FORMÁLIS NYELVEK SZEREPE A TERMÉSZETES NYELVEK MODELLEZÉSÉBEN

FORMAL GRAMMARS IN MODELING OF NATURAL LANGUAGES

Kovács László[*], Zsolt Tóth[*], Tamás Meszticzky[*],Gábor Szemán[*]

ABSTRACT

A core module in natural language processing engines is the grammatical parsing unit. This unit determines the grammatical roles of the incoming words and it converts the sentences into semantic models. knowledge about the structure of grammar of the language. Theory of formal languages and grammars give the mathematical background to model natural and artificial languages. The traditional, automata-based parsers are usually not very effective in the parsing of inflection transformations. The paper presents the main difficulties in modeling the grammar of natural languages. As these grammar symptoms can't be modeled with the traditional mechanism of formal grammars, heuristic grammar parsers are developed for the practical application.

1. TERMÉSZETES NYELVI INTERFÉSZEK

Az emberi kommunikáció természetes módja a nyelv mind írott mind beszélt formában. A történelem folyamán különböző nyelvek alakultak ki, melyeket a nyelvészek különböző csoportokba sorolnak. A természetes nyelvek mindegyikéről elmondható, hogy viszonylag nagy szókinccsel és absztrakt szabályokkal rendelkeznek, melyek elsajátítása nem egyszerű feladat. Az elmúlt években bekövetkezett robbanásszerű fejlődés az informatikában, telekommunikációban a felhasználók széles köre számára tette lehetővé az információs technológiák használatát miközben a meglévő rendszerek egyre összetettebbeké, komplexebbeké váltak. Ezért a korszerű rendszerek használta gyakran magas szintű jártasságot igényel, amit a felhasználóknak el kell sajátítaniuk a rendszer megfelelő használatához, ami gyakran nehéz és időigényes feladatnak bizonyul.

Az ember – gép kommunikáció területén folyó kutatások egyik iránya a természetes nyelvi interfészek (NLI) kialakítására irányul. A természetes nyelvi interfész valósítja meg a kapcsolatot a felhasználó és a rendszer között úgy, hogy a felhasználó természetes nyelvi parancsokkal, utasításokkal, kérdésekkel tudja kezelni az adott rendszert. Az NLI modulok egyik fontos eleme a nyelvtani értelmező, amely ellenőrzi a bejövő mondatok formai helyességét, majd felépít egy szemantika tartalmat hordozó leírást. A feladat nehézségét a nagy szókincsen, a bonyolult absztrakt nyelvtani szabályokon túl fokozza az egyes nyelvek közötti különbségek. Jelenleg már érhetőek el olyan kereső alkalmazások főként angol nyelven, melyek képesek egyszerű kérdések megválaszolására.

A modern számítógépek elterjedését követően az emberek felismerték, hogy az információ minden formáját - legyenek azok számok, képek, hangok - megfeleltethetők sztringeknek. Ezek a sztringek képezik a "nyelvek" halmazát, ami az egyik központi eleme lett az informatika tudományának. Ezt a területet érinti számos alapvető matematikai tulajdonsága a nyelveknek, nyelvgeneráló rendszerek, mint például a nyelvtanok. A számítógépes nyelvek mindegyike pontosan leírható nyelvtanok segítségével. Továbbá a nyelvtan segítségével programokat is írhatunk (szintaktikai elemzőket), melyek képesek eldönteni egy adott sztringről, hogy szintaktikailag helyes e az adott programozási nyelvben. Sokan remélték, hogy a természetes nyelvek is elemezhetőek olyan pontosan, hogy hasonló elemző programokat készíthessünk, amelyek el tudják dönteni egy mondatról, hogy nyelvtanilag helyes-e. Napjainkban az ilyen célú programok teljesítménye jóval elmarad az elvárttól. A legfőbb probléma, hogy nincs közös megállapodás arról, hogy mi számít nyelvtanilag helyes mondatnak; senki sem tudott még előállni egy olyan nyelvtannal, amely elég precíz leírást adna ahhoz, hogy véglegesnek tekinthetnénk.

A természetes nyelvek modellezésére formális nyelvek témaköre nyújtja a matematikai hátteret. A formális nyelvek elméletét az 1950es évek táján dolgozták ki. Szeretném kiemelni Gold és Chomsky munkásságát, akik kiemelkedő eredményeket értek el ezen a területen. Habár a formális nyelvek témaköre több évtizedre tekint vissza, jelenleg is számos nyitott kérdést tartogat. A következőkben a formális nyelvek

* *Miskolci Egyetem, Általános Informatikai Tanszék*

Chomsky féle osztályozását vesszük alapul és vizsgáljuk meg, hogy az egyes nyelvtani osztályok az egyes természetes nyelveket, milyen mértékben fedik le.

2 .FORMÁLIS NYELVEK

Az egyes nyeleket egy adott ∑ véges ábécé felett értelmezzük az alábbi módon L≤{ω | ∑*}, azaz a nyelv az ábécé elemeiből képzett tetszőlegesen hosszú sorozatok egy részhalmaza. Az egyes nyelvek felsorolással történő megadása meglehetősen időigényes feladat lenne, ezért az egyes nyelveket az őket leíró generatív nyelvtanuk segítségével szokás megadni. A nyelvtant az G=<T,N,P,S> négyessel szokás megadni [11], ahol:

T: a terminális szimbólumok véges halmaza (a,b,c)
N: a nem terminális szimbólumok véges halmaza (A,B,C)
P: képzési szabályok véges halma P = {α→β | α,β in {T union N}*}
S: Mondatkezdő szimbólumok halmaza S in N

A formális nyelveket a helyettesítési szabályaik alapján különböző kategóriákba lehet sorolni. Ezen csoportosítások legismertebbje a már fent is említett Chomsky féle hierarchia [5], ahol az egyes nyelvtani osztályok között tartalmazási relációs is fenn áll.

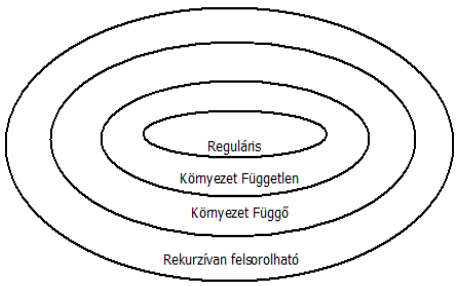
Nyelv	Megvalósítás	Képzési szabály
Rekurzívan felsorolható nyelvek	Turing gép	$\alpha \rightarrow \beta$
Környezet függő nyelvek (CSG)	Linear-bounded automata	$\alpha A \beta \rightarrow \alpha \gamma \beta$
Környezet független nyelvek (CFG)	Push-down automata	$A \rightarrow \alpha$
Reguláris nyelvek	Véges állapotú automata	$A \rightarrow aB$ vagy $A \rightarrow Ba$ és $A \rightarrow a$

Táblázat 1., Formális nyelv kategóriák

A leggyakrabban alkalmazott reprezentációs módszer a formális nyelvek területén a véges automaták (FSA) eszközrendszere, amely a reguláris nyelvek leírására szolgál. Az automatában az egyes csomópontok rendszerint az egyes nem-terminális szimbólumoknak felelnek meg, míg az egyes élek a terminál szimbólumokhoz rendelődnek. Egy átmenet akkor következhet be, ha az élhez tartozó terminál szimbólum a soron következő jel a vizsgált mondatban.

A verem automata egyik fontos jellemzője, hogy a rendszer egy saját verem struktúrájú memóriával rendelkezik, ezáltal lehetővé válik a CFG jellegű nyelvtani elemek modellezése is. Ezen nyelvcsaládba tartoznak többek között a $V^nB^mC^n$ alakú nyelvek is, ahol az automata meg tudja jegyezni a korábban beérkezett jelek jellemzőit.

A gyakorlatban sokszor használt eszköz a TAG (Tree Adjoining Grammar) [9] mechanizmus, mely egy nyelvtan fák kapcsolódó rendszerének feleltethető meg. A TAG rendszerben a nyelvtan is egy (N,T,I.A) négyesnek feleltethető meg. Ebben a szerkezetben az I elem az induló nyelvtan fák halmazát jelöli, míg A a kiegészítő fák halmaza. Az induló fák kibővíthetők az által, hogy a fa egyes levelei egy külső fával helyettesíthetőek. A külső fák levelei szintén újabb külső fákhoz rendelhetőek, ezáltal egymásba ágyazott modulok jönnek létre a nyelvben.



1. Ábra, Formális nyelvek kategóriái

A kognitív nyelvészet gyökerei a 1970-es évekig nyúlnak vissza. Az első rendszerek egyik fő képviselője a Langacker [10] modellje. A kognitív irányzat egyik fő jellemzője, hogy az emberi nyelv az ember általános kognitív, tanulási képességeit tükrözi. Az ember születésekor egy általános tanulási képességgel rendelkezik, melynek segítségével tanulja meg a nyelvet a tapasztalataira építve. Napjainkban egyre nagyobb szerepet kapnak a kognitív irányzatok a hagyományos generatív modellekkel szemben. Az irodalomban is több modell változat jelent meg ebben a körben. Ezekből többek között megemlíthető a Word Grammar (WG) modell, melyet Hudson [12] alkotott meg. E modellben a nyelvtan egy gráffal írható le, mely lefedi a tudásábrázolás mind a négy fő rétegét:a szemantikai szintet, a szintaktikai réteget, a morfológiai réteget és a fonológiai réteget.

Ezen elképzeléssel részben rokon a Dependency Grammar (DG), Tesneire [14] modell.. Ebben a reprezentációban az alapvető egységek a mondat egységek, a szavak közötti függőségek lesznek. A függőség irányított kapcsolatot jelent a mondat elemi között és bizonyos nyelvtani és jelentésbeli viszonyt szimbolizál.

3. NYELVI JELENSÉGEK MODELLEZÉSE

A természetes és a formális nyelvek között több hasonlóságot is felfedezhetünk. Mind a természetes mind a formális nyelvek szemantikai jelentést - értelmet - is adnak az adott nyelvnek. A természetes nyelveknél ez elkerülhetetlen, azonban a formális / mesterséges nyelveknél ez egy nem kívánatos információ. Emellett mindkettő elválasztja egymástól a szintaktikát és a szemantikát. Ezen hasonlóságok ellenére mégis jelentős különbségek vannak a természetes és a formális nyelvek között. Először is a természetes nyelvek már több ezer éve léteznek, és nem tudjuk ki tervezte őket; míg a formális nyelveket logikával és számítástechnikával foglalkozó tudósok alkották, hogy megfeleljen bizonyos tervezési kritériumoknak. A legjelentősebb különbség pedig a tény, hogy a formális / mesterséges nyelvek teljes mértékben körülhatárolhatók és tanulmányozhatók. [8]

A természetes nyelvek jelenségeinél a formális nyelvek alkalmazása során több problémával is találkozhatunk.

Nem-reguláris elemek:

- Nem kötött szórend: egyes nyelvekben a szavak sorrendje nem kötött a mondaton belül, nincs domináns szórend. Például a ."Peti könyvet olvas" és s "Könyvet olvas Peti" nyelvtanilag helyes mondatok. Ugyan ezen sorrend változatok leírhatók reguláris elemekkel is, az egyes sorrendek explicit megadásával; de nincs lehetőség a különböző sorrendváltozatok együttes kezelésére. Az irodalomban fellelhetők javaslatok a sorrendfüggetlenség leírására, mint például a DAwtl nyelvtan.

- Rugalmas ragozások: a ragozó nyelvekben a szavak ragozásai mutatják a fogalom szemantikai szerepét. Például a "kosár" alapszóból képzett "kosaraitokkal" szónál utalunk a többes számra, a birtoklásra és a felhasználási szerepre. A ragozás rendszerint nem egyszerűen csak az elemi ragok láncolata; az egyes ragok egymásra hatnak és illeszkednek a hangok és betűk. Ez a rugalmasság mögött egyfajta környezet függőség is megjelenik.

- Memória alapú jelenségek: a nyelv olyan konstrukciós készlettel rendelkezik, melyben egyes elemek tetszőleges sokszor ismételhetők, de az ismétlések számát meg kell jegyeznie a rendszernek. Példaként vehetjük az alábbi mondatot: 'Gabi azt mondta, hogy Zoli azt mondta, hogy ... , hogy Tibi azt mondta, hogy Feri iszik és igaza van, és igaza van,..., és igaza van.". Ez a mondat a klasszikus $V^nB^mC^n$ nem reguláris mintára illeszkedik.

- Környezet függő elemek: a természetes nyelvekben a szemantika kihat a szintaktikára, azaz a nyelvtani elem függ a tartalomtól. Például a mondatban egyeztetni kell a főnév német a névelőjével: a névelő alakja függ egy másik elem szemantikájától.

- Kivételek rugalmas kezelése: az alapszabályok mellett a nyelvek kivételeket is tartalmaznak. A kivétel oka lehet például a szóalak foglaltsága vagy éppen a hagyomány.

Nyelvtaniság kritériumai:

- Nyelvtanisági (nyelvtani helyesség) fokozatok: a nyelvtani helyesség több szinten értelmezhető. Egyrészt a helyesség az elfogadott hivatalos nyelvtannal mérhető, másrészt a köznapi érhetőség mércéjével is ellenőrizhető. Ez utóbbi erősen egyén és korszak függő.
- A nyelv használói alakítják a nyelvet az új szituációknak, helyzeteknek megfelelően, felülírva a nyelvtanisági nézeteket. A nyelvek szoros egymásra hatásban fejlődnek. Napjainkban jelentősen felgyorsult a nyelvek keveredése az információs technológia térhódítása és a gyors technológiai fejlődés által. Az átvett vagy beilleszkedő szavak sokszor magukkal hozzák a másik nyelv szemantikai világát és sokszor a nyelvtanát is.
- A nyelvtan a nyelvben több formai szinten jelenik meg egyidejűleg. A ragozó nyelvek esetében az alábbi szintek különíthetőek el egymástól: a mondatlánc szintje; a mondat szinte, ahol a szavak sorrendiségét ellenőrzik; a szóösszetételek szintje; valamint a szavak ragozási szintje.

Jelentés:

- A szavaknak nincs rögzített értelmezési tartománya. A szavak jelentése folyamatosan bővülhet és módosulhat. Az egyes zsargonokban a hétköznapi értelemtől eltérő jelentéssel párosulhatnak a szavak. A tartalmi módosulás nyelvtani módosulással is járhat, hiszen új szófajként is megjelenik az adott szó. Így a helyes nyelvtani használat csak a kontextus pontos ismeretében válik egyértelművé.
- Ugyanaz lehet a neve a különböző fogalmaknak és személyeknek: A hononímák miatt a szemantikai és szintaktikai értelmezés csak a környezet, a kontextus pontos ismeretében válik lehetővé. Ez a jelenség a felhasznált nyelvtan környezet függőségét igényli.
- A természetes nyelvekben sokszor implicit alakban jelennek meg a mondat egyes elemei. Például

nem ismételjük meg az alanyt, ha a soron következő mondatban újból róla lesz szó. Egyes nyelvekben nem kötelező a mondat minden szemantikai alapegységének explicit megadása. Ezen esetekben ismét a kontextus ismeret ad támpontot a hiányzó elemről.

A felsorolt nehézségek miatt sokszor nem az alap formális nyelvi eszközöket alkalmazzák, ehelyett speciális optimalizált rendszerek kerülnek megvalósításra. A ragozások esetében például szokás minta adatbázisokat is használni [6], amely nyelv lehetséges ragozási mintáit tartalmazza. Ebből az adatbázisból gyorsan lekérdezhető a vizsgált szóhoz tartozó szótő. Egy másik elterjedt megoldás az IF- THEN alapú mintaillesztésen alapuló átalakítási szabályok rendszere. Ezen az elven működik az ismert Porter szótövező[1] és magyar nyelvre kifejlesztett Tordai szótövező [4] rendszerek is. A magyar nyelvi interfészre épülő robotvezérlési rendszerben is egy speciális toldalékolási adatbázis [6] került felhasználásra. Az adatbázisban egy adott toldalékhoz az alábbi információk kerülnek letárolásra:

Mező neve	Leírás
ID	A toldalék azonosítója
Érték	A toldalék szöveges alakja, pl. *'-ban'*.
Hangillesztés	A csatlakozó részek közötti illesztési szabály
Típus	A toldalék jellege: *'képző'*, *'jel' 'rag'*.
Kód	A toldalékolás szimbóluma *cas<acc>*.
Kezdő osztály	A forrás elem osztálya.
Eredmény osztály	A cél alak nyelvtani osztálya
Záróelem	A toldalék lehet-e záróelem vagy sem

Táblázat 2., Toldalékolási attribútumok

A fenti adatbázis a köznapi nyelvi szituációk gyűjteményére épül. A feldolgozás egy további problémáját az jelenti, hogy a természetes nyelv korpuszának (egy nyelv adott időpontban használt változatára vonatkozó szövegek összességének) legyen az akár egyetlen dialógus nem kell következetesnek lennie sem nyelvtani sem jelentéstani szempontból. Hiszen minden embernek megvan a saját nyelvtana, a saját alkalmazási kontextusa. A nyelvtani szabályok ugyan konvergációs szerepet töltenek be, de ez a konvergáció nem eredményez azonosságot. Ezáltal a dialógus végeredményben inkább kapcsolódó nyelvtanok egy gyűjteményén alapul, mintsem egyetlen nyelvtanon.

4. ÖSSZEFOGLALÁS

A tanulmány a természetes nyelvek modellezési lehetőségeit foglalja össze a formális nyelvek szemszögéből. A lehetséges formális nyelvi elemek a nyelvi jelenségeknek csak egy részét képesek lefedni. Emiatt a gyakorlati természetes nyelvi interfészekben rendszerint heurisztikus megközelítéseket alkalmaznak.

KÖSZÖNETNYÍLVÁNÍTÁS

A bemutatott kutató munka a TÁMOP-4.2.1.B-10/2/KONV-2010-0001 jelű projekt részeként az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

IRODALOMJEGYZÉK

[1] Porter MF (1980) An algorithm for suffix stripping. Program, 14: 130-137.
[2] Paice, C. D. (1994). An evaluation method for stemming algoritms. In Proceedings of ACM-SIGIR94, pages 42–50.
[3] Krovetz, R. (1993). Viewing morphology as an inference process. In Proceedings of ACM-SIGIR93, pages 191–203.
[4] Tordai, A., de Rijke, M.: Hungarian monolingual retrieval at clef (2005)
[5] Chomsky, N.: Aspects of the Theory of Syntax. Cambridge, MIT Press, 1965.
[6] Kovács L. Barabás P: Efficient Enconding of Inflecion Rules in NLP Systems, Bulletin of Univ. Targu Mures, submitted.
[7] Gildea, D., Jurafsky, D.: Automatic Induction of Finite State Transducer for Simple Phonological Rules, Meeting of ACL, 1995.
[8] Harris, Z.: Methods in Structural Linguistics, University of Chicago Press, 1951.
[9] Manning, C., Schütze, H.: Foundations of Statistical Natural language Processing, MIT Press, 1999.
[10] Wallace, C.:Seneca Morphology, International Journal of American Linguistic, 1960.
[11] Jurafsky D., Martin J. H.: Speech and Language Processing, Prentice Hall, 2000.
[12] Krenn, B., Samuelsson C.:The Linguistic's Guide to Statistics, 1997.
[13] Hudson, R.: Language Networks: The new Word Grammar, Oxford University Press, 2007
[14] Tesniere,L.:Elements de syntaxe structurale, Paris, Klincksieck, 1959