathrm{DOM}(A)\}.$$

Ha $\beta$ egy adatbázisban reprezentálható birtokos szerkezet birtok szerepű tagja, akkor jelöljük

$\gamma$-val alábbi kifejezések egyikét $\beta$ tulajdonságainak függvényében:

$$\gamma = \begin{cases} \psi(\beta) & \text{ha } \beta \subseteq \mathcal{I} \\ \|\kappa(\beta)\| & \text{ha } \beta \in \mathbf{R} \\ \|\beta\| & \text{ha } \beta \in \mathcal{A} \end{cases}$$

Az általam bevezetett jelölések mellett az $\alpha{\twoheadrightarrow}\beta$ birtokos kifejezés jelentését a következőképp modelláltam és (V)ISA-modellnek neveztem el [J3][C8, C2, C7] (43–44. oldal): ha a létezik olyan $A, B \in \mathcal{A}$, amelyre $A \in \psi(\alpha)$, $B \in \gamma$ úgy, hogy valamely $R \in \mathbf{R}$ sémára $R \in \Psi(\psi(\alpha)) \cap \Psi(\gamma)$ és $\{A, B\} \in R$, továbbá $\Pi(A, B)$ is igaz, akkor $\alpha{\twoheadrightarrow}\beta = \mathrm{DOM}(B) \cap \mathcal{I}$, feltéve, hogy $\alpha$ egyetlen individuumot jelöl. Amennyiben $\alpha$ individuumok egy halmazát reprezentálja, akkor

$$\alpha{\twoheadrightarrow}\beta = \bigcup_{i \in \alpha} i{\twoheadrightarrow}\beta.$$

A (V)ISA-modell által definiált adatbázisban reprezentálható birtokos szerkezetek SQL-nyelvre fordítására megalkottam (V)ISA-algoritmust (1. algoritmus)[J3][C8, C2, C7]. A pszeudokódban szereplő $\kappa$ függvény egy séma természetes kulcsát adja eredményül (44–46. oldal). Jól kivehető, hogy a pszeudokód a (V)ISA-modell egyes elemeit lépésről lépésre számítja, így triviálisan igaz, hogy a (V)ISA-algoritmus a (V)ISA-modellre nézve helyes eredményt szolgáltat. Disszertációmban példákkal illusztráltam a (V)ISA-modell működését, így megmutattam, hogy a 4.2. táblázatban szereplő birtokos kifejezésekre a (V)ISA-algoritmus a táblázatban szereplő SQL-lekérdezésekkel ekvivalens lekérdezést állít elő.

Az NNDB adatbázis-technológiának a felhasználásával a feldolgozás átlagos esetben $O(n(\log|\mathcal{I}| + \log|\mathcal{A}|))$, ahol $n$ a keresések száma (44. oldal). Adatbázis-technológiával az individuumhalmazok és az attribútumok elemei, megfelelő ábrázolással és indexeléssel logaritmikusan kereshetőek – erre a disszertációmban mutattam példát (35–37. oldal) – , illetve egyetlen lekérdezéssel elérhetőek. Ha feltételezzük, hogy a keresés eredménye az esetek túlnyomó többségében korlátos, akkor az indexelés logaritmikus komplexitása miatt a kijelenthetjük, hogy (V)ISA-algoritmus is logaritmikus komplexitású. Egyes eseteket vizsgálva azonban a feldolgozási idő függ attól, hogy a birtokos szerkezet birtok szerepű tagja séma, attribútum vagy individuumhalmaz, illetve attól is, hogy a találati halmaz hány individuumot tartalmaz. Legrosszabb esetben tehát $O(|\mathcal{I}|)$ komplexitásról beszélhetünk.

A modell jósági paramétereit A szavak hájában (NKFP-0019/2002) keretében, közvetett módon tudtam csak megvizsgálni. Egyrészt azért, mert a rendszer teljesítőképessége nagyban függ az NLIDB egyéb komponenseitől is, így a szerepviszonyok megfelelő jelölésétől, feldolgozásától. Másrészt azért, mert a modell tesztelése adatbázisok, tanító adatok és a feldolgozás szempontjából nyelvészetileg megfelelően annotált korpuszok hiányában csak közvetett úton – a rendszer egészét tekintve – lehetett elvégezni. Itt jegyzem meg, hogy bár a szövegfeldolgozás számára korpuszok rendelkezésre állnak, természetes nyelvű kérdések, nyelvi szerkezetek és adatbázis-lekérdezések viszonylatában sem korpusz, sem adatbázis nem

**Algoritmus 1**: (V)ISA–algoritmus

```
1  Adott DB = <R, r, Σ, λ, Ξ> NNDB I individuumokkal és A attribútumokkal.
2  function VISA( α, β) returns SQL;
3  begin
4      if  β ∈ A then  γ := ‖β‖;
5      else if  β ∈ R then  γ := ‖κ(β)‖;
6      else  γ := ψ(β);
7      forall  R ∈ Ψ(ψ(α)) ∩ Ψ(γ) do
8          forall  A ∈ ψ(α) ∩ R do
9              forall  B ∈ γ ∩ R do
10                 if  Π(A, B) then
11                     post := ( β ∈ A ) ? " and 'B = β'" : "'";
12                     head := 'select B from R ';
13                     if  Halmaz( α) then
14                         return head + 'where A = α' + post;
15                     end
16                     else
17                         δ := ( α = γ↠ε ) ? VISA( γ, ε) : α;
18                         if  α ∈ R then
19                             return
                               head + 'where A in ( select κ(α) from α )' + post;
20                         end
21                         else
22                             return head + 'where A in (δ)' + post;
23                         end
24                     end
25                 end
26             end
27         end
28     end
29     return '';
30 end
```

ismeretes egyetlen idegen nyelvre sem, ezért összehasonlítási vizsgálatokra csak a kifejezőerő szintjén van lehetőség.

Megvizsgáltam, hogy az NLIDB felépítése, az elérhető forráskódok, illetve tudományos közleményekben publikált eredmények alapján melyik megoldás milyen típusú birtokos szerkezet feldolgozására alkalmas rögzített adatbázis-szerkezet esetében. Az egyes megoldások sajátosságait a 4.3. táblázat foglalja össze. A táblázat megfelelő mezőjébe *i* jel került akkor, ha lehet találni egy adott birtokos szerkezettípushoz több olyan konkrét esetet mint tanút, amelyet a megoldás képes feldolgozni. A vizsgálat alapján megállapítottam, hogy jelenleg a (V)ISA-algoritmus az egyetlen szakirodalomban tárgyalt modell és eljárás, amely az összetett, láncolt birtokos szerkezetek feldolgozására alkalmas.

A tesztelés során tipikus felhasználói kérdéseket és „ellenséges", azaz tudatosan nehezen értelmezhető szerkezeteket is vizsgáltam a birtokos szerkezetek tematikus szerepeinek összefüggésében. Azt tapasztaltam (47–49. oldal), hogy az adatbázisban reprezentálható birtokos szerkezetekre a modell 78,87%-os pontosságban a fenti modellre illeszkedik, a gyakori lekérdezési minták alapján pedig 91,83%-os pontosságot ért el. A tesztelésből megállapítható, hogy a bonyolult, jelentéstömörítő kifejezések azok, amelyek jellemzően nem írhatóak le a modell segítségével (pl. *filmek Mekkája*, *mérkőzés győztese*). Részletesebb elemzés a disszertációmban olvasható.

### 4.3 Hasonlósági keresések adatbázisokban

- Kapcsolódó publikációk: [J1, J2, J4, J5][C4, C6, C1, C13].
- Ismert függelten hivatkozások száma: 7.
- Elismerések: az Association for Computing Machinery (ACM) éves rendes világversenyén, a Knowledge Discovery and Data Mining Cup (KDDCup) rendezvényen 2. díjat kapott az indexelési eljárást tartalmazó megoldásunk kreatív ötlet és találati pontosság kategóriákban.

A felhasználók és az adatbázist építő-szervező szakértők közötti számottevő fogalomhasználatbeli eltérés lehet, különös tekintettel a fogalmak által jelölni szándékozott jelentéstartalomra. Természetesen, a jelenség nem csak lekérdező és adatbázis, hanem akár hasonló tematikájú, tartalmú adatbázisok közötti viszonyokra is fenn állhat. Különösen problémás a helyzet az olyan elnagyolt jelentésű fogalmakkal és nyelvi szerkezetekkel, mint amilyen pl. az olcsó, magyar, nagy vagy a közel lenni valamihez. Az eltérő, bizonytalan fogalomhasználatból azonban adódik, hogy a szakszerűen létrehozott és kezelt adatbázis-tartalom nem feltétlenül egyezik meg a felhasználói kérdésben rögzített fogalmakkal. Adatbázisokban csak és kizárólag egzakt kereséseket lehet végrehajtani – tipikusan néhány keresési kulcs alapján, ennek megfelelően egy a jelentéstartalmi különbség kezeléséhez szükség van a közelítő keresés támogatására adatbázisokban.

Az NLIDB természetes kiterjesztése, illetve a felhasználói viselkedés általános támogatása lehet egy olyan rendszer, amely a felhasználótól nem igényli még az adatbázisban kereshető, megfelelő szakkifejezések, névelemek ismeretét sem. Kutatásaim arra irányultak, hogy ezt a fajta kiterjesztést általánosan, egy új indexelési eljárással valósítsam meg. Az általam javasolt megoldás, a természetes nyelvi feldolgozás strukturált adatszerkezetek, azon belül is, NLIDB gerincét képező adatbázisok általi támogatása, egy teljesen új irányt képvisel.

**IV. Tézis**

*Eldönthető, többkulcsos, intenzionális attribútumokra is kiterjeszthető új, előrendezési relación alapuló indexelési eljárást alkottam a nyelvi jelentéstartalmi átfedések és fokozások számítógépes modellálására. Megmutattam, hogy a modell irányított körmentes gráfként reprezentálható. Igazoltam, hogy a gráf alapú indexelésre épített keresési eljárás az attribútumok sorrendjére invariáns, illetve támogatja az indexelt attribútumok részhalmazának, azaz részkulcsok keresését is $O(k \log_k n)$ költséggel $k > 1$ esetén, ahol $k$ a gráf átlagos fokszáma és $n$ a gráf pontjainak a száma. A modellt kiterjesztettem hasonló, közelítő és korlátos értékek több kulcs alapján való keresésére. A modellhez konstruktív módon megalkottam egy új keresési eljárást, amelyről igazoltam, hogy helyes és teljes a modellre nézve.*

Több attribútum szerinti hasonlóság keresés támogatására az adatbázisok esetében rangsoroló (ranking) eljárásokat szoktak alkalmazni, ezek azonban elsősorban a találati pontosságot, az adatbázisban tárolt és a lekérdezésben rejlő kifejezéseket közelíti egymáshoz. Sokkal általánosabb megoldást kapunk, ha az adatbázisban tárolt értékek között egy általános rendezési relációt definiálunk. Az alkalmazott indexelési eljárás újszerűsége az, hogy az adatbázisok elméletében kevéssé kutatott ún. előrendezési (pre-order) reláció tulajdonságaira építi fel a kereséshez használt indexfát, amely reláció jóval általánosabb a nemzetközi szakirodalomban használt rendezési relációknál (51–52. oldal). A modellnek előnye társaihoz képest, hogy képes jelentésbeli eltolódásokat, fokozódások, résztulajdonságok kezelésére éppúgy, mint származtatott, intenzionális attribútumok rendezésére – eddig egyedülálló módon. A modellről elmondható – bár ez nem része a tézisnek –, hogy az Bernays-Schönfinkel-Ramsey osztályba, ennek megfelelően az eldönthető modellek közé tartozik [J2].

Legyen adott egy $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma>$ adatbázis. Jelölje $ri(\Omega)$ $\mathcal{DB}$ adatbázis reprezentatív példányát, ahol $\Omega$ a $\mathcal{DB}$ adatbázis univerzális sémájának felel meg. Az $\Omega$ sémáról feltételezzük, hogy 0NF, azaz attribútumértékei értékhalmazok is lehetnek, míg a `NULL` kitöltéseket mindenhol üres halmaznak ($\emptyset$) tekintjük.

Legyen $t_1, t_2$ egy–egy ennese az $\Omega$ sémára illeszkedő $ri(\Omega)$ relációnak, továbbá legyen $\varrho$ egy olyan bináris, tranzitív és reflexív (előrendezési) reláció, amelyre

$$\varrho : \mathrm{DOM}(A) \times \mathrm{DOM}(A) \rightarrow \{\top, \bot\},$$

ahol $\top$ a logikai igaz, míg $\bot$ a logikai hamis szimbólumot jelöli, és $A \in \Omega$. Bevezettem, hogy

$t_1$ helyettesíthető $t_2$-vel valamely $A$ attribútumot és a $\varrho$ relációt tekintve, ha $\varrho(t_1[A], t_2[A]) = \top$, és ezt a tényt $t_1 \preccurlyeq_A^{\varrho} t_2$-vel jelöltem.

A helyetesíthetőséget attribútumhalmazokra is kiterjesztettem (52–53. oldal), ezt elneveztem nevezzük fedésnek. Legyen $t_1, t_2$ egy–egy ennese valamely $\Omega$ sémára illeszkedő $r_i(\Omega)$ relációnak, továbbá legyen $\Phi = \{\varrho_1, \varrho_2, \ldots, \varrho_n\}$ egy olyan bináris, tranzitív és reflexív relációkból halmaz, amelyre

$$\varrho_i : \mathrm{DOM}(A_i) \times \mathrm{DOM}(A_i) \to \{\top, \bot\},$$

bármely $i$ indexű $\varrho_i$ relációra és különböző $A_i \in \Omega$ ($A_i \neq A_j$, ha $i \neq j$) attribútumra. Azt mondjuk, hogy $t_1$ fedhető $t_2$-vel valamely $R \subseteq \Omega$ attribútumhalmazra és $\Phi$ relációhalmazt tekintve, ha

$$\forall A_i \quad A_i \in R \Rightarrow t_1 \preccurlyeq_{A_i}^{\varrho_i} t_2,$$

, ahol $\varrho_i$ az $A_i$ attribútumnak megfelelő reláció. Ezt a tényt $t_1 \sqsubseteq_R^{\Phi} t_2$-vel jelöljük.

Az előrendezési reláció gráfként reprezentáltam (53. oldal). Legyen $\mathcal{G}_R^{\Phi} = <\mathbf{V}, \mathbf{E}>$ egy irányított gráf, amelyben $\mathbf{V}$ elemei attribútumértékek egy rendezett halmazai, míg $<v_i, v_j> \in \mathbf{E}$ akkor és csak akkor, ha $v_i, v_j \in \mathbf{V}$, $v_i \neq v_j$ és $v_i \sqsubseteq_R^{\Phi} v_j$. Az ilyen tulajdonságú $\mathcal{G}_R^{\Phi}$ gráfot teljes $\Phi/R$–katalógusnak neveztem el. Mivel az előrendezés tranzitív, ezért értelemszerűen a teljes $\Phi/R$–katalógus erősen összefüggő részei klikkeket alkotnak.[J2][C6]. Az tranzitivitás redundáns tárolást jelent a gráf tárolása esetében, ezért érdemes a kikövetkeztethető élekkel redukálni a gráfot.

Legyen $\mathcal{G}_R^{\Phi} = <\mathbf{V}, \mathbf{E}>$ egy teljes katalógus, amelyből a redukált $\mathcal{G'}_R^{\Phi} = <\mathbf{V}', \mathbf{E}'>$ katalógust (röviden katalógust) úgy kapjuk meg, hogy $\mathbf{V}'$ elemei $\mathcal{G}_R^{\Phi}$ klikkjeinek felelnek meg, és $v'_i, v'_j \in \mathbf{E}'$ akkor és csak akkor, ha $v'_i \neq v'_j$ és

$$\exists v_i \exists v_j \forall v_k \qquad v_i \in C_i \ \wedge \ v_j \in C_j \ \wedge \ v_k \in C_k \ \wedge \ C_i \neq C_j \ \wedge \ C_i \neq C_k \ \wedge \ C_j \neq C_k \ \wedge$$
$$\wedge <v_i, v_j> \in \mathbf{E} \Rightarrow <v_i, v_k> \notin \mathbf{E} \vee <v_k, v_j> \notin \mathbf{E},$$

ahol $v'_i, v'_j$ a $C_i, C_j \subseteq \mathbf{V}$ klikkeknek megfelelő csomópontok és $C_k \subseteq \mathbf{V}$ tetszőleges klikk (53–54. oldal). Az így kapott gráf irányított körmentes gráf (DAG - Directed Acyclic Graph).

Könnyen belátható, hogy a katalógus támogatja a részkulcs alapú keresést is, hiszen a rendezési reláció az attribútumok összességére, így azok részhalmazaira is érvényes. A gráf segítségével szemlélethetjük a keresés legfontosabb lépéseit. Tegyük fel, hogy $R' \subset R$ nem üres attribútumhalmaz mint részkulcs alapján szeretnénk keresni. Ha a tetszőleges $v_j$ csomópontban vagyunk, akkor a részkulcs alapú kereséshez minden olyan $v_k$ csomópontot meg kell vizsgálnunk, amelyhez létezik $<v_j, v_k> \in \mathbf{E}$ irányított él, és amelyre a $v_j \sqsubseteq_{R'}^{\Phi} v_k$. Legrosszabb esetben valamennyi $v_k$ gyökerű részfát be kell járnunk, ami akár a teljes gráfot is jelentheti. Átlagos esetben, ha a gráf átlagos fokszáma $k$, akkor $k/2$ elágazást járunk be minden csomópontban. Mivel az elágazások számával a gráfban található irányított utak átlagos hossza logaritmikusan csökken – kiegyensúlyozott gráfot feltételezve –, ennek alapján a részkulcs alapú keresés átlagos komplexitására $O(k \log_k n)$ adódik, amennyiben $k > 1$. Ha

$k \leq 1$, akkor $O(n)$ átlagos költsége van a részkulcskeresésnek, hiszen ebben az esetben legfeljebb lineáris rendezésről beszélhetünk (54–56. oldal).

A relációs adatbázisok érték szerint kapcsolnak össze különböző valós világbeli megfigyeléshez, entitáshoz tartozó rekordokat. Ennek megfelelően a relációs adatbázisok megkövetelik a felhasználóktól, hogy a világbeli elemeket pontosan úgy nevezzék, ahogyan az adatbázisban tárolva, rögzítve van. Az NLIDB világában jellemzően több adatbázist kapcsolunk össze, így nem várható el sem az adatbázist létrehozóktól, sem a felhasználóktól, hogy minden körülmények között egyértelmű, félreértethetlen fogalmakat használjanak, hiszen ebben az esetben az embert rendelnénk a gép szolgálatába, ami nem mondható kívánatosnak.

Példának okáért, ha valaki tengerparti nyaralást tervez, és egy olyan keresőkifejezést szeretne megfogalmazni, hogy a szállodája legyen közel a tengerhez, ne legyen túl drága, legyen elérhető számos szórakozási lehetőség, akkor bizony egy utazási iroda jól felkészített adatbázisa sem tudna megfelelő ajánlatot adni az ügyfél számára. Akkor sem, ha a közelit mondjuk 200 méternél kisebbnek, a nem túl drágát kb. 100.000 forintos árnak, a számos fogalmát pedig legalább négynek értelmezzük. Az adott kérdésre könnyen előfordulhat, hogy nem fogunk választ kapni egy lekérdezéssel, hiszen pontosan ilyen ajánlat nem feltétlenül van az adatbázisban; valahogyan modellálni kellene, hogy mi lehet a rekord- vagy adathasonlóság.

Az adathasonlóságot kizárólag speciális, adott attribútum szemantikáját teljesen kihasználó rangsoroló algoritmusok segítségével szokás osztályozni. A rangsoroló algoritmusok azonban nem alkalmazhatóak hatékonyan többkulcsos keresés, illetve intenzionális és halmazértékű esetében sem. Az adathasonlóságnak ezért egy jóval általánosabb modelljét [J2] alkottam meg, a $\mathcal{G}_R^\Phi$-katalógusok felhasználásával (56–61. oldal).

Legyen $\mathcal{G}_R^\Phi = <\mathbf{V}, \mathbf{E}>$ egy redukált katalógus és $K$ egy keresőkifejezés. Vezessük be az alábbi jelöléseket és fogalmakat:

- Jelölje $\min(K)$ azon csomópontok maximális halmazát – ezt $K$ keresőkifejezés alsó korlátjának fogjuk nevezni –, amelynek elemei az alábbi tulajdonsággal bírnak:

  $$\min(K) = \{v \mid v \in \mathbf{V} \wedge \exists v_i \quad v_i \in \mathbf{V} \wedge <v, v_i> \in \mathbf{E} \wedge \neg v_i \sqsubseteq_R^\Phi K \wedge v \sqsubseteq_R^\Phi K \sqsubseteq_R^\Phi v_i\},$$

  ahol $\neg$ a logikai tagadás jele.

- Hasonlóan jelölje

  $$\max(K) = \{v \mid v \in \mathbf{V} \wedge \exists v_i \quad v_i \in \mathbf{V} \wedge <v_i, v> \in \mathbf{E} \wedge \neg K \sqsubseteq_R^\Phi v_i \wedge v_i \sqsubseteq_R^\Phi K \sqsubseteq_R^\Phi v\}$$

  a gráf adott tulajdonságú csomópontjainak maximális halmazát, amelyet $K$ keresőkifejezés felső korlátjának fogunk nevezni.

- Jelölje végül $\mathrm{sim}(K)$ azon csomópontok maximalis halmazát, ezeket $K$-hoz hasonló

adatelemeknek nevezzük, amelyekre ha $\max(K) \cap \min(K) = \emptyset$, akkor

$$\text{sim}\ (K) = \{v \mid v \in \mathbf{V} \wedge \exists v_i \quad (v_i \in \max(K) \cup \min(K)) \vee (v_i \in \max(K)\ \wedge\ <v, v_i> \in \mathbf{E})\ \vee\ (v_i \in \min(K)\ \wedge\ <v_i, v> \in \mathbf{E})\ \},$$

egyébként pedig $v_K = \max(K) \cap \min(K)$ $(v_k \neq \emptyset)$ esetén

$$\text{sim}\ (K) = \{\ v \mid v \in \mathbf{V} \wedge \exists v_i \quad ((<v_i, v_K> \in \mathbf{E}\ \wedge\ <v_i, v> \in \mathbf{E}) \vee (<v_K, v_i> \in \mathbf{E})\ \wedge\ <v, v_i> \in \mathbf{E})\ \}.$$

Az egyszerű keresési algoritmus [J2][C4] (55. oldal) működési elve szerint – a gyökérből indulva – egy olyan elemet keres, amelyet a keresőkifejezés fed. Ha ilyen elem nincs, akkor értelemszerűen a gráf csomópontjai által reprezentált ennesek egyike sem felel meg a keresési feltételnek. Ha mégis akadna ilyen, akkor a tranzitív tulajdonság felhasználásával az új kezdőpontnak a talált csomópontot veszi. Ha az adott csomópont fedi a keresőkifejezést, akkor a csomópont által reprezentált enneseket kerestük. Ha nem, akkor rekurzívan folytatjuk a keresést a talált csomópontból mint új gyökérből kiindulva.

A hasonló adatok feltérképezésének algoritmusát az általam alkotott hasonlósági kereső eljárás [J2][C1, C6] írja le (58. oldal), amely persze részben felhasználja az egyszerű keresésnél már alkalmazott `Trunc` eljárást. Az algoritmus megkeresi a gráfban azt a helyet, ahol a keresőkifejezésnek megfelelő csomópont a gráfban lennie kellene. Ha van olyan csomópont, amely a keresőkifejezésnek megfeleltethető, akkor az adott csomóponton kívül a hozzá kapcsolódó elemeket, ha ilyen csomópont nincs, akkor pedig a neki megfelelő csomópontot virtuálisan, a gráf szemantikájának megfelelően „beszúrva" a gráfba a virtuális csomópont környezetét adja vissza.

Disszertációban bizonyítottam, hogy ha a $\mathcal{G}_R^\Phi = <\mathbf{V}, \mathbf{E}>$ egy redukált katalógus, akkor a $v_i, v_j \in \mathbf{V}$ csomópontjaira $v_i \sqsubseteq_R^\Phi v_j$ akkor és csak akkor áll fenn, ha van a gráfban egy irányított út $v_i$ és $v_j$ között (54. oldal). Kimondtam, ha létezik irányított $v_i, v_j \in \mathbf{V}$ csomópontjai között úgy, hogy $v_i \sqsubseteq_R^\Phi v_j$, akkor a gráfban levő $v_i$ és $v_j$ között található összes irányított út tartalmazza az összes olyan $v_k \in \mathbf{V}$ csomópontot, amelyre $v_i \sqsubseteq_R^\Phi v_k \sqsubseteq_R^\Phi v_j$ (56. oldal). Az állítás bizonyítását a disszertációban megtalálható. E két állítás segítségével igazoltam, hogy a hasonlósági keresés algoritmusa helyes, azaz megfelelő $K$ keresőkifejezésre pontosan az adathasonlóság definíciója szerinti $\text{sim}(K)$ elemhalmazokat adja eredményül (57–59. oldal).

Az eljárás átlagos költségét nagyon nehéz meghatározni, hiszen ez nagyban függ a gráf és a keresőkifejezés sajátosságaitól. Annyi azonban elmondható, hogy legrosszabb esetben a gráf összes élét be kell járni; ekkor $O(|E|)$ komplexitásról beszélhetünk (59–60. oldal). Ha a keresendő kifejezés az adatbázisban megtalálható, akkor annak átlagos keresési költsége az egyszerű kereséssel azonos.

A KDDCup 2005 verseny kiadott adathalmazain tesztekkel igazoltam, hogy tudásszegény környezetben, dokumentumok hierarchikus kategóriafába való rendezése esetében ha rögzítjük az osztályozó beállítási paramétereit, akkor a katalógusok alkalmazása 12%-os

javulást eredményez (64–68. oldal). Az alkalmazott megoldásban a katalógusok előrendezési relációja az egyes kategóriákra jellemző szavak ún. tf-idf (term frequency, inverse document frequency) értékeire vonatkozó rendezési reláció volt [J4, J5][C13].

## 5 Eredmények alkalmazhatósága

A 4.1 és 4.2 szakaszokban bemutatott eredményeket a NKFP–0019/2002 jelű, sikeresen lezárult Szavak hálójában projekt keretében értem el. Az eredmények javítása, bővítése jelenleg is folyamatban van, hiszen két Gazdasági Versenyképességi Operatív Program, egy Jedlik Ányos pályázat és a Mobil Innovációs Központ keretei között folytatott kutatásai tevékenységemben is döntően ezeket az eredményeket használom fel olyan területeken, mint pl.:

- gazdasági tartalmú hírek összetartozó szálainak felderítése,
- webkereső motorok indexelési technikájának javítása,
- gazdasági tartalmú hírek tényadatainak automatikus kinyerése, formalizálása,
- távközlési hívásbejegyzések problémacentrikus tematikus osztályozása,
- mobilszolgáltatások dinamikus, futási idejű összekötettése,
- hely– és környezetfüggő mobilszolgáltatások biztosítása.

Mindegyik problémakör kulcseleme az ún. kontextusazonosítás és -felismerés, ahol az NNDB-re épülő, de nem feltétlenül természetes nyelvi, hanem pl. lokalitás, erőforrás, téma, kategória fogalmakat központba állító architektúrák sikeresen vizsgáznak.

Az algoritmusokat Java programnyelven, míg a relációs adatbázishoz köthető megoldásokat, teszteléseket Oracle adatbázis-kezelővel valósítottam meg. A megoldás időközben integrálódott egy ún. vizuális tezauruszon alapuló metakereső rendszerrel is. Jelenleg a teljes implementáció nyilvánosan még nem elérhető.

A 4.3 szakaszban bemutatott megoldást az ACM éves rendes adatbányászati rendezvényén, a KDDCup versenyen 2005–ben második helyezést elért tudásszegényv szövegkategorizáló [J4, J5][C13] egyik fontos elemét képezte. Az algoritmus elnyerte a kreatív ötlet kategóriában is a II. díjat.

## Irodalomjegyzék


[1] Allen, J. *Natural Language Understanding ($2^{nd}$ edition)*. The Benjaming/Cummings Publishing, Redwood City, CA, US, 1995.

[2] Amble, T. BusTUC: a natural language bus route oracle. In *Proceedings of the $6^{th}$ conference on Applied natural language processing* (San Francisco, CA, USA, 2000), Morgan Kaufmann Publishers Inc., pp. 1–6.

[3] Androutsopoulos, I., Ritchie, G. D., and Thanisch, P. Natural language interfaces to databases – an introduction. *Journal of Natural Language Engineering 1*, 1 (July 1995), 29–81.

[4] Armstrong, W. W. Dependency structures of data base relationships. In *IFIP Congress* (1974), J. L. Rosenfeld, Ed., North-Holland, pp. 580–583.

[5] Barker, C. *Possessive Descriptions*. PhD thesis, University of Carolina, Santa Cruz, Department of Linguistics, 1995.

[6] Barker, C., and Dowty, D. R. Non-verbal thematic proto-roles. In *Proc. of NELS 23 Conference* (Amherst, Massachusetts, 1992), North-Eastern Linguistics Conferences, GLSA Publications, pp. 49–62.

[7] Beeri, C., Dowd, M., Fagin, R., and Statman, R. On the structure of armstrong relations for functional dependencies. *Journal of ACM 31*, 1 (1984), 30–46.

[8] Chan, E. P. F., and Mendelzon, A. O. Independent and separable database schemes. In *PODS'83: Proceeding of the 2$^{nd}$ ACM SIGACT-SIGMOD symposium on Principles of database systems* (New York, NY, USA, 1983), ACM Press, pp. 288–296.

[9] Chen, P. P.-S. The entity-relationship model – toward a unified view of data. *ACM Transactions on Database Systems 1*, 1 (1976), 9–36.

[10] Codd, E. F. A relational model of data for large shared data banks. *Communications of ACM 13*, 6 (1970), 377–387.

[11] Copestake, A., and Sparck-Jones, K. Natural language interfaces to databases. *Knowledge Engineering Review 5*, 4 (1990), 225–249.

[12] Fagin, R. A., and Vardi, M. Armstrong databases for functional and inclusion dependencies. *Information Processing Letters 16*, 1 (1983), 13–19.

[13] Kim, J.-Y., Lander, Y. A., and Partee, B. H., Eds. *Possessives and Beyond: Semantics and Syntax*. GLSA Publications and Booksurge LLC, Amherst, MA, USA, Mar. 2005.

[14] Maier, D. *The Theory of Relational Databases*. Computer Science Press, Rockville, USA, 1983.

[15] Martin, P., Appelt, D., and Pereira, F. *Transportability and generality in a natural-language interface system*. Morgan Kaufmann Publishers Inc., Los Altos, CA, USA, 1986, pp. 585–593.

[16] Meng, X., and Wang, S. Nchiql: The chinese natural language interface to databases. In *DEXA'01: Proceedings of the 12$^{th}$ International Conference on Database and Expert Systems Applications* (London, UK, 2001), vol. 2113 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 145–154.

[17] Meng, X., Wang, S., and Wong, K.-F. Overview of a chinese natural language interface to databases: Nchiql. *International Journal of Computer Processing of Oriental Languages 14*, 3 (Sept. 2001), 213–232.

[18] Rappaport, G. C. *The Syntax of Possessors in the Nominal Phrase: Drawing the Lines and Deriving the Forms*. In Kim et al. [13], Mar. 2005, pp. 243–262.

[19] Sagiv, Y. A characterization of globally consistent databases and their correct access paths. *ACM Transactions on Database Systems 8*, 2 (1983), 266–286.

[20] Sheth, A. P., and Larson, J. A. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys 22*, 3 (1990), 183–236.

[21] Storto, G. *Possessives in Context – Issues in the Semantics of Possessive Constructions*. PhD thesis, University of California, Los Angeles, USA, Linguistics, 2003.

[22] Storto, G. *Possessives in Context*. In Kim et al. [13], Mar. 2005, pp. 59–86.

## 6 Publikációk

### Könyvfejezetek

[B1] Kardkovács, Z. T. *Szövegbányászat*. TypoTeX, Budapest, Hungary, June 2007, ch. Válaszkereső rendszerek, pp. 217–237.

[B2] Tikk, D., Kardkovács, Z. T., Magyar, G., Babarczy, A., and Szakadát, I. *Intelligent Systems at the Service of Mankind*, vol. 2. UBooks, Augsburg, Germany, Jan. 2006, ch. Natural Language Question Processing for Hungarian Deep Web Searcher.

### Folyóiratcikkek

[J1] Gajdos, S., Kardkovács, Z. T., and Surányi, G. Deduktív objektum-orientált adatbáziskezelők tervezése és megvalósítása. *Híradástechnika L*, 11 (Nov. 1999), 18–24. Journal on C5.

[J2] Kardkovács, Z. T., Surányi, G., and Gajdos, S. On the integration of large data banks by a powerful cataloguing method. *Periodica Polytechnica 48*, 1–2 (2004), 61–70.

[J3] Kardkovács, Z. T., and Tikk, D. On the transformation of sentences with genitive relations to sql queries. *Data & Knowledge Engineering 61*, 3 (2007), 406–416.

[J4] Kardkovács, Z. T., Tikk, D., and Bánsághi, Z. The ferrety algorithm for the kdd cup 2005 problem. *SIGKDD Explorations 7*, 2 (2005), 111–116.

[J5] Kardkovács, Z. T., Tikk, D., and Bánsághi, Z. A 2005-ös kdd kupa feladatának megoldása a fürkész algoritmussal. *Híradástechnika LXI*, 8 (2006), 50–58. Journal on C5.

[J6] Tikk, D., Kardkovács, Z. T., and Magyar, G. Deep web searcher for hungarian. *Internation Journal on Information Technology 1*, 1–4 (Dec. 2004), 191–197.

[J7] Tikk, D., Kardkovács, Z. T., and Magyar, G. A szavak hálójában: szabadszavas mélyháló-kereső program. *Híradástechnika LX*, 5 (May 2005), 2–8. Journal on C5.

[J8] Tikk, D., Kardkovács, Z. T., and Szidarovszky, F. Szótári névelemek felismerése és morfológiai annotálása. *Híradástechnika LXI*, 1 (2006), 29–34. Journal on C5.

[J9] Tikk, D., Kardkovács, Z. T., and Szidarovszky, F. Voting with a parameterized veto strategy: solving the kdd cup 2006 problem by means of a classifier committee. *SIGKDD Explorations 8*, 2 (2006), 53–62.

## Bírált konferenciacikkek és előadások

[C1] Surányi, G., Kardkovács, Z. T., and Gajdos, S. Catalogues from a new perspective: A data structure for physical organisation. In $8^{th}$ *East European Conf. on Advances in Databases and Information Systems, ADBIS 2004* (Budapest, Hungary, 2004), G. Gottlob, A. Benczúr, and J. Demetrovics, Eds., vol. 3255 of *Lecture Notes in Computer Science*, Springer Verlag, pp. 204–214.

[C2] Kardkovács, Z. T. On the transformation of sentences with genitive phrases to sql statements. In *Proceedings of the* $10^{t}$ *International Conference on Applications of Natural Language to Information Systems (NLDB)* (Alicante, Spain, June 2005), vol. 3513 of *Lecture Notes in Computer Science*, Springer Verlag, pp. 10–20.

[C3] Kardkovács, Z. T., Lejtovicz, E. K., and Kovács, G. Context identification: A relational database approach. In *Proceedings of the* $3^{nd}$ *Language & Technology Conference* (Poznan, Poland, Oct. 2007), Z. Vetulani, Ed., pp. 211–215.

[C4] Kardkovács, Z. T., Surányi, G., and Gajdos, S. Application of catalogues to integrate heterogeneous data banks. In *Proceedings of On The Move to Meaningful Internet Systems 2003* (Nov. 2003), vol. 2889 of *Lecture Notes in Computer Science*, Springer Verlag, pp. 1045–1056.

[C5] Kardkovács, Z. T., Surányi, G., and Gajdos, S. Ubiquitous access to deep content via web services. In *Web Engineering, International Conference, ICWE 2003* (Oviedo, Spain, 2003), J. M. C. Lovelle, B. M. G. Rodríguez, L. J. Aguilar, J. E. L. Gayo, and M. del Puerto Paule Ruíz, Eds., vol. 2722 of *Lecture Notes in Computer Science*, Springer Verlag, pp. 208–211.

[C6] Kardkovács, Z. T., Surányi, G., and Gajdos, S. Towards building knowledge centres on the world wide web. In $3^{rd}$ *Int. Conf. on Advances in Information Systems, ADVIS 2004* (Izmir, Turkey, 2004), T. Yakhno, Ed., vol. 3261 of *Lecture Notes in Computer Science*, Springer Verlag, pp. 139–149.

[C7] Kardkovács, Z. T., and Tikk, D. Szintaktikailag elemzett birtokos kifejezések algoritmizált fordítása adott formális nyelvre. In *Magyar Számítógépes Nyelvészeti Konferencia* (2005), D. Csendes and Z. Alexin, Eds., Szegedi Tudományegyetem Informatikai Tanszékcsoport, pp. 267–276.

[C8] Kardkovács, Z. T., Tikk, D., and Magyar, G. (V)ISA: A model for transforming genitive phrases into sql statements. In *Proceedings of the* $2^{nd}$ *Language & Technology Conference* (Poznan, Poland, Apr. 2005), pp. 58–62.

[C9] Tikk, D., Bíró, G., Szidarovszky, F., Kardkovács, Z. T., Héder, M., and Lemák, G. Magyar internetes gazdasági tematikájú tartalmak keresése. In *Magyar*

*Számítógépes Nyelvészeti Konferencia* (2006), Z. Alexin and D. Csendes, Eds., Szegedi Tudományegyetem Informatikai Tanszékcsoport, pp. 3–14.

[C10] Tikk, D., Szidarovszky, F., Kardkovács, Z. T., and Magyar, G. Entity recognizer in hungarian question processing. In *Proceedings of the* $9^{th}$ *Congress of the Italian Association for Artificial Intelligence* (Milano, Italy, Sept. 2005), vol. 3673 of *Lecture Notes on Artificial Intelligence*, Springer Verlag, pp. 535–546.

[C11] Tikk, D., Szidarovszky, F., Kardkovács, Z. T., and Magyar, G. Named entity recognition in a hungarian nl based qa system. In *Proceedings of the* $14^{th}$ *International Conference on Information System Developement* (Karlstad, Sweden, Aug. 2005), Kluwer Academic/Plenum Publishers.

[C12] Tikk, D., Kardkovács, Z. T., Andriska, Z., Magyar, G., Babarczy, A., and Szakadát, I. Natural language question processing for hungarian deep web searcher. In $2^{nd}$ *IEEE International Conference on Computational Cybernetics, ICCC 2004* (Vienna, Austria, Sept. 2004), W. Elmenreich, W. Haidinger, and T. Machado, Eds., pp. 303–309.

[C13] Tikk, D., Kardkovács, Z. T., and Bánsághi, Z. Topic mapping: a tool for finding the meaning of internet search queries. In *INES'06 Proceedings of IEEE Int. Conf. on Intelligent Engineering Systems, 2006* (2006), pp. 227–232.

[C14] Tikk, D., Kardkovács, Z. T., and Magyar, G. The hungarian deep web searcher project. In *International Conf. on Information Technology, ICIT 2004* (Istanbul, Turkey, Dec. 2004), A. Okatan, Ed., pp. 76–79.

[C15] Tikk, D., Kardkovács, Z. T., and Magyar, G. Qa system for hungarian based on deep web search. In *Proceedings of the* $2^{nd}$ *Romanian-Hungarian Joint Symposium on Applied Computational Intelligence* (Temesvár, Romania, May 2005), pp. 399–410.

[C16] Tikk, D., Kardkovács, Z. T., and Magyar, G. Searching the deep web: The WoW project. In *ISD'2006 Proceedings of the* $15^{th}$ *International Conference on Advances in Information Systems Development* (Budapest, Hungary, 2006), W. Wojtkowski, W. G. Wojtkowski, J. Zupancic, G. Magyar, and G. Knapp, Eds., Springer Verlag, pp. 493–504.

[C17] Tikk, D., Kardkovács, Z. T., Magyar, G., Babarczy, A., and Szakadát, I. Natural language module of a hungarian deep web searcher. In *IEEE* $4^{th}$ *International Conference on Intelligent Systems Design and Application* (Budapest, Hungary, Aug. 2004), pp. 73–77.

## Egyéb közlemények

[O1] Lejtovicz, K. E., and Kardkovács, Z. T. Anaforafeloldás természetes nyelvű szövegekben. In *Magyar Számítógépes Nyelvészeti Konferencia* (2006), Z. Alexin and D. Csendes, Eds., Szegedi Tudományegyetem Informatikai Tanszékcsoport, pp. 362–363.

[O2] Kardkovács, Z. T. Deduktív objektumorientált adatbázis-kezelő modellezése és fizikai rétegének hatékony megvalósítása. Master thesis, Budapest University of Technology and Economics, Department of Telecommunications and Telematics, June 2001.

[O3] Kardkovács, Z. T. Az információábrázolás matematikai módszerei. Tech. Rep. MEK-00854, BME Informatikai Központ – Axelero Internet Rt., Magyar Elektronikus Könyvtár, June 2003.

[O4] Kardkovács, Z. T. Fejlett, strukturált információs rendszerek. Tech. rep., BME Informatikai Központ – Axelero Internet Rt., Magyar Elektronikus Könyvtár, May 2003.

[O5] Kardkovács, Z. T., Surányi, G., and Gajdos, S. On the integration of large data banks by a powerful cataloguing method. In *John von Neumann PhD Conference* (Budapest, Hungary, Oct. 2003), pp. 47–50.

[O6] Kardkovács, Z. T., Surányi, G., and Gajdos, S. An axiomatic model for deductive object-oriented databases. In *Proceedings of the* $5^{th}$ *International Symposium of Hungarian Researchers* (Budapest, Hungary, Nov. 2004), pp. 325–336.

[O7] Tikk, D., Bíró, G., Szidarovszky, F., Kardkovács, Z. T., Héder, M., and Lemák, G. Categorization-based topic-oriented internet search engine. In *HUCI'2006 Proceedings of the* $7^{th}$ *Symposium of Hungarian Researchers on Computational Intelligence* (Budapest, Hungary, 2006), pp. 233–246.

[O8] Tikk, D., Kardkovács, Z. T., and Magyar, G. Wow: the hungarian deep web searcher. In *Proceedings of the* $5^{th}$ *International Symposium of Hungarian Researchers* (Budapest, Hungary, Nov. 2004), pp. 135–146.